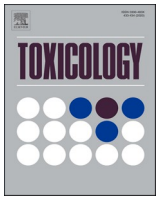
Toxicology 511 (2025) 154064

Contents lists available at ScienceDirect

Toxicology

journal homepage: www.elsevier.com/locate/toxicol



InterDIA: Interpretable prediction of drug-induced autoimmunity through ensemble machine learning approaches

Lina Huang, Peineng Liu, Xiaojie Huang *

Department of Clinical Pharmacy, Jieyang People's Hospital 522000, China

ARTICLE INFO

Edited by Mathieu Vinken

Keywords:
Drug-induced autoimmunity
Predictive toxicology
Interpretable machine learning
Ensemble resampling approaches
SHAP analysis

ABSTRACT

Drug-induced autoimmunity (DIA) is a non-IgE immune-related adverse drug reaction that poses substantial challenges in predictive toxicology due to its idiosyncratic nature, complex pathogenesis, and diverse clinical manifestations. To address these challenges, we developed InterDIA, an interpretable machine learning framework for predicting DIA toxicity based on molecular physicochemical properties. Multi-strategy feature selection and advanced ensemble resampling approaches were integrated to enhance prediction accuracy and overcome data imbalance. The optimized Easy Ensemble Classifier achieved robust performance in both 10-fold cross-validation (AUC value of 0.8836 and accuracy of 82.81 %) and external validation (AUC value of 0.8930 and accuracy of 85.00 %). Paired case studies of hydralazine/phthalazine and procainamide/N-acetylprocainamide demonstrated the model's capacity to discriminate between structurally similar compounds with distinct immunogenic potentials. Mechanistic interpretation through SHAP (SHapley Additive exPlanations) analysis revealed critical physicochemical determinants of DIA, including molecular lipophilicity, partial charge distribution, electronic states, polarizability, and topological features. These molecular signatures were mechanistically linked to key processes in DIA pathogenesis, such as membrane permeability and tissue distribution, metabolic bioactivation susceptibility, immune protein recognition and binding specificity. SHAP dependence plots analysis identified specific threshold values for key molecular features, providing novel insights into structure-toxicity relationships in DIA. To facilitate practical application, we developed an open-access web platform enabling batch prediction with real-time visualization of molecular feature contributions through SHAP waterfall plots. This integrated framework not only advances our mechanistic understanding of DIA pathogenesis from a molecular perspective but also provides a valuable tool for early assessment of autoimmune toxicity risk during drug development.

1. Introduction

Drug-induced autoimmunity (DIA) is a severe adverse drug reaction wherein medication acts as an environmental trigger that impairs immune tolerance, leading to an inappropriate immune attack on host tissues (Dedeoglu, 2009). One of the most significant manifestations is drug-induced lupus (DIL), accounting for approximately 10 % of systemic lupus erythematosus (SLE) cases (Vedove et al., 2009). Since the first reported case of sulfadiazine-induced SLE in 1945, over 100 drugs have been identified as potential DIA triggers, causing various autoimmune conditions including autoimmune hemolytic anemia, autoimmune hepatitis, Sjogren's syndrome, rheumatoid arthritis, polymyositis, and dermatomyositis (Chang and Gershwin, 2011).

Unlike typical drug allergies characterized by IgE-mediated immediate hypersensitivity reactions, DIA represents an idiosyncratic, non-IgE immune response that develops insidiously, often manifesting months or years after drug initiation (Chang and Gershwin, 2010). The development of DIA involves complex interactions between genetic polymorphisms, underlying health conditions, and environmental factors, making it highly unpredictable (Xiao and Chang, 2014). The diagnosis remains challenging due to diverse clinical manifestations and the absence of specific serological markers, primarily relying on patient history, physical examination, and exclusion of other causes.

Several studies have proposed key mechanisms underlying DIA, including DNA demethylation, T-cell signal transduction blockage, adenosine diphosphate ribosylation, histone deacetylase inhibition, macrophage activation, antigen modification (haptenization), defective de novo T-cell development, and apoptosis (Chang and Gershwin, 2010;

* Corresponding author.
E-mail address: huangxj46@mail3.sysu.edu.cn (X. Huang).

https://doi.org/10.1016/j.tox.2025.154064
Received 11 December 2024; Received in revised form 15 January 2025; Accepted 23 January 2025
Available online 25 January 2025

Szyper-Kravitz and Shoenfeld, 2008). Among these mechanisms, the formation of reactive drug metabolites deserves particular attention (Rubin, 2015; Uetrecht, 2005). Despite structural and pharmacological differences among drugs, their in vivo metabolism can generate products with similar immune-perturbing properties. This metabolic convergence helps explain how structurally diverse drugs can trigger comparable autoimmune responses (Rubin, 2015, 2021; Vedove et al., 2009). These reactive metabolites can disrupt immune tolerance through multiple pathways, such as inducing apoptosis and interacting with macromolecules, ultimately leading to autoimmunity.

The toxicity of a chemical can be related directly to its structure; thus, studies of quantitative structure-activity/toxicity relationships (QSAR/QSTR) should shed light on understanding the molecular mechanisms of action (Sun et al., 2012). Given that DIA represents an idiosyncratic toxicity reaction, computational chemistry methods are essential to elucidate how structurally diverse drugs can induce similar immune-perturbing properties. Recently, machine learning approaches have demonstrated great advancement for accelerating chemical risk evaluation and toxicity prediction by automatically identifying potential associations between molecular properties or structural features and compound toxicity (Vo et al., 2020; Wambaugh et al., 2019; Yang et al., 2018). Unlike traditional in vivo toxicity testing characterized by high-cost animal models with low-throughput readouts, machine learning-based predictive toxicology offers incomparable advantages in terms of throughput, cost, and expandability to virtual compounds (Sun et al., 2012; Vo et al., 2020). Recent bibliometric analysis has underlined the significance of modern cheminformatics and QSAR modeling in drug discovery and toxicity prediction (Banerjee et al., 2024).

Inspired by these advances in machine learning-based toxicity prediction, researchers have begun to explore structure-based approaches for predicting DIA toxicity (Guo et al., 2022; Wu et al., 2021). Wu et al., developed a machine learning model incorporating structural alerts (particularly nitrogen-containing benzene substituents) and daily dose information, achieving an area under the curve (AUC) of 70 %. Guo et al., proposed a molecular fingerprint-based model achieving 76.26 % accuracy and 0.84 AUC on their validation set. These structure-based predictions demonstrate that DIA prediction, previously considered unpredictable, can indeed be assessed through computational approaches, providing valuable insights for future research. However, these studies face several limitations. The limited number of DIA-positive drugs creates substantial dataset imbalance, challenging the development of robust machine learning models. Notably, Wu's model showed significant prediction bias with a sensitivity of only 40 %, largely due to unaddressed data imbalance issues, while Guo's strategy of random undersampling DIA-negative drugs potentially compromises the model's ability to learn from non-toxic compounds. Additionally, both studies lack applicability domain analysis and model interpretability investigation, making it difficult to determine reliable prediction boundaries and understand the molecular mechanisms underlying their predictions, thus limiting their practical applications.

In this study, we present InterDIA, a comprehensive and interpretable machine learning framework for predicting DIA toxicity. Unlike previous approaches that relied heavily on structural alerts and fingerprints, we systematically investigated the physicochemical properties of drugs using multiple molecular descriptor sets. To address the inherent data imbalance challenge, we implemented advanced ensemble resampling techniques combined with multi-strategy feature selection approaches. Most importantly, we pioneered the application of SHAP (SHapley Additive exPlanations) analysis to provide mechanistic insights into molecular determinants of DIA, offering both global patterns and compound-specific interpretations. Furthermore, we developed a freely accessible web platform (https://drug-induced-autoimmunity-predictor.streamlit.app/) that enables batch prediction with real-time visualization of feature contributions, facilitating practical application in drug development. To facilitate future research, we have made our complete dataset, source code, and trained models openly accessible via our GitHub repository: https://github.com/Huangxiaojie2024/InterDIA. This integrated approach not only advances our understanding of DIA mechanisms from a molecular perspective but also provides a valuable tool for early assessment of autoimmune toxicity risk during drug discovery. A schematic representation of our proposed interpretable machine learning framework is presented in Fig. 1.

2. Materials and methods

2.1. Methodological overview

In this work, we propose InterDIA, an interpretable machine learning framework for predicting drug-induced autoimmunity. By integrating state-of-the-art ensemble learning approaches, we developed and implemented an accurate and reliable prediction pipeline, which is freely available through our free online prediction platform. The platform incorporates SHAP waterfall plots for each predicted compound, enabling intuitive visualization of molecular features that contribute to DIA toxicity prediction.

As illustrated in Fig. 1, our framework consists of three main components: (i) feature representation, which characterizes compounds through diverse molecular descriptor sets; (ii) learning framework, which integrates feature preprocessing and multi-strategy selection with ensemble resampling techniques to enhance model performance and address data imbalance; and (iii) explainability analysis, which utilizes SHapley Additive exPlanations (SHAP) methodology to elucidate model predictions at both global and local levels. This integrated approach not only enables accurate DIA toxicity prediction but also reveals molecular mechanisms underlying autoimmune responses, thereby supporting rational drug design and toxicity evaluation in both pre-clinical and post-marketing stages.

2.2. Dataset collection and feature representation

The dataset used in this study was obtained from previous research (Guo et al., 2022), where DIA-positive drugs were compiled by combining compounds from the Side Effect Resource (SIDER) database (Kuhn et al., 2016) with documented autoimmune adverse reactions (incidence rate $\geq$0.1 %) and additional positive cases from Wu et al.'s study (Wu et al., 2021). DIA-negative drugs were also extracted from Wu et al.'s work (Wu et al., 2021). All chemical structures were standardized using canonical Simplified Molecular Input Line Entry System (SMILES) representation. The final dataset comprised 597 drugs (148 DIA-positive and 449 DIA-negative), which was randomly partitioned into training (477 drugs) and external validation (120 drugs) sets at an 8:2 ratio while maintaining the class distribution (Table 1). The complete dataset with canonical SMILES and corresponding labels is available in the Supporting Information (Table S1 and S2).

