Simen Owen Academic Proceedings Series

4th International Conference on Electronics, Engineering, Computer Science and Applied Development (EESD 2026)

Article

Drug-Induced Liver Injury Prediction and Structural Risk Factor Identification Using Multi-Source Molecular Feature Fusion and Explainable Machine Learning

Haotian Zhang[1,*]

[1] School of Pharmacy, China Pharmaceutical University, Nanjing, China
* Correspondence: Haotian Zhang, School of Pharmacy, China Pharmaceutical University, Nanjing, China

Abstract: Drug-induced liver injury (DILI) is a critical issue in drug safety research, and its mechanism is rather complex, posing significant challenges for DILI risk prediction. Existing studies mostly employ single-molecule representation or fixed feature fusion methods, with less consideration given to the consistency among different molecular information, the reliability of prediction results, and the interpretability of the model. To address this issue, this study proposes RAFF-DILI, a multi-source learning framework, which comprehensively utilizes physical-chemical descriptors, Morgan fingerprints, and molecular graph representations to describe molecular structural features from different aspects. The model adopts sample adaptive gated fusion and combines methods such as cross-view consistency regularization, probability calibration, and consensus attribution to improve the stability and interpretability of DILI risk prediction. The FDA DILIst test results show that RAFF-DILI maintained excellent predictive performance in five independent runs. The AUROC, AUPRC, and MCC values were 0.892 ± 0.012, 0.918 ± 0.010, and 0.671 ± 0.021 respectively, where the AUROC being 0.019 higher than the strongest baseline model.In the DILIrank 2.0 structure repetition external evaluation, the model still achieved an AUROC of 0.841 ± 0.017, indicating its stable prediction performance across different datasets. Further analysis revealed that some physical-chemical features and substructure patterns repeatedly appeared in the prediction results of higher DILI risks. In conclusion, RAFF-DILI can combine molecular information from different sources, maintaining good prediction performance while enhancing the reliability and interpretability of the prediction results, and can provide computational methods for risk screening and priority ranking in drug safety evaluation.

Keywords: Drug-induced liver injury; Multi-source molecular feature fusion; Graph neural network; Explainable machine learning; Drug-safety prediction

Received: 05 September 2026
Revised: 22 October 2026
Accepted: 03 November 2026
Published: 07 November 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Drug-induced liver injury (DILI) is a significant adverse reaction to drugs and has been a long-term concern in drug development and post-marketing surveillance. The occurrence of DILI involves multiple factors. The generation of active metabolites, mitochondrial dysfunction, changes in bile acid homeostasis, immune response, drug exposure, and individual susceptibility all may contribute to liver injury [1]. Therefore, it is difficult to comprehensively reflect the clinical mechanism of DILI solely based on molecular structure. However, structural information still plays an important role in early computational screening. The physical and chemical properties of molecules and specific substructures can reflect their metabolic activity, lipophilicity, and molecular interactions.

Public resources such as DILIst provide standardized human DILI labels, providing a data foundation for conducting molecular-level machine learning research [2].

Computational DILI prediction has gradually evolved from descriptor-based and fingerprint methods to molecular graph learning. Studies have found that specific chemical fragments can help improve the performance of DILI classification and also provide a basis for structural analysis [3]. Graph neural networks can directly learn the connection relationships between atoms and chemical bonds in molecules, thereby obtaining molecular representations [4]. In recent years, studies have also incorporated biological graphs and toxicogenomic information, indicating that combining information from different sources is helpful for DILI prediction [5]. However, existing methods still have some limitations. The information contained in different molecular representations is not the same, and simply concatenating them makes it difficult to determine which information is more reliable. When different representations are learned separately, the results may also differ, and this problem becomes more obvious for compounds outside the training data. Moreover, the accuracy of model predictions does not necessarily mean that the structural interpretation is necessarily reliable. The attribution results obtained from different representations may vary greatly, making it difficult to directly use them as mechanism evidence.