For molecular feature representation, we adopted a comprehensive descriptor-based approach in contrast to previous studies that primarily relied on structural alerts, daily dose, and molecular fingerprints. To capture diverse chemical characteristics, we calculated four sets of molecular descriptors using multiple computational platforms: RDKit descriptors via the open-source ChemDes platform (http://www.scbdd.com/chemdes/) (Dong et al., 2015), Mold2 descriptors through Mold2 software (Hong et al., 2008), Discovery Studio (DS) descriptors using Discovery Studio 2019, and MOE descriptors via MOE software (version 2022.02). This integrated multi-platform strategy generated a total of 1622 molecular descriptors (Table 2), enabling thorough characterization of both physicochemical properties and structural features.

2.3. Learning framework

The proposed learning framework integrates a comprehensive feature processing pipeline with advanced ensemble learning approaches to address the data imbalance challenge. The feature

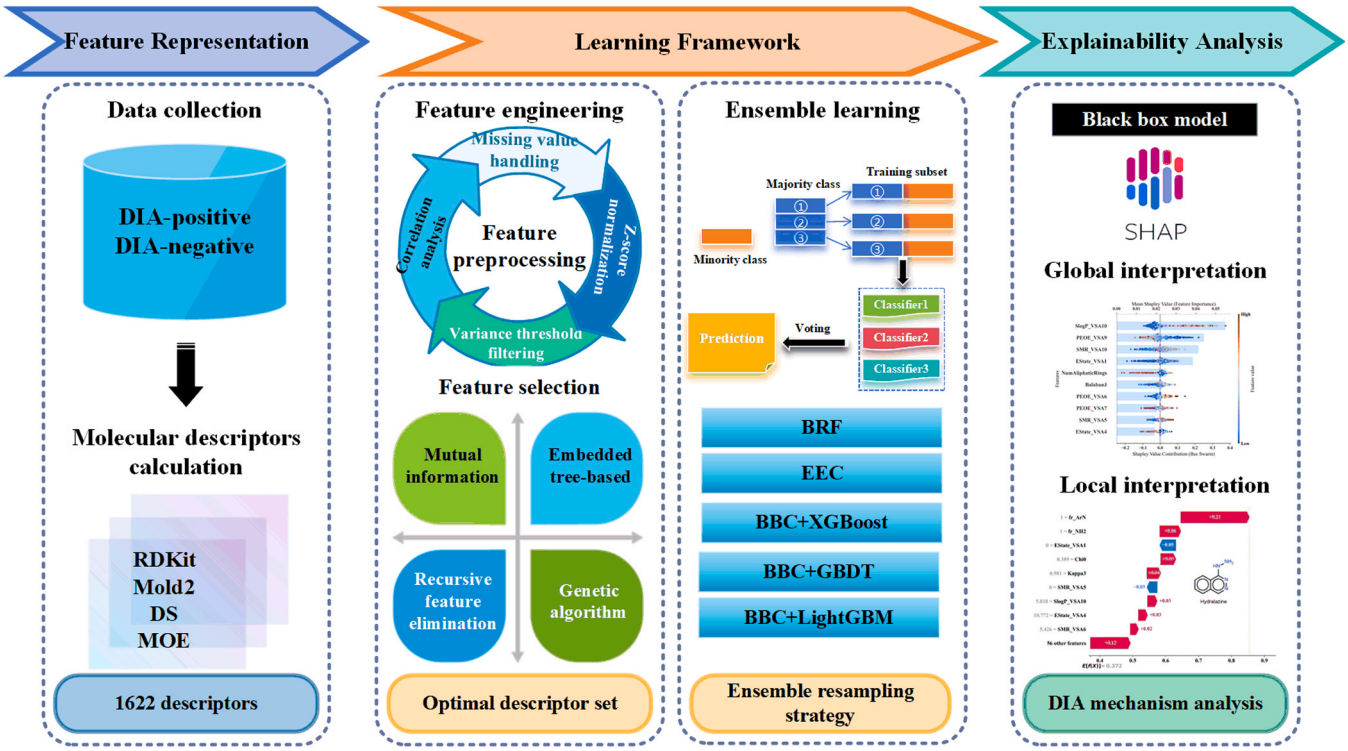


Fig. 1. Schematic illustration of the InterDIA framework for predicting drug-induced autoimmunity.

Table 1
Distribution of DIA-positive and DIA-negative drugs in the training and external validation sets.

Datasets	DIA-positive drugs	DIA-negative drugs	Total
Training set	118	359	477
External validation set	30	90	120
Total	148	449	597

Table 2
Overview of molecular descriptor sets for DIA toxicity prediction.

Descriptor set	Number of descriptors
RDKit	196
Mold2	777
DS	440
MOE	209
Total	1622

processing pipeline comprised two sequential steps: feature preprocessing and feature selection. In the feature preprocessing stage, we implemented four procedures including missing value handling, z-score normalization of molecular descriptors, variance threshold filtering for zero-variance features removal, and correlation analysis to eliminate multicollinearity (threshold: 0.9).

After preprocessing, we employed four complementary feature selection techniques to identify the most relevant molecular descriptors. Mutual Information (MI) analysis captured both linear and nonlinear feature-target relationships, retaining features with MI values above zero (Vergara and Estévez, 2014). Embedded tree-based feature selection (ETB) (Genuer et al., 2010) using balanced random forest (BRF) classifier (Chen et al., 2004) preserved features with importance scores exceeding 0.003. Recursive Feature Elimination with Cross-Validation (RFECV) iteratively removed less important features while monitoring model performance (Darst et al., 2018). Additionally, a Genetic Algorithm (GA) (Leardi et al., 1992) optimization approach (population size: 50, generations: 40, crossover rate: 0.5, mutation rate: 0.2) with BRF classifier explored potential optimal feature combinations.

Given the significant class imbalance in our dataset (DIA-positive: DIA-negative ratio = 1:3), we utilized three advanced ensemble resampling algorithms from the imbalanced-learn package (version 0.12.3) (Lemaître et al., 2017). Balanced Random Forest (BRF) creates balanced training subsets by combining bootstrap samples from the minority class with randomly selected majority class samples (Chen et al., 2004). Easy Ensemble Classifier (EEC) generates multiple balanced subsets through random undersampling and trains an ensemble of AdaBoost learners (Liu et al., 2009). Balanced Bagging Classifier (BBC) extends traditional bagging by creating balanced bootstrap samples through majority class undersampling (Maclin and Opitz, 1997).

Based on these ensemble resampling strategies, we constructed five distinct machine learning models. BRF and EEC were implemented directly with their embedded base learners (decision trees and AdaBoost, respectively). For BBC, we integrated three different base learners: gradient boosting decision tree (GBDT) (Friedman, 2001), light gradient boosting machine (LightGBM) (Ke et al., 2017), extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016). The implementation utilized multiple Python packages: imbalanced-learn (version 0.12.3) for ensemble resampling, Scikit-learn (version 1.5.1) for GBDT, lightgbm (version 3.3.5) for LightGBM, and xgboost (version 1.6.1) for XGBoost.

Model hyperparameters were optimized using Bayesian optimization through Hyperopt (version 0.2.7) (Bergstra et al., 2013) with the tree-structured Parzen estimator (TPE). This approach efficiently explores the hyperparameter space by focusing on promising regions, with the Matthews correlation coefficient (MCC) serving as the optimization metric due to its robustness in handling imbalanced datasets.

This comprehensive learning framework leverages complementary feature selection methods, advanced ensemble resampling algorithms, and diverse base learners to achieve robust DIA prediction while effectively addressing the class imbalance challenge.

2.4. Performance evaluation

To ensure robust evaluation, model performance was primarily assessed through out-of-fold predictions from 10-fold cross-validation (cv), complemented by external validation on an independent test set. We employed multiple evaluation metrics, including accuracy (ACC), sensitivity (SEN, also known as recall), specificity (SPE), and Matthews correlation coefficient (MCC). These metrics were calculated from the confusion matrix components: true positive (TP, correctly identified DIA-positive drugs), true negative (TN, correctly identified DIA-negative drugs), false positive (FP, DIA-negative drugs misclassified as positive), and false negative (FN, DIA-positive drugs misclassified as negative). The metrics were calculated as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{1}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{3}$$

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

Additionally, we evaluated the area under the receiver operating characteristic curve (AUC) to assess the model's discriminative ability across different classification thresholds. Given the imbalanced nature of our dataset, we prioritized MCC and AUC as primary evaluation metrics due to their established robustness in handling class imbalance, thus providing a more reliable assessment of model performance (Boughorbel et al., 2017; Chicco and Jurman, 2020; Halimu et al., 2019).

2.5. Applicability domain

The applicability domain (AD) assessment is essential for defining the chemical space within which model predictions can be considered reliable, as emphasized by the Organization for Economic Co-operation and Development (OECD) guidelines for structure-activity relationship (SAR) models (Co-operation and Development, 2014). To establish the model's AD, we employed the Euclidean distance method implemented in Euclidean Applicability Domain 1.0 software (Kar et al., 2018). This approach calculates normalized mean distance scores (ranging from 0 to 1) based on the molecular descriptors of compounds in the training set to define the prediction reliability boundary. Compounds with descriptor values falling outside this established domain are considered structurally distinct from the training set and thus potentially unreliable for prediction, ensuring scientifically sound application of our model in real-world scenarios.

2.6. Model explainability analysis

To enhance model transparency, we implemented the SHAP-based interpretation method at both global and local levels (Lundberg and Lee, 2017). Given the ensemble nature of our models, we employed Kernel SHAP, a model-agnostic approach that can interpret predictions from any machine learning model, making it particularly suitable for analyzing complex ensemble learning algorithms. For global interpretation, we generated SHAP summary plots with bee swarm visualization to reveal the overall impact and distribution of molecular descriptors on model predictions. The top 10 influential features were further analyzed using SHAP dependence plots to elucidate their specific relationships with DIA toxicity prediction. At the individual compound level, we employed SHAP waterfall plot to provide detailed visualization of feature contributions for single predictions. Additionally, we integrated SHAP waterfall plot into our online prediction platform, offering an interactive and intuitive visualization tool for individual drug assessments. This multi-level explainability approach not only uncovers the general patterns of molecular characteristics influencing DIA toxicity but also provides transparent, compound-specific prediction rationales to support practical decision-making.

3. Results and discussion

3.1. Identification of the optimal molecular descriptor subset for DIA prediction

Feature preprocessing and selection are crucial steps in developing robust machine learning models, particularly when dealing with high-dimensional molecular descriptors (1622 features from four platforms) that may contain redundant or irrelevant information (Guyon and Elisseeff, 2003). To identify the most informative descriptor subset for DIA prediction, we systematically evaluated various feature selection strategies using a BRF classifier, with performance assessed through out-of-fold predictions from 10-fold cross-validation on the training set. The MCC and AUC were chosen as primary evaluation metrics due to their robustness in handling imbalanced datasets.