In response to these issues, this study proposes a multi-source molecular feature fusion method for the prediction of DILI and structural risk analysis. The research mainly includes the following contents: (1) Combining physical-chemical descriptors, Morgan fingerprints, and molecular graph information, and adjusting the weights of different features according to the prediction effect on specific samples to avoid directly concatenating all features together. The prediction effect of this method is evaluated through comparative experiments and ablation experiments. (2) Adding consistency constraints in model training to reduce the differences between the prediction results obtained from different features, while retaining the effective information in various features. This is verified through ablation experiments, training process, probability calibration, and feature perturbation methods. (3) Comparing the influence of different features on the prediction results, and identifying potential risk structures that can be supported by multiple features. Then, using the structurally unique compounds from DILIrank 2.0 for external testing to examine the model's prediction effect on new compounds. DILIrank 2.0 is the updated human drug-induced liver injury (DILI) risk classification resource by the FDA [6].

The model first standardizes the molecules, then generates different types of molecular features, encodes and fuses them, classifies them, calibrates the prediction probabilities, and finally analyzes the structural features related to the prediction results. The dataset is divided using the scaffold-aware method and multiple repeated experiments are conducted to reduce information leakage caused by similar structures. The external test is used to further examine the model's prediction effect on unseen chemical structures.

This study focuses on whether the combination of different molecular features can improve the accuracy and stability of DILI prediction, and examines whether the structural explanations given by the model are reliable. The model's predictions are comprehensively judged by comparing the prediction probabilities, the results after feature perturbation, and the structural information obtained from different features. This method can be used for prioritization and structural risk screening in in vitro drug safety evaluation.

2. Related Works

2.1. Descriptor-Based and Interpretable Machine Learning

The early prediction of DILI mainly relied on physical-chemical descriptors, molecular fingerprints, and structural fragments to describe compounds. These methods have low computational requirements and are suitable for DILI datasets with only a few

hundred to several thousand labeled compounds. Previous studies have shown that combining physical-chemical properties, scaffold structures, molecular fingerprints, etc., helps to distinguish hepatotoxic drugs from non-hepatotoxic drugs [7]. Further research utilized feature importance and attention mechanisms to identify the features and local structures that had a significant impact on the prediction results, making the model's prediction basis clearer [8]. Additionally, the integration method that combines multiple descriptors can reduce reliance on a single algorithm and has achieved good prediction results on some public DILI datasets [9].

However, these methods are still limited by the way of feature representation. Molecular fingerprints mainly describe local structures and may miss some chemical environment information. Physical-chemical descriptors more reflect the overall properties of the molecule and are difficult to reflect the specific connection relationships between atoms and chemical bonds. Feature selection and combination can reduce redundant information, but they cannot compensate for the structural information that has been lost during the representation process. It should also be noted that what the model considers important as a feature only indicates its relevance to the prediction results, and it cannot be used to determine that it is the cause of hepatotoxicity.

2.2. Graph-Based and Biological Representation Learning

Graph neural networks provide another direction for the representation of molecular structures. It treats atoms as nodes and chemical bonds as edges, and then learns molecular features based on the connection relationships between them. Recent DILI research has added bond lengths and partial charges to this basis, and the results show that this more complete molecular representation has better classification effects than simple molecular graphs [10]. Some studies have also added genetic interactions and toxicological data to the model to enable DILI prediction to consider the biological changes caused by drugs [11]. Related studies have also attempted to combine molecular structures with disease-related networks, and the results show that structural and biological information can provide different references [12].

Graph neural networks and biological data can offer more information, but they also bring new limitations. Graph models have many parameters and are prone to overfitting on smaller DILI datasets. Toxicology and pathway data are not available for every compound, and processing such data requires additional data handling. For cases where only molecular structure information is available and new compounds need to be predicted, relying on these additional data further limits the applicability of the model.

2.3. Multi-Source Molecular Fusion and Explainability

In recent years, research has begun to integrate different types of molecular information to overcome the limitations of a single representation. For instance, when chemical embedding is combined with biological detection-response fingerprints, it shows better stability under the framework of scaffolds and clustering [13]. When structural fingerprints are combined with SMILES representations, they can simultaneously utilize local structure and molecular sequence information [14]. Additionally, some studies have combined molecular graphs, fingerprints, and drug-related knowledge graphs for DILI prediction and risk classification [15].

However, existing methods mainly focus on how to integrate these information, while paying less attention to the reliability of different representations for specific compounds. Different representations may yield different prediction results, especially for structures that are rare in the training data. At this point, even if the model's predictions are accurate, it is difficult to determine whether the features it relies on are stable.

2.4. Research Gap and Position of This Study

Descriptors, fingerprints, and graph models each have their advantages and disadvantages. Descriptors and fingerprints are simple to calculate but have limited descriptions of molecular structures. Graph models can utilize the relationships between

atoms and chemical bonds, but are more susceptible to data scale and new structures. Current multimodal methods can combine more information, but it is difficult to determine the reliability of different representations for specific compounds.

To address this issue, this study adaptively assigns weights based on the performance of different molecular representations on specific compounds and incorporates consistency constraints to reduce the prediction differences between different representations. At the same time, it calibrates the DILI prediction probabilities and compares the structural explanations given by different representations, retaining only the risk structures that are mutually supportive. Finally, it combines scaffold-aware data partitioning, feature perturbation, and structural deduplication for external validation to evaluate the model's prediction performance, stability, and structural interpretation.

3. Methodology

3.1. Overall Architecture

This study develops RAFF-DILI, a multi-source molecular learning framework for reliability prediction, integrating molecular representation, reliability estimation, prediction, and structural interpretation into a single process. As shown in Figure 1, the input SMILES string is standardized and then generates physical-chemical descriptors, 2048-bit Morgan fingerprints with a radius of 2, and atomic-bond molecular graphs. Descriptors and fingerprints are encoded by multi-layer perceptrons, and molecular graphs are encoded by a three-layer graph isomorphism network (GIN). The three representations are ultimately mapped to a unified 128-dimensional latent space.

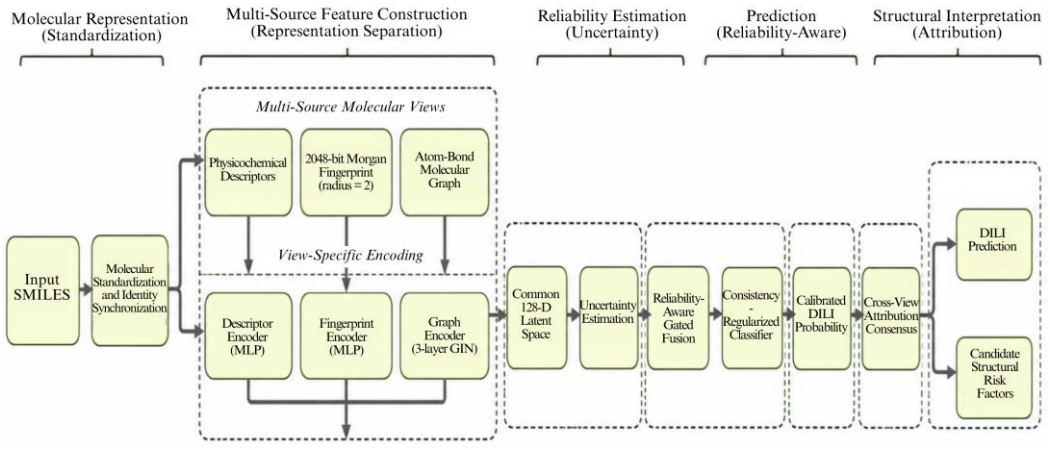


Figure 1. Overall Architecture of Raff-dili.

RAFF-DILI dynamically adjusts the contribution of different representations based on the uncertainty of specific compounds, and reduces the prediction differences between different representations through cross-view consistency constraints. Subsequently, the calibration layer adjusts the classifier output to obtain the DILI prediction probability. Finally, the attribution results of descriptors, fingerprints, and molecular graphs are aligned, and structural factors related to a higher DILI risk are selected.

3.2. Molecular Standardization and Data Synchronization

Each record is matched with a standard SMILES and InChIKey to ensure consistency in the identity of the compounds. For salts and isolated chemical components, only the largest effective chemical fragment is retained, and the formal charge is normalized. Subsequently, compounds that are completely identical are removed, and compounds with contradictory DILI labels are excluded. The DILI classification labels are encoded in a binary manner. DILI positive compounds are denoted as $y_i = 1$, and DILI negative

compounds as $y_i = 0$. The compounds in DILIrank 2.0 do not participate in the optimization of model parameters.

Continuous descriptors are standardized using the statistics of the training set:

$$\tilde{x}ij = \frac{xij-\mu_j}{\sigma_j+\epsilon}, \quad (1)$$

Among them, x_{ij} represents the j -th descriptor of compound i , and μ_j and σ_j respectively denote the mean and standard deviation of this descriptor in the training set. ϵ is a constant set to prevent numerical instability. The validation set and test set are both standardized using the statistics of the training set to avoid information leakage.