3.1.1. Comparative analysis of individual descriptor sets performance

The predictive performance of four individual molecular descriptor sets (RDKit, Mold2, DS, and MOE) was assessed using four feature selection methods: MI, ETB, RFECV, and GA. Detailed results are provided in Table S3, while Table 3 summarizes the optimal feature subset and corresponding model performance for each descriptor type. GA consistently identified the best-performing feature subsets across all descriptor types. RDKit descriptors demonstrated superior performance, with the GA-selected subset of 65 optimal features achieving the highest predictive accuracy (AUC = 0.8764, MCC = 0.5841) and balanced performance across metrics (ACC = 81.76 %, SEN = 83.05 %, SPE = 81.34 %). The MOE descriptor set with 58 GA-selected features ranked second in terms of overall performance (AUC = 0.8467, MCC = 0.5560) and notably achieved the highest sensitivity (84.75 %), followed by the Mold2 descriptor set comprising 157 features (AUC = 0.8230, MCC = 0.4899). Despite its compact size (57 features), the DS descriptor set exhibited the lowest performance (AUC = 0.8127, MCC = 0.4699). These results suggest that the GA-selected subset of 65 RDKit descriptors provides the most informative and discriminative molecular representation for DIA prediction.

3.1.2. Comparative analysis of hybrid descriptor sets performance

Motivated by the superior performance of RDKit descriptors, we explored seven hybrid combinations with other descriptor sets to further enhance the predictive power (Table S3 and Table 4). Among these combinations, the RDKit+DS+MOE subset optimized through RFECV demonstrated robust performance (AUC = 0.8627, MCC = 0.5522) with remarkable feature reduction efficiency. Comparative analysis revealed that while the RFECV-optimized subset showed marginally lower MCC (0.5522 vs. 0.5538) compared to its GA-optimized counterpart, it achieved higher discriminative capability (AUC = 0.8627 vs. 0.8481) with significantly fewer molecular descriptors (43 vs. 171). Furthermore, this model achieved the highest sensitivity (83.90 %) among all hybrid feature subsets, suggesting its particular strength in identifying DIA-positive drugs. The RDKit+DS and RDKit+MOE combinations also showed promising performance (AUC = 0.8541 and 0.8574, MCC = 0.5463 and 0.5452, respectively). However, more complex combinations incorporating all four descriptor types (RDKit+Mold2 +DS+MOE) did not yield better performance, suggesting that the increased descriptor complexity does not necessarily translate to enhanced predictive capability. These findings highlight the critical role of efficient feature selection in identifying optimal molecular descriptor subsets for DIA prediction.

Table 3
Performance comparison of individual molecular descriptor sets. (Bold values indicate the best performance for each metric.).

Descriptor set	Feature selection method	Optimal No. of descriptors	AUC	ACC	SEN	SPE	MCC
RDKit	**GA**	**65**	**0.8764**	**81.76 %**	83.05 %	**81.34 %**	**0.5841**
Mold2	GA	157	0.8230	76.52 %	79.66 %	75.49 %	0.4899
DS	GA	57	0.8127	75.68 %	77.97 %	74.93 %	0.4699
MOE	GA	58	0.8467	79.45 %	**84.75 %**	77.72 %	0.5560

Table 4
Performance comparison of hybrid molecular descriptor sets. (Bold values indicate the best performance for each metric.)

Descriptor set	Feature selection method	Optimal No. of descriptors	AUC	ACC	SEN	SPE	MCC
RDKit+Mold2	GA	204	0.8404	79.25 %	77.97 %	79.67 %	0.5228
RDKit+DS	GA	105	0.8541	**80.50 %**	78.81 %	**81.06 %**	0.5463
RDKit+MOE	Feature Preprocessing	249	0.8574	79.25 %	83.05 %	77.99 %	0.5452
RDKit+Mold2 +DS	GA	255	0.8266	78.62 %	77.97 %	78.83 %	0.5131
RDKit+Mold2 +MOE	GA	278	0.8428	78.20 %	81.36 %	77.16 %	0.5220
RDKit+DS+MOE	**RFECV**	**43**	**0.8627**	79.45 %	**83.90 %**	77.99 %	**0.5522**
RDKit+Mold2 +DS+MOE	ETB	120	0.8468	79.04 %	81.36 %	78.27 %	0.5345

Based on these evaluations, two feature subsets were identified as optimal candidates for DIA prediction model development: the GA-selected RDKit subset with 65 features (RDKit_GA_65) and the RFECV-optimized combination of RDKit, DS, and MOE descriptors with 43 features (RDKit+DS+MOE_RFECV_43). The detailed composition of these feature subsets is provided in Table S4.

3.2. Performance evaluation of ensemble machine learning models for DIA prediction

To tackle data imbalance issue and develop robust prediction models for DIA toxicity, we employed five machine learning algorithms embedded with ensemble resampling methods: BRF, EEC, BBC+XGBoost, BBC+GBDT, and BBC+LightGBM. The models were trained using the two optimal feature subsets (RDKit_GA_65 and RDKit+MOE+DS_RFECV_43) with hyperparameters optimized through Bayesian optimization (Table S5). Model performance was evaluated via out-of-fold predictions of 10-fold cross-validation and external validation. The comprehensive performance metrics are presented in Table 5, while the receiver operating characteristic (ROC) curves for external validation are illustrated in Fig. 2.

Among the models built with the RDKit_GA_65 feature subset, the EEC model demonstrated superior performance across all evaluation metrics. In 10-fold cross-validation, it achieved outstanding results (AUC = 0.8836, ACC = 82.81 %, MCC = 0.5978) with well-balanced sensitivity (SEN = 82.20 %) and specificity (SPE = 83.01 %). This strong performance was further validated on the external validation set (AUC = 0.8930, ACC = 85.00 %, MCC = 0.6413), demonstrating robust generalization capability. While the BBC+GBDT model also showed promising performance on the external validation set (AUC = 0.8974, MCC = 0.6093), its relatively low sensitivity (74.58 % in cross-validation and 73.33 % in external validation) indicated limited capability in identifying DIA-positive drugs.

In the RDKit+MOE+DS_RFECV_43 feature subset, the EEC model demonstrated complex performance patterns. During 10-fold cross-validation, it achieved the highest AUC (0.8690) but ranked second in MCC (0.5737) behind BBC+GBDT (0.5793). However, in external validation, the EEC model exhibited superior performance across multiple metrics (AUC = 0.8915, ACC = 82.50 %, SEN = 83.33 %, MCC = 0.5985). In contrast, while BBC+GBDT showed the highest specificity in both cross-validation (SPE = 87.19 %) and external validation (SPE = 86.67 %), it demonstrated considerably lower sensitivity (cross-validation: 72.88 %, external validation: 70.00 %), indicating a potential bias toward negative predictions. Given this comprehensive evaluation, the EEC model emerged as the most balanced performer within this feature subset, offering more reliable predictions across both positive and negative cases. Notably, all models built with this feature subset showed inferior performance compared to their counterparts using RDKit_GA_65 feature subset, suggesting the latter's superior molecular representation capability for DIA prediction.

Based on these evaluations, the EEC model built with the RDKit_GA_65 feature subset emerged as the optimal model for DIA prediction, demonstrating superior performance and robust generalization ability. Consequently, this model was chosen as the basis for further in-depth analysis, including mechanistic interpretations and the development of a web platform for practical applications.

3.3. Applicability domain analysis

To verify our model's reliability, we conducted an applicability

Table 5
Performance evaluation of ensemble models based on out-of-fold predictions of 10-fold cross-validation and external validation. (Bold values indicate the best performance for each metric within each feature subset.)

Feature subset	Model name	Out-of-fold predictions of 10-fold cv					External validation set				
		AUC	ACC	SEN	SPE	MCC	AUC	ACC	SEN	SPE	MCC
RDKit_GA_65	BRF	0.8764	81.76 %	**83.05 %**	81.34 %	0.5841	0.8878	80.83 %	76.67 %	82.22 %	0.5444
	EEC	**0.8836**	82.81 %	82.20 %	83.01 %	**0.5978**	0.8930	**85.00 %**	**83.33 %**	85.56 %	**0.6413**
	BBC+XGBoost	0.8499	79.87 %	77.97 %	80.50 %	0.5327	0.8763	80.83 %	73.33 %	83.33 %	0.5313
	BBC+GBDT	0.8773	**83.65 %**	74.58 %	**86.63 %**	0.5850	**0.8974**	**85.00 %**	73.33 %	**88.89 %**	0.6093
	BBC+LightGBM	0.8653	82.81 %	74.58 %	85.52 %	0.5694	0.8915	84.17 %	73.33 %	87.78 %	0.5926
RDKit+MOE+DS_RFECV_43	BRF	0.8627	79.45 %	**83.90 %**	77.99 %	0.5522	0.8544	78.33 %	**83.33 %**	76.67 %	0.5344
	EEC	**0.8690**	81.55 %	81.36 %	81.62 %	0.5737	**0.8915**	**82.50 %**	**83.33 %**	82.22 %	**0.5985**
	BBC+XGBoost	0.8438	76.94 %	77.12 %	76.88 %	0.4840	0.8756	78.33 %	80.00 %	77.78 %	0.5192
	BBC+GBDT	0.8499	**83.65 %**	72.88 %	**87.19 %**	**0.5793**	0.8826	**82.50 %**	70.00 %	**86.67 %**	0.5495
	BBC+LightGBM	0.8538	81.13 %	69.49 %	84.96 %	0.5204	0.8759	77.50 %	70.00 %	80.00 %	0.4623