3.3. Multi-Source Molecular Representation

The descriptor vector x_i^d consists of physicochemical descriptors, mainly including molecular weight, lipophilicity, topological polar surface area, hydrogen bond number, aromaticity, ring structure statistics, and some non-redundant descriptors. The fingerprint vector (x_i^f is a 2048-bit Morgan fingerprint with a radius of 2, used to describe the local structure. The two types of features are encoded independently by separate multi-layer perceptrons and are not directly concatenated:

$$h_i^m = \mathrm{LN}\,![W_{m2}\,\mathrm{GELU}(W_{m1}x_i^m + b_{m1}) + b_{m2}], \quad m \in \mathrm{d,f,} (2)$$

After encoding, both types of molecular features are mapped to 128-dimensional representations, denoted as $h_i^m \in \mathbb{R}^{128}$. Here, the descriptor branch has 256 hidden units, and the Morgan fingerprint branch has 512 hidden units. During the encoding process, GELU is used to introduce nonlinearity, and LN is used to stabilize the feature distribution. W and b in the network are obtained through training.

The molecular graph is denoted as (G_i=(V_i,E_i)), where the nodes represent atoms and the edges represent chemical bonds between atoms. Atoms and chemical bonds are described by corresponding structural information. Based on this, GIN updates the node representation layer by layer by aggregating the information of adjacent atoms. The update process of the (l)th layer is:

$$h_v^{(l+1)} = \mathrm{MLP}^{(l)}\left[(1+\varepsilon_l)h_v^{(l)} + \sum_{u\in\mathcal{N}(v)} \phi^{(l)}\left(h_u^{(l)}, e_{uv}\right)\right]. (3)$$

Among them, $h_v^{(l)}$ represents the node (v) at the (l)th layer, $\mathcal{N}(v)$ are its neighboring nodes, and ε_l are learnable parameters. After three layers of GIN encoding, the node representations are pooled by mean and maximum values, resulting in a 128-dimensional molecular graph representation $h_i^g \in \mathbb{R}^{128}$.

3.4. Reliability Estimation and Control-Parameter Updating

Through $K = 10$ random discard iterations, different view prediction results were obtained, and the reliability of the representation was estimated based on the prediction fluctuations. RAFF-DILI estimated the uncertainty for each compound separately, without considering it as a fixed attribute of the entire mode. To reduce the weight fluctuations caused by mini-batch training, exponential weighted moving average was used to smooth the uncertainty, and the gating temperature was gradually reduced with the number of optimization steps:

$$\hat{u}_{i,m} = \frac{1}{K-1}\sum_{k=1}^{K}\left(p_{i,m}^{(k)} - \bar{p}_{i,m}\right)^2, \quad u_{i,m}^{(t)} = \rho u_{i,m}^{(t-1)} + (1-\rho)\hat{u}_{i,m}^{(t)}, \quad \tau_t = \max(\tau_{\min}, \tau_0\beta^t)$$
(4)

Among them, $p_{i,m}^{(k)}$ represents the k -th random prediction of view m , $\bar{p}_{i,m}$ is its mean, ρ is the smoothing factor, and τ_t is the gated temperature that varies with the optimization step number t . A higher temperature helps maintain balanced fusion in the early training stage, and as the temperature decreases, it gradually highlights the views with higher reliability. The reliability weight of view m is:

$$\alpha_{i,m} = \frac{\exp[(s_{i,m}-\gamma u_{i,m})/\tau_t]}{\sum_{r\in\{d,f,g\}}\exp[(s_{i,r}-\gamma u_{i,r})/\tau_t]} \quad (5)$$

The view weights are also influenced by both compatibility and uncertainty. $s_{i,m}$ represents the compatibility score of view m , and γ controls the extent of the influence of uncertainty. For views with large prediction fluctuations, the model will reduce their final weights.

3.5. Reliability-Aware Feature Fusion and Prediction

The potential representations of each view are first mapped to a unified feature space, then fused based on the reliability weights, and the DILI prediction probability is obtained:

$$z_i = \sum_{m\in d,f,g} \alpha_{i,m} P_m h_i^m, \quad p_i = \sigma(W_c z_i + b_c), \tag{6}$$

Among them, P_m represents the projection matrix of view, z_i represents the fused representation, W_c and b_c are classifier parameters, and $\sigma(\cdot)$ represents the Sigmoid function. Unlike direct concatenation, this fusion method dynamically adjusts the contribution of each view based on the reliability of the view of different compounds, and suppresses the representations with higher uncertainty.

3.6. Cross-View Consistency-Regularized Optimization

To avoid significantly inconsistent predictions of the same compound by different views, RAFF-DILI adds cross-view consistency constraints while optimizing the classification loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1-y_i)\log(1-p_i)] + \lambda_c \frac{1}{3N}\sum_{i=1}^{N}\sum_{\substack{m<r \\ m,r\in\{d,f,g\}}}[1 - \cos(P_m h_i^m, P_r h_i^r)] + \lambda_r \|\Theta\|_2^2 \tag{7}$$

Among them, N represents the batch size, λ_c and λ_r respectively control the intensity of consistency constraints and weight regularization, and Θ represents all trainable parameters. The cosine constraint does not require the representations of different views to be completely identical; instead, it focuses on restricting significant deviations in directions, thereby reducing unreasonable inconsistencies while maintaining the differences in features of each view. The model is optimized using AdamW, with the initial learning rate set to 10^{-3}, a batch size of 64, and is trained for a maximum of 150 cycles, with early stopping based on the AUROC of the validation set.

3.7. Probability Calibration and Attribution Consensus

The temperature parameter $T > 0$ is only fitted on the validation set. After the probability calibration is completed, an integrated gradient analysis is conducted on the three molecular representations to identify the main features that the model predicts. The attribution results of the fingerprint branch are further mapped to specific cyclic atomic environments. The graph structure branch integrates the attribution results by considering the connection relationships between atoms. The calibrated probability and consensus score of the candidate factor k are calculated as follows:

$$p_i^{\text{cal}} = \sigma\left(\frac{\text{logit}(p_i)}{T}\right), \quad R_k = \sum_{m\in\{d,f,g\}} \omega_m A_{k,m}, \quad \sum_{m\in\{d,f,g\}} \omega_m = 1 \tag{8}$$

Among them, $A_{k,m}$ represents the attribution value of factor k under view m, and ω_m represents the adjusted interpretation weight based on the reliability of the view. To enhance the stability of the results, the candidate factors need to show a positive risk contribution at least four times in five independent runs, and receive support from at least two types of representations. Only when the conditions are met will they be retained. The finally selected factors are used to characterize the candidate structural risk factors.

3.8. Optimization and Inference Procedure

To clearly illustrate the connection between each step, the complete training and reasoning process of RAFF-DILI is summarized in Algorithm 1.

Algorithm 1. RAFF-DILI Training, Calibration, and Structural Attribution

Input: SMILES set S, labels Y, train/validation partitions, $\lambda_c, \lambda_r, \gamma, \rho, \tau_0, \tau_{\min}$

Output: calibrated probabilities P^{cal}, predicted labels $\hat{Y}$, ranked candidate structural factors R

1: Standardize SMILES and generate canonical molecular identities.

2: Remove duplicates, label conflicts, salts, and invalid structures.

3: Construct descriptor, Morgan-fingerprint, and molecular-graph views.

4: Fit descriptor normalization parameters using training compounds only.

5: Encode descriptor and fingerprint views with view-specific MLPs.

6: Encode molecular graphs using the three-layer GIN.

7: Perform K dropout passes and estimate view-specific uncertainty.

8: Update smoothed uncertainty and gate temperature parameters.

9: Compute reliability-aware weights for the three molecular views.

10: Fuse projected representations and obtain raw DILI probabilities.

11: Calculate classification, consistency, and regularization losses.

12: Update network parameters with AdamW and apply early stopping.

13: Fit temperature T on validation predictions and calibrate probabilities.

14: Compute multi-view attributions and map them to molecular factors.

15: Return calibrated predictions and consensus-ranked structural risk factors.

4. Results and Analysis

4.1. Experimental Setup

The experimental environment is Ubuntu 22.04, equipped with an Intel Core i9-13900K processor, 64 GB of memory, and NVIDIA RTX 4090 (with 24 GB of video memory). Raff-Dili is built based on Python 3.11, PyTorch, and RDKit. The model training uses AdamW, with an initial learning rate of 1×10^{-3} , a batch size of 64, a dropout ratio of 0.20, and a maximum training period of 150 cycles. The consistency coefficient λ_c , uncertainty penalty coefficient γ , and smoothing coefficient ρ are set to 0.15, 0.8, and 0.9 respectively. The gating temperature during training is gradually reduced from 1.5 to 0.5. Training is stopped when the AUROC of the validation set does not improve for 15 consecutive cycles.