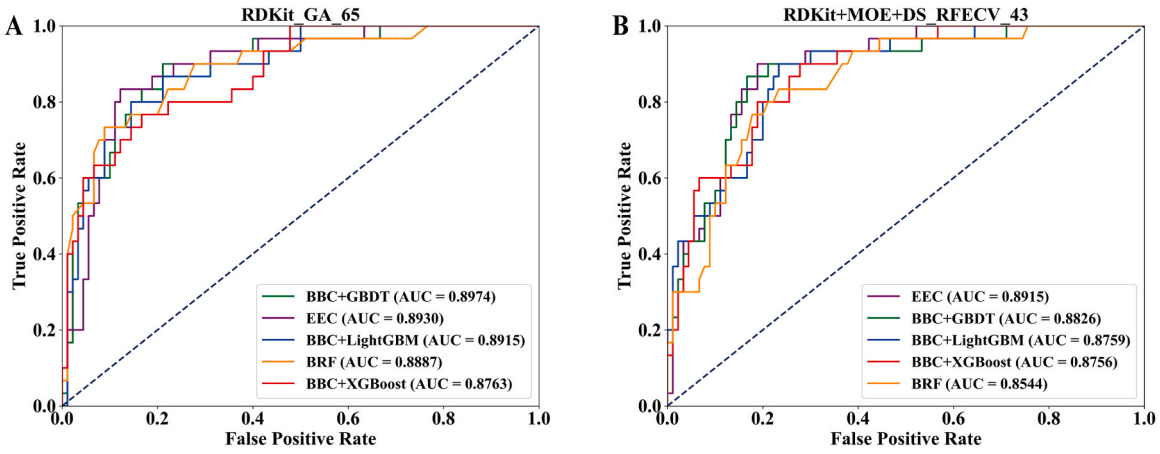


Fig. 2. ROC curves comparing the performance of different ensemble models on the external validation set using (A) RDKit_GA_65 and (B) RDKit+MOE+DS_RFECV_43 feature subsets.

domain (AD) analysis using the Euclidean distance method, following OECD guidelines for SAR model validation (Co-operation and Development, 2014; Kar et al., 2018). The AD defines the chemical space within which the model's predictions can be considered reliable and helps prevent extrapolation beyond the model's training domain. Using the optimized 65 RDKit molecular descriptors, normalized mean distance scores were calculated for compounds in the training and external validation sets. The distributions of these normalized distances are visualized in Fig. 3. The AD boundary (0−1) was established based on the training set compounds. The analysis revealed that only one external validation compound fell outside the AD boundary, indicating that most compounds share similar molecular characteristics with the training set and lie within the model's predictive domain. Therefore, the predictions for the external validation set in this study have been demonstrated to be highly reliable. Establishing a well-defined AD and rigorously assessing the model's performance within this domain are crucial for building confidence in its robustness and predictive accuracy. For practical applications in DIA toxicity prediction, we recommend that predictions for compounds falling outside the AD should be treated with caution and may require additional validation.

3.4. Mechanistic interpretation of molecular features for DIA prediction

3.4.1. Global analysis of DIA-associated molecular properties

DIA represents a complex idiosyncratic adverse drug reaction whose underlying molecular mechanisms remain incompletely understood (Guo et al., 2022; He and Sawalha, 2018; Uetrecht, 2005). Little attention has been paid to the fundamental physicochemical properties of drugs that may initiate or influence DIA development in previous studies. To bridge this knowledge gap, we employed SHAP (SHapley Additive exPlanations) analysis on our optimal EEC model to systematically investigate DIA mechanisms from a molecular physicochemical perspective.

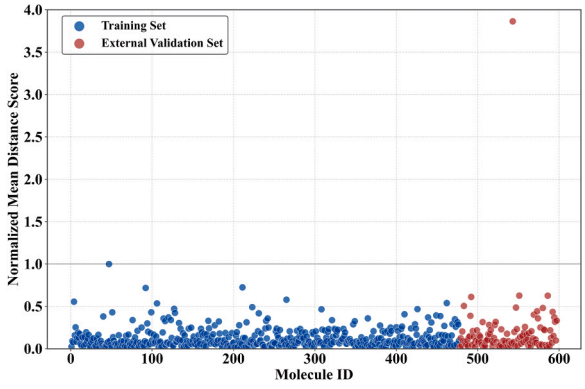


Fig. 3. Applicability domain analysis of the EEC model.

The SHAP summary plots for both training (Fig. 4A) and external validation sets (Fig. 4B) revealed remarkable consistent patterns of feature importance. Notably, identical top 10 molecular descriptors were identified as crucial determinants across both sets, validating our model's robust capability in capturing intrinsic molecular characteristics associated with DIA toxicity rather than dataset-specific patterns. These key molecular descriptors encompass diverse physicochemical properties including lipophilicity (SlogP_VSA10), partial charge distribution (PEOE_VSA6, PEOE_VSA7, PEOE_VSA9), electronic state (EState_VSA1, EState_VSA4), molecular polarizability (SMR_VSA5, SMR_VSA10), and topological features (NumAliphaticRings, BalabanJ). Together, these descriptors highlight the critical association between specific molecular properties and potential DIA toxicity mechanisms, providing insights into how physicochemical characteristics influence autoimmune responses.

The SHAP dependence plots (Fig. 4C), ordered by the top 10 feature importance derived from the training set, further elucidate specific relationships between molecular features and DIA prediction, revealing distinct threshold effects and nonlinear patterns. SHAP values higher than zero correspond to a positive class prediction in the model, indicating a higher risk of DIA toxicity, while negative SHAP values suggest a negative class prediction and lower DIA toxicity risk. The detailed description, importance ranking of these features, and their SHAP value thresholds (positive/negative) for DIA toxicity prediction are summarized in Table 6.

(1) Lipophilicity-mediated effects

The SlogP_VSA10 descriptor (rank 1), representing the van der Waals surface areas of atoms with logP contributions in [0.4, 0.5), was identified as the most influential feature in DIA prediction. The SHAP dependence plot reveals a clear threshold effect at 6.18, where values exceeding this threshold significantly increase DIA risk (SHAP value > 0). This relationship suggests that regions with moderate lipophilicity may promote DIA through potential multiple mechanisms: (1) Tissue distribution: Enhanced transmembrane permeability and tissue distribution, particularly to immune-rich organs, resulting in prolonged immune surveillance exposure; (2) Protein interaction: Direct interaction with self-proteins, leading to protein modification and altered immune tolerance, triggering autoimmune responses through epigenetic regulation and T cell-mediated autoimmunity (Chang and Gershwin, 2010); (3) Metabolic activation: Enhanced

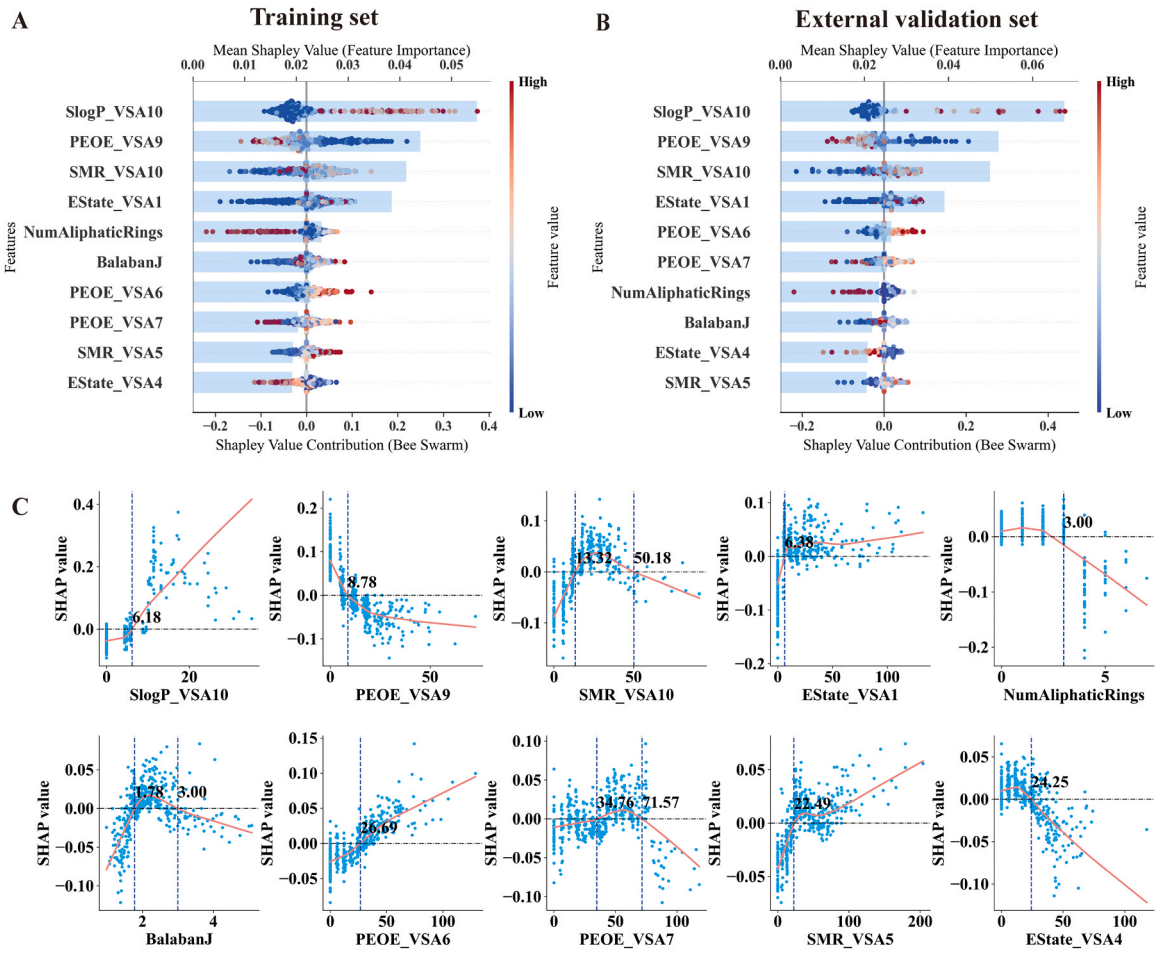


Fig. 4. Global interpretation of molecular features contributing to DIA prediction using SHAP analysis. (A) SHAP summary plot for training set and (B) external validation set showing consistent feature importance patterns across the top 10 molecular descriptors. The x-axis represents the SHAP value contribution (impact on model output), and color indicates the feature value (red for high, blue for low). (C) SHAP dependence plots ordered by the top 10 feature importance from training set, illustrating the relationship between feature values (x-axis) and their impact on model prediction (y-axis), with identified threshold values indicated by vertical dashed lines. The red curves represent LOWESS (Locally Weighted Scatterplot Smoothing) fit to visualize the overall trends. SHAP values higher than zero correspond to a positive class prediction in the model, indicating higher DIA toxicity risk, while negative values suggest lower risk.

Table 6
Summary of top 10 molecular descriptors and their contributions to DIA prediction identified by SHAP analysis.

Descriptor type	Descriptor	Description	Importance rank	SHAP value< 0	SHAP value≥ 0
Lipophilicity	SlogP_VSA10	Sum of van der Waals surface areas of atoms with logP contribution in [0.4, 0.5)	1	[0, 6.18]	(6.18, 35.12)
Partial charge	PEOE_VSA6	Sum of van der Waals surface areas of atoms with PEOE partial charge in [−0.10, −0.05)	7	[0, 26.69]	(26.69, 128.92)
	PEOE_VSA7	Sum of van der Waals surface areas of atoms with PEOE partial charge in [−0.05, 0.00)	8	[0, 34.76] ∪ (71.57, 118.02)	(34.76, 71.57]
	PEOE_VSA9	Sum of van der Waals surface areas of atoms with PEOE partial charge in [0.05, 0.10)	2	(8.78, 72.68)	[0, 8.78]
Electronic state	EState_VSA1	Sum of van der Waals surface areas of atoms with Estate (electrotopological state) indices < −0.39	4	[0, 6.38]	(6.38, 132.39)
	EState_VSA4	Sum of van der Waals surface areas of atoms with Estate (electrotopological state) indices in [0.72, 1.17]	10	(24.25, 117.81)	[0, 24.25]
Polarizability	SMR_VSA5	Sum of van der Waals surface areas of atoms with molecular refractivity in [2.45, 2.75)	9	[0, 22.49]	(22.49, 203.97)
	SMR_VSA10	Sum of van der Waals surface areas of atoms with molecular refractivity ≥ 4.00	3	[0, 13.32] ∪ (50.18, 91.18)	(13.32, 50.18]
Topological features	NumAliphaticRings	Count of aliphatic ring systems (rings containing at least one non-aromatic bond) in the molecule	5	(3, 7]	[0,3]
	BalabanJ	A topological index characterizing molecular connectivity and degree of branching	6	(0.99, 1.78] ∪ (3.00, 5.08)	(1.78, 3.00]

susceptibility to metabolic bioactivation (e.g., by cytochrome P450 enzymes), promoting the formation of reactive metabolites capable of hapten formation (Cho and Uetrecht, 2017; Grattagliano et al., 2009). These patterns demonstrate that molecular lipophilicity plays a multifaceted role in DIA pathogenesis by modulating drug distribution, protein interactions, and metabolic activation pathways, highlighting its significance as a key determinant in immune-mediated adverse reactions.

(2) Partial charge distribution-mediated effects

The electronic property descriptors based on Partial Equalization of Orbital Electronegativity (PEOE) reveal the crucial role of partial charge distribution in DIA development (Gasteiger and Marsili, 1980). Specifically, PEOE_VSA9 (rank 2) characterizes regions with positive partial charges [0.05, 0.10), showing DIA-negative predictions above 8.78, suggesting these positively charged regions may have protective effects against DIA risk; PEOE_VSA6 (rank 7) represents regions with negative partial charges [-0.10, −0.05), showing positive correlation with DIA risk, with DIA-positive predictions above 26.69; while PEOE_VSA7 (rank 8) describes regions with weak negative charges [-0.05, 0.00), displaying a complex pattern with DIA-positive predictions within (34.76, 71.57]. The influence of these charge distributions on DIA development can be explained through two key mechanisms: (1) Electrostatic interactions with immune-related proteins: The positive contribution of PEOE_VSA6 (negative charges) and negative contribution of PEOE_VSA9 (positive charges) suggest that negatively charged regions facilitate binding to positively charged active sites of immune-related proteins. This charge complementarity may enhance immune recognition and subsequent autoimmune responses. (2) Metabolic activation susceptibility: Regions with negative charges may promote oxidation by myeloperoxidase and cytochrome P450 enzymes, leading to formation of reactive metabolites. For instance, aromatic amine-containing drugs can undergo oxidation to generate reactive aromatic hydroxylamine species, a process facilitated by the electronic properties of these moieties. These metabolites can trigger autoimmunity through various pathways including protein modification and immune cell activation (Rubin, 2015). These patterns demonstrate that molecular partial charge distribution may influences both protein-drug recognition and metabolic bioactivation in DIA pathogenesis.

(3) Electronic state-mediated effects

The electronic state descriptors based on EState (Electrotopological State) indices combine electronic properties and topological structure information, providing insights into atomic electronic accessibility and reactivity. Specifically, EState_VSA1 (rank 4) represents the van der Waals surface areas of atoms with EState $<$ -0.39, showing positive correlation with DIA risk with DIA-positive predictions above 6.38. A higher EState_VSA1 value indicates larger surface areas of electron-deficient atoms (regions with strong electron-withdrawing properties) with high topological accessibility (Zhu et al., 2024). Conversely, EState_VSA4 (rank 10) characterizes regions with EState values in [0.72, 1.17], showing DIA-negative predictions above 24.25, suggesting that regions with moderate electron-rich properties may have protective effects. These electronic state characteristics likely influence DIA development through several mechanisms: (1) Protein-binding modification: Electron-withdrawing groups create electrophilic sites in adjacent regions, facilitating covalent interactions with nucleophilic residues on proteins. The resulting modified protein adducts can be recognized as neoantigens by the immune system, subsequently triggering autoantibody production and immune-mediated responses (Chipinda et al., 2011; Zhou et al., 2005). (2) Immune protein recognition: The electron density distribution governs molecular recognition by immune-related proteins, determining the specificity and strength of these interactions. These findings underscore the critical role of atomic electronic and topological properties in determining covalent modifications and protein-binding pathways in the context of DIA development.

(4) Molecular polarizability-mediated effects

The SMR_VSA descriptors characterize the distribution of van der Waals surface areas based on atomic contributions to molecular refractivity, quantifying molecular polarizability (Balaji et al., 2004). Specifically, SMR_VSA10 (rank 3) represents regions with molecular refractivity $\geq$ 4.00, exhibiting a complex triphasic relationship: protective effects in [0, 13.32], DIA-promoting effects within (13.32, 50.18], suggesting optimal polarizability for drug-protein interactions, and reverting to protective effects when exceeding 50.18. In contrast, SMR_VSA5 (rank 9) characterizes regions with moderate molecular refractivity [2.45, 2.75), showing positive correlation with DIA risk with DIA-positive predictions above 22.49. These molecular refractivity patterns may influence DIA development through several mechanisms: (1) Drug-protein interaction: Mediating binding interactions through induced dipole-dipole and van der Waals forces; (2) Molecular recognition: Facilitating binding site complementarity through appropriate electronic polarization (Muzet et al., 2003); (3) Distribution characteristics: Determining membrane permeability and tissue distribution, thus regulating drug exposure to immune system components. These findings highlight how the balance and distribution of molecular polarizability critically influence immune recognition processes in DIA development.

(5) Topological features-mediated effects

The topological descriptors characterize molecular scaffold features and their influence on DIA development. Specifically, NumAliphaticRings (rank 5), representing the count of aliphatic ring systems (Chen et al., 2022), shows negative correlation with DIA risk with DIA-negative predictions above 3, suggesting that incorporation of more cyclic structures might reduce immunogenicity. This protective effect may arise through multiple mechanisms:(1) Conformational constraint: Increased molecular rigidity from cyclic structures limits conformational flexibility, potentially reducing non-specific interactions with immune-related proteins; (2) Metabolic shielding: Cyclic structures sterically protect susceptible groups from metabolic enzymes, decreasing reactive metabolite formation; (3) Spatial orientation control: Rigid cyclic scaffolds establish specific three-dimensional arrangements that determine recognition by immune proteins. BalabanJ (rank 6), a topological index reflecting molecular connectivity and branching degree (Balaban, 1982), exhibits a complex triphasic effect: showing DIA-negative predictions in (0.99, 1.78] and (3.00, 5.08), while DIA-positive predictions within (1.78, 3.00]. This relationship indicates that moderate molecular branching provides optimal spatial arrangement for immune recognition, while excessive branching may disrupt these interactions through steric effects. These findings demonstrate how molecular scaffold characteristics critically determine both conformational properties and molecular recognition processes in DIA pathogenesis.

Collectively, these key molecular descriptors reveal multiple interconnected mechanisms underlying DIA development. Lipophilicity (SlogP_VSA10) and molecular polarizability (SMR_VSA) primarily influence drug distribution and protein interactions, while partial charge distribution (PEOE_VSA) and electronic states (EState_VSA) determine both metabolic activation potential and immune protein recognition. The topological features (NumAliphaticRings, BalabanJ) regulate these processes through conformational and steric effects. Together, these findings demonstrate that DIA risk is governed by a precise balance of physicochemical properties that affect three critical aspects: (1) membrane permeability and tissue distribution, (2) metabolic bioactivation

susceptibility, and (3) immune protein recognition and binding specificity. These mechanistic insights not only advance our understanding of DIA pathogenesis but also provide a rational framework for designing drugs with minimized autoimmune toxicity risk during development.

3.4.2. *Activity cliffs analysis: insights from structural analog pairs with contrasting DIA risks*

Activity cliffs are defined as pairs of structurally similar molecules that exhibit large differences in their biological activities or properties (Dimova et al., 2014; Van Tilborg et al., 2022). To validate our model's effectiveness in capturing such discontinuities in structure-toxicity relationships, we analyzed two pairs of structural analogs representing classical activity cliffs in drug-induced autoimmunity: procainamide/N-acetylprocainamide and hydralazine/phthalazine. Both procainamide and hydralazine are established high-risk drugs for drug-induced lupus (DIL), primarily operating through DNA methylation inhibition, which triggers CD4 + T cell autoreactivity via LFA-1 overexpression (Chang and Gershwin, 2010; Richardson, 2019). These structural analog pairs demonstrate distinctive activity cliffs, characterized by sharp transitions in immunogenicity despite high structural similarity. Specifically, N-acetylprocainamide (procainamide's acetylated metabolite) fails to induce DIL flares in patients with prior procainamide-induced lupus, while phthalazine (lacking hydralazine's hydrazine side chain) shows diminished capacity to induce T-cell LFA-1 overexpression and autoreactivity in both in vitro and in vivo studies (Kluger et al., 1981; Yung et al., 1997).

Our model effectively captured these activity cliffs, assigning substantially higher risk probabilities to known DIL-inducing drugs (hydralazine: 85.4 %, procainamide: 52.0 %) compared to their structural counterparts (phthalazine: 4.6 %, N-acetylprocainamide: 25.1 %) (Fig. 5A-D).