The experiment mainly used FDA DILIst [2], which contains 1,279 compounds, of which 768 are DILI positive and 511 are DILI negative. The external evaluation used DILIrank 2.0 [6], which contains 1,336 FDA-approved drugs and is divided into Most, Less, No, and Ambiguous-DILI-concern categories. Both datasets are from publicly available FDA resources. The external binary classification evaluation only uses compounds from the Most and No-DILI-concern categories, and compounds overlapping with DILIst are removed based on Canonical SMILES and InChIKey. The Less and Ambiguous-DILI-concern categories do not participate in the binary classification evaluation.

The preprocessing stage first standardizes the compounds, including removing salts, retaining the largest fragments, normalizing charges, and removing duplicate compounds, and checking the validity of the molecular structure. The statistics required for descriptor standardization are calculated only from the training set. To reduce the possibility of structural leakage, DiliSt is divided into 70/15/15 according to the Bemis–Murcko scaffold. The model is run independently 5 times, each time using a different random seed. The experiment selects RF-Morgan, XGBoost-Descriptor, GIN, and ChemBERTa as baselines, and evaluates the model performance using AUROC, AUPRC, and Matthews correlation coefficient (MCC). The results are presented as mean ± standard deviation. A two-sided paired test is used to compare different methods, with ($p<0.05$) considered statistically significant.

4.2. Comparative Performance

As shown in Table 1, RAFF-DILI achieved the highest average values in all three evaluation metrics. Its AUROC was 0.892 ± 0.012, which was higher than that of the strongest baseline ChemBERTa (0.873 ± 0.013) ($p = 0.018$). The AUPRC increased from 0.909 ± 0.011 to 0.918 ± 0.010, and the MCC increased from 0.637 ± 0.020 to 0.671 ± 0.021.

Table 1. Predictive Performance on the DILIst Scaffold Test Set (Mean ± SD, N = 5)

Model	AUROC	AUPRC	MCC	p, AUROC vs. RAFF-DILI
RF-Morgan	0.846 ± 0.018	0.883 ± 0.015	0.596 ± 0.028	0.003
XGBoost-Descriptor	0.858 ± 0.016	0.895 ± 0.013	0.612 ± 0.024	0.006
GIN	0.866 ± 0.014	0.903 ± 0.012	0.625 ± 0.022	0.011
ChemBERTa	0.873 ± 0.013	0.909 ± 0.011	0.637 ± 0.020	0.018
RAFF-DILI	0.892 ± 0.012	0.918 ± 0.010	0.671 ± 0.021	-

The improvement in MCC of RAFF-DILI was greater than that in AUROC, indicating that the model's improvement was not only reflected in sample ranking. RF-Morgan and XGBoost-Descriptor also achieved good results, suggesting that traditional chemical representations still have certain predictive capabilities under the current data scale. RAFF-DILI adjusted its weights based on the reliability of different molecular representations, while combining physical-chemical properties, local structure, and molecular graph information, and reducing the influence of unreliable representations on the prediction results.

4.3. Ablation and Mechanism Verification

As shown in Table 2, when any of the main modules were removed, the model performance decreased. After removing the reliability gating, the AUROC decreased from 0.892 ± 0.012 to 0.879 ± 0.015 ((p = 0.031)), indicating that the weights of each view need to be adjusted according to the actual situation for different compounds. After removing the consistency regularization, the AUROC dropped to 0.883 ± 0.014 ((p = 0.044)), and at the same time, the differences between different run results increased, suggesting that consistency regularization can make the model's predictions more stable.