The structural variations between these pairs critically determine their immunogenic potential through specific molecular features: (1) Reactive aromatic amine group: Both compounds contain aromatic amines (fr_ArN) that undergo oxidative metabolism via neutrophil activation and cytochrome P450-mediated Phase I hydroxylation, generating unstable reactive metabolites that readily form macromolecular adducts (Huang et al., 2024; Rubin, 2015, 2021). These

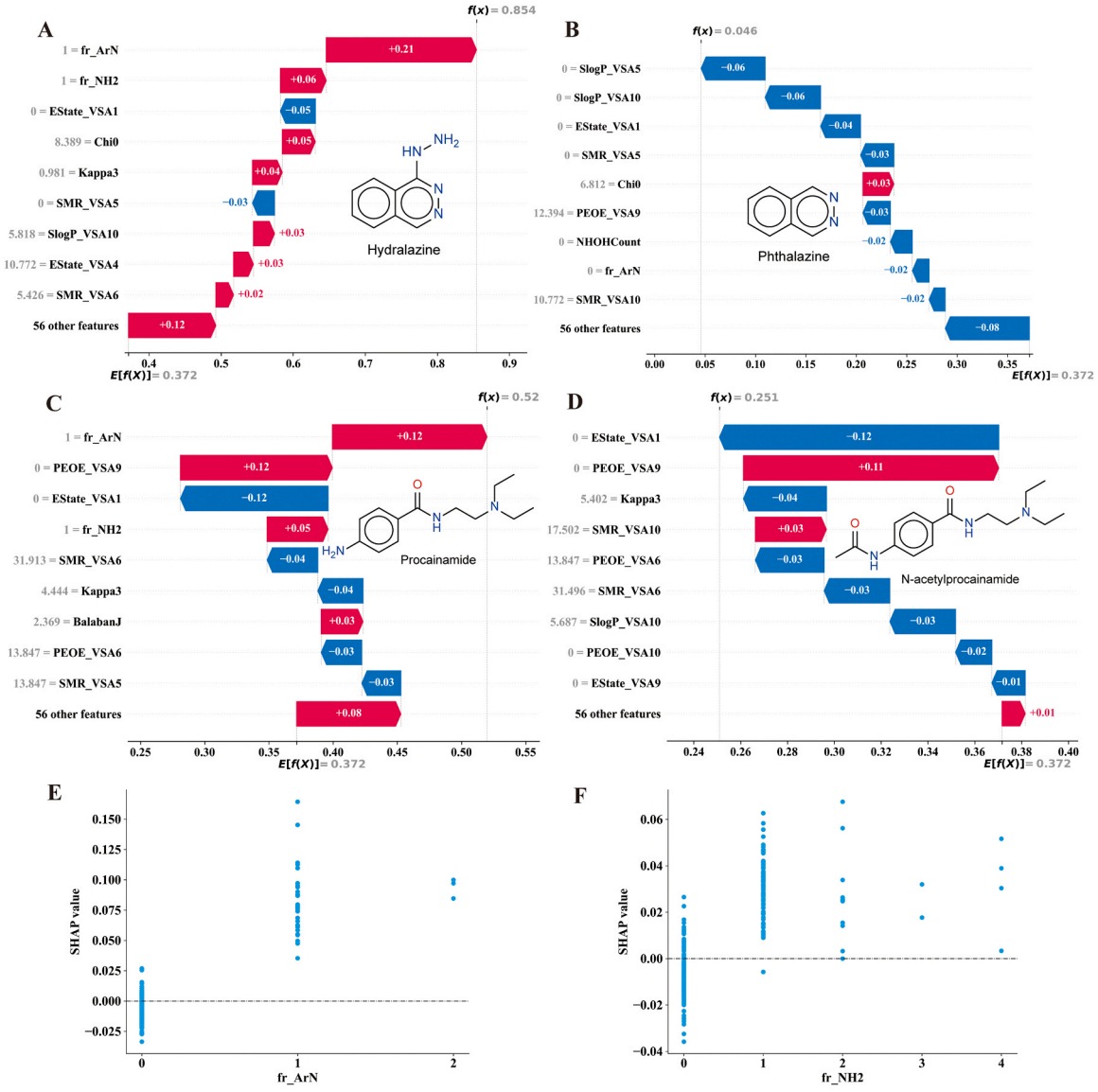


Fig. 5. SHAP analysis of structural features contributing to DIA prediction for drug pairs. (A-D) SHAP waterfall plots showing feature contributions for hydralazine (A), phthalazine (B), procainamide (C), and N-acetylprocainamide (D). Red and blue bars indicate positive and negative contributions to DIA risk, respectively. Chemical features are shown alongside their respective plots. (E-F) SHAP dependence plots for fr_ArN and fr_NH2, demonstrating the relationship between feature presence and their impact on model predictions. The x-axis represents feature values, and the y-axis shows corresponding SHAP values.

modifications can disrupt immune tolerance and initiate autoimmunity. N-acetylation blocks this oxidative pathway, explaining the protective effect of rapid acetylator phenotype against hydralazine- and procainamide-induced lupus (Rubin, 2015, 2021). (2) Surface chemistry: Primary amine groups (fr_NH2) form irreversible bonds with APC (Antigen-Presenting Cells) surface aldehydes, disrupting normal immune cell communication and triggering immune activation (Uetrecht, 2005).

These structure-based mechanisms underlying these activity cliffs are clearly captured by our SHAP analysis (Fig. 5). The SHAP dependence plots (Fig. 5E, F) reveal that both fr_ArN and fr_NH2 features (value ≥1) consistently correlate with high positive SHAP values, confirming their significant contribution to DIA toxicity prediction.

The hydralazine/phthalazine pair exemplifies a particularly steep activity cliff. Hydralazine's high prediction probability (85.4 %, Fig. 5A) is driven by strong positive contributions from both fr_ArN (+0.21) and fr_NH2 (+0.06), reflecting its immunogenic moieties. The removal of the hydrazine group in phthalazine (4.6 %, Fig. 5B) generates a dramatic activity cliff that not only eliminates these reactive features but also significantly alters key molecular properties (Table S6), including electronic state (EState_VSA9), partial charges (PEOE_VSA9), polarizability (SMR_VSA10) and lipophilicity (SlogP_VSA10), among others. These changes in physicochemical properties together reduce phthalazine's ability to undergo oxidative metabolism and interact with immune system components, resulting in significantly lower immunogenic potential.

Similarly, the procainamide/N-acetylprocainamide pair demonstrates another significant activity cliff, where N-acetylation of the primary aniline group fundamentally changes the compound's immunogenicity. Procainamide's moderate risk prediction (52.0 %, Fig. 5C) reflects contributions from both fr_ArN (+0.12) and fr_NH2 (+0.05), while N-acetylprocainamide's lower risk (25.1 %, Fig. 5D) results from the elimination of reactive amine groups, preventing oxidative metabolite formation.

These activity cliffs provide crucial insights into structure-immunogenicity relationships. The observed DIA probability differences (80.8 % between hydralazine/phthalazine and 26.9 % between procainamide/N-acetylprocainamide pairs) exemplify how subtle structural modifications can profoundly impact immunological responses. Notably, these activity cliffs arise from structural changes that simultaneously modulate multiple physicochemical properties critical for DIA development, creating local discontinuities in the structure-toxicity landscape. The concurrent alterations in electronic states, charge distribution, molecular polarizability, and lipophilicity demonstrate the complex interplay of molecular features in determining immunogenic potential. Our model's ability to accurately capture these sharp transitions validates its utility in analyzing structure-toxicity relationships. However, the presence of such activity cliffs also highlights fundamental challenges in computational toxicology, where minor structural modifications can lead to unexpectedly large changes in biological effects. Understanding and predicting these discontinuities remains essential for developing reliable computational approaches for DIA toxicity assessment in drug discovery.

3.5. Development of an interpretable online platform for DIA prediction

To enable practical application of our findings, we developed an open-access online platform (https://drug-induced-autoimmunity-predictor.streamlit.app/) implementing the optimized EEC model with RDKit_GA_65 feature subset. The complete source code for the web application, along with deployment instructions and all necessary dependencies, is available in our GitHub repository (https://github.com/Huangxiaojie2024/InterDIA). As demonstrated in Fig. 6, this platform facilitates rapid DIA risk assessment with transparent prediction interpretation.

The platform utilizes a streamlined two-step prediction process: (1) Users generate RDKit descriptors through the ChemDes platform (http://www.scbdd.com/rdk_desc/index/), and (2) Upload the resulting CSV file. The model automatically selects the 65 optimal descriptors identified in our study and performs DIA toxicity prediction. Notably, the platform supports batch processing for efficient large-scale compound screening.

A key innovation of our platform is the integration of SHAP waterfall plots for individual predictions (Fig. 6B), providing mechanistic insights into how specific molecular features influence the predicted DIA risk. This interpretability not only elucidates prediction rationale but also guides strategic structural modifications to minimize immunogenicity risk.

3.6. Comparison with previous studies for DIA prediction

To evaluate the capabilities of our proposed model in the context of existing approaches, we compared the performance metrics with two recent studies focused on DIA prediction (Guo et al., 2022; Wu et al., 2021). The first study by Wu et al. developed a CatBoost model combining structural alerts and daily dose information, while Guo et al. constructed a MACCS fingerprint-based SVM model. Table 7 presents a comprehensive comparison of model characteristics and performance metrics across these studies.

The dataset construction strategies exhibited notable differences among these studies. Wu et al. utilized a relatively constrained dataset comprising 50 DIA-positive and 357 DIA-negative drugs, employing a

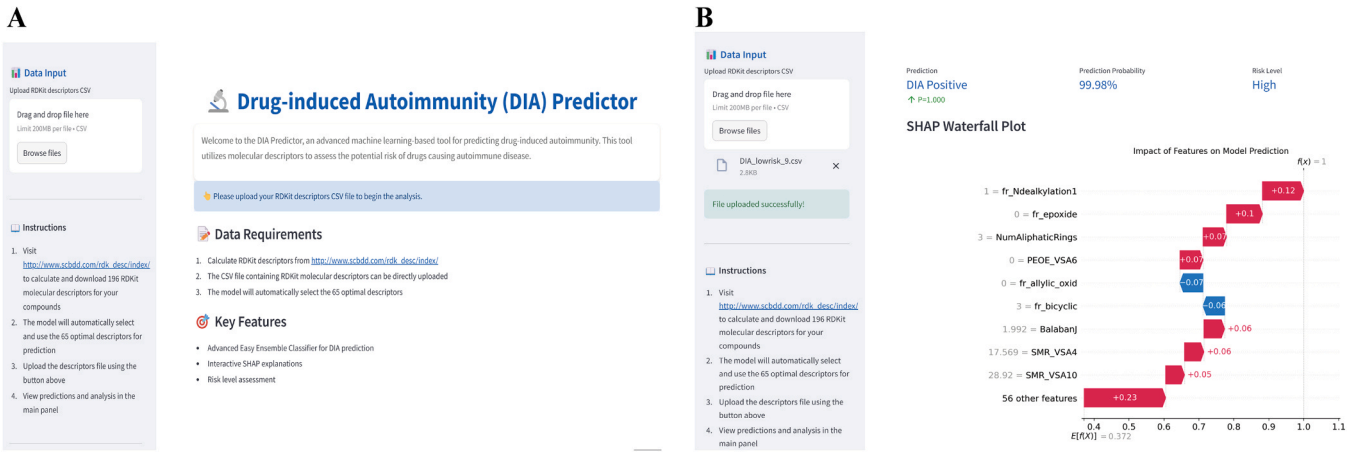


Fig. 6. Online platform for DIA toxicity prediction. (A) Platform homepage demonstrating essential features and input requirements. (B) Example prediction output displaying DIA risk probability and corresponding SHAP waterfall plot illustrating molecular feature contributions to the prediction.

Table 7
Comparison of model performances among different studies for DIA prediction.

Model name	Feature	No. of compounds	Test method	AUC	ACC	SEN	SPE	MCC
CatBoost (Wu et al., 2021)	Structural alerts and daily dose	325	5-fold cv	-	-	-	-	-
		82	External validation	0.70	90.24 %	40.00 %	97.22 %	0.47
MACCS_SVM (Guo et al., 2022)	MACCS fingerprint	240	5-fold cv	0.86	77.50 %	76.52 %	78.40 %	0.55
		358	External validation	0.84	76.26 %	75.76 %	76.31 %	0.33
		73	Overlapping subset[a]	-	75.34 %	75.00 %	75.38 %	0.3418
Proposed EEC Model	RDKit descriptor	477	10-fold cv	0.8836	82.81 %	82.20 %	83.01 %	0.5978
		120	External validation	0.8930	85.00 %	83.33 %	85.56 %	0.6413
		73	Overlapping subset[a]	0.9346	86.30 %	87.50 %	86.15 %	0.5562

[a] The overlapping subset contains 73 compounds (8 DIA-positive and 65 DIA-negative) present in both external validation sets, enabling direct head-to-head comparison. For Guo's model, TP = 6, FN = 2, TN = 49, FP = 16; For our proposed EEC model, TP = 7, FN = 1, TN = 56, FP = 9.

conventional 8:2 training-test partitioning. Guo et al. enhanced positive sample representation by incorporating SIDER database entries with documented autoimmune adverse reactions (incidence ≥0.1 %), resulting in 148 positive and 450 negative samples. However, their asymmetric splitting strategy (80:20 for positives, 25:75 for negatives) potentially introduced learning bias due to limited negative training samples during model construction. Our study addressed these limitations through systematic 8:2 partitioning combined with advanced ensemble resampling techniques, ensuring comprehensive learning from both positive and negative classes while maintaining dataset integrity.

In terms of model development strategy, distinct approaches were adopted by each study. Wu et al. primarily relied on structural alerts and daily dose information in their CatBoost model, which may not capture the full spectrum of molecular characteristics influencing DIA. The SVM model by Guo et al. utilized molecular fingerprints exclusively, providing detailed structural information but potentially overlooking critical physicochemical properties. Neither study conducted applicability domain analysis to define reliable prediction boundaries, nor provided mechanistic interpretations of their predictions. Our study implemented more sophisticated and comprehensive strategies. We first integrated RDKit descriptors and other molecular descriptors to comprehensively characterize both structural features and physicochemical properties. Through multi-strategy feature selection combining MI, ETB, RFECV, and GA, we identified the most informative molecular characteristics while maintaining model interpretability. Most importantly, we established a well-defined applicability domain to ensure reliable predictions and employed SHAP analysis to provide mechanistic insights into molecular determinants of DIA toxicity, aspects not addressed in previous studies.

Performance evaluation revealed distinct patterns across these models. The CatBoost model demonstrated high overall accuracy (90.24 %) and specificity (97.22 %) but significant limited sensitivity (40.00 %). This severe class imbalance issue was not properly addressed in their study, leading to a model heavily biased towards negative samples. Such imbalanced performance (sensitivity of only 40.00 % versus specificity of 97.22 %) raises concerns about the model's reliability and generalization capability, as it essentially fails to identify more than half of the DIA-positive compounds while showing a strong bias towards negative predictions. The SVM model achieved more balanced but moderate performance (AUC = 0.84, ACC = 76.26 %, SEN = 75.76 %, SPE = 76.31 %), likely constrained by the asymmetric dataset splitting strategy. In contrast, our EEC model exhibited superior and well-balanced performance across all metrics (AUC = 0.8930, ACC = 85.00 %, SEN = 83.33 %, SPE = 85.56 %), indicating robust prediction capability for both DIA-positive and DIA-negative compounds. To enable a fair and direct comparison between our model and Guo's approach, we identified an overlapping subset of 73 compounds (8 DIA-positive and 65 DIA-negative) that were present in both external validation sets. The detailed prediction results for these compounds are provided in Table S7. Since Guo's model is only available through an online predictor (http://diad.sapredictor.cn/) without source code, we compared performance metrics excluding AUC. Our EEC model significantly outperformed their approach across all metrics (ACC: 86.30 % vs. 75.34 %, SEN: 87.50 % vs. 75.00 %, SPE: 86.15 % vs. 75.38 %, MCC: 0.5562 vs. 0.3418). This substantial improvement, particularly the 11 % increase in accuracy and 0.21 increase in MCC, demonstrates our model's enhanced capability in DIA toxicity prediction.

The superior performance of our model primarily stems from several key innovations: the integration of Easy Ensemble Classifier to address data imbalance, ensuring balanced learning from both positive and negative classes; comprehensive molecular characterization and multi-strategy feature selection, which captures the full spectrum of molecular properties relevant to DIA development; and establishment of a well-defined applicability domain for reliable predictions, which ensures the model's predictions are only made within its validated chemical space. These advancements collectively enable more robust and reliable DIA toxicity prediction compared to previous approaches.

4. Strengths and limitations of this study

This study represents a significant advancement in drug-induced autoimmunity prediction through several key innovations. First, we developed an integrated approach that combines ensemble learning strategies with multi-strategy feature selection to effectively address data imbalance while capturing crucial molecular characteristics of DIA. This comprehensive methodology enables balanced learning from both positive and negative samples and reliable identification of DIA-related molecular features, resulting in superior model performance compared to previous studies (Guo et al., 2022; Wu et al., 2021). Second, we pioneered the application of explainable artificial intelligence (XAI) techniques in DIA prediction. While XAI techniques have been successfully applied to various fields including drug discovery and toxicity prediction (Harren et al., 2022; Togo et al., 2022; Wu et al., 2023), their potential in DIA prediction remained largely unexplored. Through SHAP analysis, we systematically identified critical molecular properties and provided mechanistic interpretations of their roles in DIA development, advancing our understanding of structure-toxicity relationships. Third, our model demonstrated strong discriminative capability in differentiating structurally similar compounds with distinct immunogenic potentials, as validated through paired case studies of hydralazine/phthalazine and procainamide/N-acetylprocainamide. The accurate prediction of their differential DIA risks supports the model's practical utility in drug development. Fourth, we developed an open-access web platform that enables batch predictions with real-time visualization of molecular feature contributions through SHAP waterfall plots. This platform provides not only prediction results but also mechanistic insights into how specific molecular features influence DIA prediction, facilitating rational drug design. Finally, to promote transparency and support further research in DIA toxicity prediction, we have made our dataset, model development codes, and web platform construction codes openly accessible via our GitHub repository (https://github.com/Huangxiaojie2024/InterDIA).

However, this study also has several limitations. The primary

limitation of this study lies in the relatively small number of DIA-positive samples, which were collected based on an incidence rate threshold of $\geq$ 0.1 %. When tested on an external set of 28 drugs associated with very low risk (<0.1 %) of drug-induced lupus (DIL), our model showed moderate performance with an accuracy of 64.28 % (data not shown). This differential performance can be explained from both mechanistic and computational perspectives. DIA-positive drugs with incidence rates $\geq$ 0.1 % likely share distinct and consistent molecular features that robustly trigger autoimmune responses, making these strong molecular signatures more readily identifiable by our machine learning model. In contrast, very low-incidence cases (<0.1 %) might involve more subtle and diverse molecular patterns, coupled with complex patient-specific factors beyond pure chemical properties. These characteristics pose significant challenges for our current molecular descriptor-based approach, as the model is inherently optimized to detect pronounced structural patterns that consistently appear in drugs with incidence rates $\geq$ 0.1 % rather than the more nuanced features associated with very low-incidence cases.

In future research, we plan to expand the dataset by incorporating more diverse DIA-positive compounds and develop specific prediction models for distinct types of autoimmune adverse reactions (such as drug-induced lupus, autoimmune hepatitis, and autoimmune hemolytic anemia). These improvements will enhance the model's ability to capture subtle molecular patterns associated with rare autoimmune reactions and provide more comprehensive predictions across the full spectrum of DIA toxicity.

5. Conclusion

In this study, we developed InterDIA, an interpretable machine learning framework for predicting drug-induced autoimmunity toxicity. Through systematic optimization of molecular descriptors and advanced ensemble learning strategies, our framework achieved robust predictive performance while effectively addressing the data imbalance challenge. Mechanistic interpretation revealed crucial molecular properties driving DIA development, including lipophilicity, partial charge distribution, electronic states, molecular polarizability, and topological features, providing valuable insights into structure-toxicity relationships. The model's practical utility was validated through successful discrimination between DIL high risk compounds and their less immunogenic structural analogs. To facilitate practical application, we integrated these findings into an accessible web platform that enables rapid DIA risk assessment with transparent mechanistic interpretation through SHAP analysis. This work not only advances our understanding of molecular mechanisms underlying DIA but also provides a valuable tool for early toxicity screening in drug development.