Table 2. Ablation Results on the DILIst Scaffold Test Set (Mean ± SD, N = 5)

Configuration	AUROC	AUPRC	MCC	p, AUROC vs. full
Full RAFF-DILI	0.892 ± 0.012	0.918 ± 0.010	0.671 ± 0.021	-
w/o reliability gate	0.879 ± 0.015	0.907 ± 0.013	0.646 ± 0.025	0.031
w/o consistency loss	0.883 ± 0.014	0.910 ± 0.012	0.654 ± 0.024	0.044
w/o graph view	0.872 ± 0.017	0.901 ± 0.015	0.631 ± 0.027	0.012
w/o fingerprint view	0.868 ± 0.018	0.898 ± 0.016	0.624 ± 0.030	0.009
w/o descriptor view	0.881 ± 0.015	0.909 ± 0.013	0.649 ± 0.026	0.037

Among the three molecular views, the performance of the model deteriorated most significantly after removing the fingerprint branches, with the AUROC dropping to 0.868 ± 0.018. After removing the graph structure branches, the AUROC was 0.872 ± 0.017; and after removing the descriptor branches, it was 0.881 ± 0.015. Thus, the fingerprints and graph structure are relatively important for the model's prediction, while the descriptors can provide certain supplementary information.

From the structural interpretation perspective, the results obtained from the three molecular representations are relatively close to the conclusions in the ablation experiments. The model paid multiple attention to molecules with higher lipophilicity, more aromatic rings or heterocycles, and also noticed compounds containing furan, thiophene, hydrazine and aniline structures. It should be noted that these structures only reflect the molecular features that the model pays attention to during prediction.

To determine whether the probabilities given by the model are consistent with the actual results, the predicted probabilities were calibrated. Before calibration, the expected calibration error was 0.079 ± 0.009; after using temperature scaling, the error decreased to 0.041 ± 0.006 ((p = 0.006)). This indicates that after calibration, the risk probabilities given by the model are closer to the actual results.

4.4. Convergence Analysis

Figure 2 presents the results of five independent training sessions. At the beginning of the training, the loss was approximately 0.691, which gradually decreased and

eventually stabilized at around 0.284. The validation AUROC improved from 0.741 to 0.892. After about 40 epochs of training, the results of the five runs were relatively close, and the validation AUROC did not continue to increase significantly thereafter.

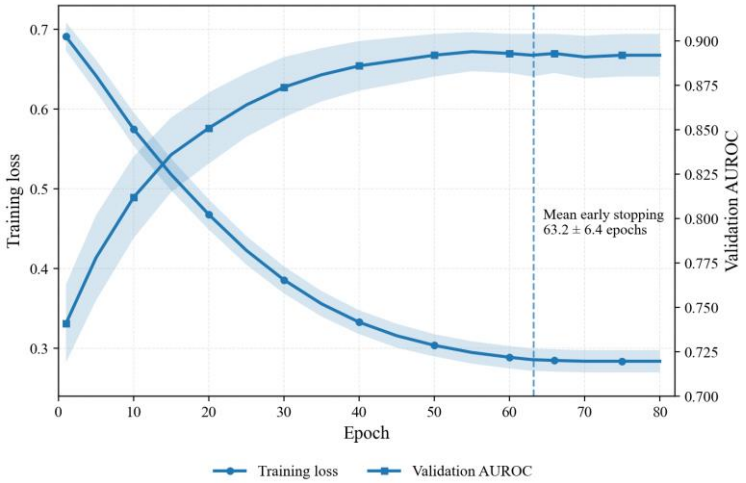


Figure 2. Training Loss and Validation Auroc of Raff-dili Across Five Independent Runs (Mean ± SD, N = 5).

The highest average value was 0.894 ± 0.011, and the model finally stopped training at approximately 63.2 ± 6.4 epochs. It can be seen that further training in the later stage mainly leads to a decrease in training loss, and the help to the performance of the validation set has become very small. Therefore, early stopping was adopted.

4.5. Robustness, Generalization, and Reliability

Figure 3 examines the model's tolerance to perturbations in input features. When the fingerprint was randomly masked by 5%, the AUROC decreased only from 0.892 ± 0.012 to 0.885 ± 0.013, with no significant difference ((p = 0.184)). When the mask was increased to 10%, the AUROC dropped to 0.869 ± 0.016 ((p = 0.022)). After perturbing the descriptors, the AUROC was 0.876 ± 0.015 ((p = 0.041)); after masking some chemical bonds in the diagram, it was 0.873 ± 0.018 ((p = 0.028)).