Funding

This research was supported by the Medical Science and Technology Research Foundation of Guangdong Province (Grant Number: A2024082).

CRediT authorship contribution statement

Huang Lina: Writing – review & editing, Visualization, Software, Resources, Methodology, Data curation. **Huang Xiaojie:** Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Liu Peineng:** Writing – review & editing, Validation, Software, Methodology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.tox.2025.154064.

Data availability

The original data presented in the study are included in the article/Supplementary Materials, and further inquiries can be directed to the corresponding author.

References

Balaban, A.T., 1982. Highly discriminating distance-based topological index. Chem. Phys. Lett. 89, 399–404.
Balaji, S., Karthikeyan, C., Moorthy, N.H.N., Trivedi, P., 2004. QSAR modelling of HIV-1 reverse transcriptase inhibition by benzoxazinones using a combination of P_VSA and pharmacophore feature descriptors. Bioorg. Med. Chem. Lett. 14, 6089–6094.
Banerjee, A., Roy, K., Gramatica, P., 2024. A bibliometric analysis of the Cheminformatics/QSAR literature (2000–2023) for predictive modeling in data science using the SCOPUS database. Mol. Divers. 1–13.
Bergstra, J., Yamins, D. and Cox, D.D. 2013. Hyperopt: a python library for optimizing the hyperparameters of machine learning algorithms. Proceedings of the 12th Python in Science Conference, Citeseer, p. 20.
Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PloS One 12, e0177678.
Chang, C., Gershwin, M.E., 2010. Drugs and autoimmunity–a contemporary review and mechanistic approach. J. Autoimmun. 34, J266–J275.
Chang, C., Gershwin, M.E., 2011. Drug-induced lupus erythematosus: incidence, management and prevention. Drug Saf. 34, 357–374.
Chen, C., Liaw, A. and Breiman, L. 2004. Using random forest to learn imbalanced data. University of California. Berkeley 110, 24.
Chen, T. and Guestrin, C. 2016. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
Chen, Y., Liu, C., Guo, G., Zhao, Y., Qian, C., Jiang, H., Shen, B., Wu, D., Cao, F., Sun, H., 2022. Machine-learning-guided reaction kinetics prediction towards solvent identification for chemical absorption of carbonyl sulfide. Chem. Eng. J. 444, 136662.
Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 21, 6.
Chipinda, I., Hettick, J.M., Siegel, P.D., 2011. Haptenation: chemical reactivity and protein binding. J. Allergy 2011, 839682.
Cho, T., Uetrecht, J., 2017. How reactive metabolites induce an immune response that sometimes leads to an idiosyncratic drug reaction. Chem. Res. Toxicol. 30, 295–314.
Co-operation, Of.E., Development, 2014. Guidance document on the validation of (quantitative) structure-activity relationship [(Q) SAR] models. OECD Publishing.
Darst, B.F., Malecki, K.C., Engelman, C.D., 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. 19, 65.
Dedeoglu, F., 2009. Drug-induced autoimmunity. Curr. Opin. Rheumatol. 21, 547–551.
Dimova, D., Stumpfe, D., Bajorath, Jr, 2014. Method for the evaluation of structure–activity relationship information associated with coordinated activity cliffs. J. Med. Chem. 57, 6553–6563.
Dong, J., Cao, D.-S., Miao, H.-Y., Liu, S., Deng, B.-C., Yun, Y.-H., Wang, N.-N., Lu, A.-P., Zeng, W.-B., Chen, A.F., 2015. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. J. Chemin. 7, 1–10.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Gasteiger, J., Marsili, M., 1980. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. Tetrahedron 36, 3219–3228.
Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognit. Lett. 31, 2225–2236.
Grattagliano, I., Bonfrate, L., Diogo, C.V., Wang, H.H., Wang, D.Q., Portincasa, P., 2009. Biochemical mechanisms in drug-induced liver injury: certainties and doubts. World J. Gastroenterol. WJG 15, 4865.
Guo, H., Zhang, P., Zhang, R., Hua, Y., Zhang, P., Cui, X., Huang, X., Li, X., 2022. Modeling and insights into the structural characteristics of drug-induced autoimmune diseases. Front. Immunol. 13, 1015409.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
Halimu, C., Kasem, A. and Newaz, S.H.S. 2019. Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification. Proceedings of the 3rd International Conference on Machine Learning and Soft Computing, pp. 1–6.
Harren, T., Matter, H., Hessler, G., Rarey, M., Grebner, C., 2022. Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence. J. Chem. Inf. Model. 62, 447–462.
He, Y., Sawalha, A.H., 2018. Drug-induced lupus erythematosus: an update on drugs and mechanisms. Curr. Opin. Rheumatol. 30, 490–497.

Hong, H., Xie, Q., Ge, W., Qian, F., Fang, H., Shi, L., Su, Z., Perkins, R., Tong, W., 2008. Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J. Chem. Inf. Model. 48, 1337–1344.
Huang, X., Xie, X., Huang, S., Wu, S., Huang, L., 2024. Predicting non-chemotherapy drug-induced agranulocytosis toxicity through ensemble machine learning approaches. Front. Pharmacol. 15, 1431941.
Kar, S., Roy, K., Leszczynski, J., 2018. Applicability domain: a step toward confident predictions and decidability for QSAR modeling. Methods Mol. Biol. 1800, 141–169.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.
Kluger, J., Drayer, D.E., Reidenberg, M.M., Lahita, R., 1981. Acetylprocainamide therapy in patients with previous procainamide-induced lupus syndrome. Ann. Intern. Med. 95, 18–23.
Kuhn, M., Letunic, I., Jensen, L.J., Bork, P., 2016. The SIDER database of drugs and side effects. Nucleic Acids Res 44, 1075–1079 (D).
Leardi, R., Boggia, R., Terrile, M., 1992. Genetic algorithms as a strategy for feature selection. J. Chemom. 6, 267–281.
Lemaître, G., Nogueira, F., Aridas, C.K., 2017. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. 18, 559–563.
Liu, X.Y., Wu, J., Zhou, Z.H., 2009. Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. B Cybern. 39, 539–550.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.
Maclin, R., Opitz, D., 1997. An empirical evaluation of bagging and boosting. AAAI/IAAI 1997, 546–551.
Muzet, N., Guillot, B., Jelsch, C., Howard, E., Lecomte, C., 2003. Electrostatic complementarity in an aldose reductase complex from ultra-high-resolution crystallography and first-principles calculations. Proc. Natl. Acad. Sci. 100, 8742–8747.
Richardson, B.C., 2019. Drug-Induced lupus erythematosus. Dubois' Lupus Erythematosus and Related Syndromes. Elsevier, pp. 377–388.
Rubin, R.L., 2015. Drug-induced lupus. Expert Opin. Drug Saf. 14, 361–378.
Rubin, R.L., 2021. Drug-induced lupus. Systemic Lupus Erythematosus, 535–547.
Sun, H., Xia, M., Austin, C.P., Huang, R., 2012. Paradigm shift in toxicity testing and modeling. AAPS J. 14, 473–480.
Szyper-Kravitz, M., Shoenfeld, Y., 2008. Drug-induced autoimmunity. Diagnostic criteria in autoimmune diseases, 59–63.
Togo, M.V., Mastrolorito, F., Ciriaco, F., Trisciuzzi, D., Tondo, A.R., Gambacorta, N., Bellantuono, L., Monaco, A., Leonetti, F., Bellotti, R., 2022. TIRESIA: an explainable artificial intelligence platform for predicting developmental toxicity. J. Chem. Inf. Model. 63, 56–66.
Uetrecht, J., 2005. Current trends in drug-induced autoimmunity. Autoimmun. Rev. 4, 309–314.
Van Tilborg, D., Alenicheva, A., Grisoni, F., 2022. Exposing the limitations of molecular machine learning with activity cliffs. J. Chem. Inf. Model. 62, 5938–5951.
Vedove, C.D., Del Giglio, M., Schena, D., Girolomoni, G., 2009. Drug-induced lupus erythematosus. Arch. Dermatol. Res. 301, 99–105.
Vergara, J.R., Estévez, P.A., 2014. A review of feature selection methods based on mutual information. Neural Comput. Appl. 24, 175–186.
Vo, A.H., Van Vleet, T.R., Gupta, R.R., Liguori, M.J., Rao, M.S., 2020. An overview of machine learning and big data for drug toxicity evaluation. Chem. Res Toxicol. 33, 20–37.
Wambaugh, J.F., Bare, J.C., Carignan, C.C., Dionisio, K.L., Dodson, R.E., Jolliet, O., Liu, X., Meyer, D.E., Newton, S.R., Phillips, K.A., 2019. New approach methodologies for exposure science. Curr. Opin. Toxicol. 15, 76–92.
Wu, Y., Zhu, J., Fu, P., Tong, W., Hong, H., Chen, M., 2021. Machine learning for predicting risk of drug-induced autoimmune diseases by structural alerts and daily dose. Int. J. Environ. Res. Public Health 18, 7139.
Wu, Z., Chen, J., Li, Y., Deng, Y., Zhao, H., Hsieh, C.-Y., Hou, T., 2023. From black boxes to actionable insights: a perspective on explainable artificial intelligence for scientific discovery. J. Chem. Inf. Model. 63, 7617–7627.
Xiao, X., Chang, C., 2014. Diagnosis and classification of drug-induced autoimmunity (DIA). J. Autoimmun. 48, 66–72.
Yang, H., Sun, L., Li, W., Liu, G., Tang, Y., 2018. In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. Front. Chem. 6, 30.
Yung, R., Chang, S., Hemati, N., Johnson, K., Richardson, B., 1997. Mechanisms of drug-induced lupus. IV. Comparison of procainamide and hydralazine with analogs in vitro and in vivo. Arthritis Rheum. Off. J. Am. Coll. Rheumatol. 40, 1436–1443.
Zhou, S., Chan, E., Duan, W., Huang, M., Chen, Y.-Z., 2005. Drug bioactivation covalent binding to target proteins and toxicity relevance. Drug Metab. Rev. 37, 41–213.
Zhu, J., Huang, Y., Yi, Q., Bu, L., Zhou, S., Shi, Z., 2024. Predicting reactivity dynamics of halogen species and trace organic contaminants using machine learning models. Chemosphere 346, 140659.