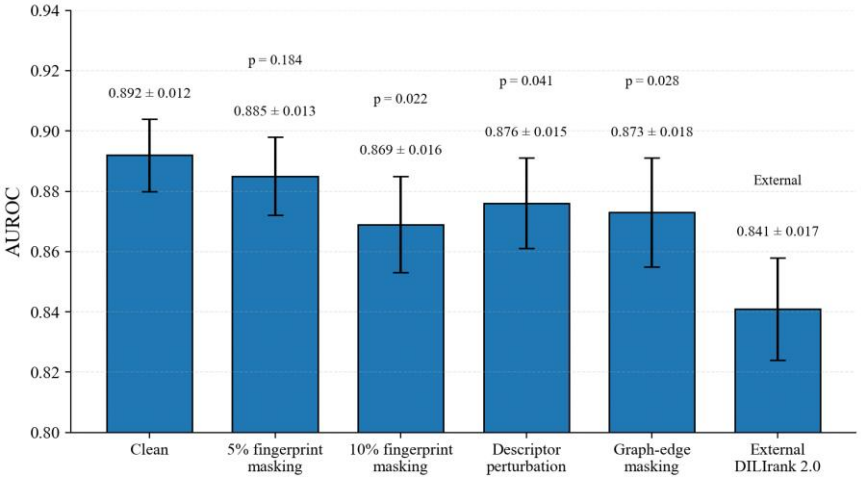


Figure 3. Robustness and Cross-Dataset Generalization of Raff-dili under Feature Perturbations and External DILIrank 2.0 Evaluation (Mean ± SD, N = 5).

From these results, it can be seen that a slight change in a certain input feature does not immediately lead to model failure. Other molecular representations still provide some information, but when the fingerprint is subjected to stronger interference, the performance degradation becomes more obvious. This indicates that the model does not

rely solely on a certain molecular representation, but can utilize the information provided by different views.

On the DILIrank 2.0 dataset after structural duplication, the AUROC of RAFF-DILI was 0.841 ± 0.017. This result was lower than the internal test set, but still maintained a good discrimination ability. Since the compounds in the external data do not have a one-to-one correspondence with the DILIst test set, the AUROCs of the two groups were not subjected to paired significance tests.

The performance improvement of RAFF-DILI mainly stems from the complementarity among different molecular representations, as well as the model's ability to adjust the fusion weights based on different compounds. The results of various experiments also show that the model performs relatively stably in terms of structural changes and external data, and the predicted probabilities it outputs have good reference value. These results provide support for the application of RAFF-DILI in drug safety screening.

5. Conclusion

In the tests based on the molecular framework, RAFF-DILI utilized both physical-chemical descriptors, molecular fingerprints, and molecular graph information, and adjusted the weights of each view according to the predicted stability of different compounds. The AUROC, AUPRC, and MCC of the model reached 0.892 ± 0.012, 0.918 ± 0.010, and 0.671 ± 0.021 respectively, with the AUROC being 0.019 higher than the best baseline model. After removing the reliability gating, the AUROC dropped to 0.879 ± 0.015, indicating that different molecular representations are not suitable for using the same weights for all compounds.

Consistency regularization had little impact on AUROC. After removing this item, the AUROC decreased from 0.892 ± 0.012 to 0.883 ± 0.014, but the result fluctuations increased between five runs. Therefore, this item mainly reduces the differences between different runs rather than directly improving the prediction indicators. After probability calibration, the expected calibration error decreased from 0.079 ± 0.009 to 0.041 ± 0.006, and the risk probabilities output by the model were also closer to the actual results.

Regardless of the different molecules, the structural features that the model focuses on show certain similarities, mainly concentrating on the hydrophobicity of the molecules as well as some cyclic and local structures. These results can help understand the basis of the model's predictions, but currently they can only serve as potential risk indicators and cannot indicate that these structures will directly cause DILI. In perturbation experiments, when the degree of change in input features is further increased, the performance drops significantly. On the DILIrank 2.0 dataset with duplicate structures removed, the model's AUROC was 0.841 ± 0.017, which was lower than the results of the internal test, and the prediction performance on new compounds decreased.

Currently, the model mainly uses molecular-level information, while DILI is also related to factors such as dose, metabolism, transporter activity, and individual differences. This information is not fully covered in the existing data. Multi-branch encoding also increases the computational cost of the model. In the future, drug exposure, metabolism, and toxicogenomic information can be further incorporated, and verified in a larger independent chemical space and actual drug screening data.

References

1. R. J. Fontana, I. Liou, A. Reuben, A. Suzuki, M. I. Fiel, W. Lee, and V. Navarro, "AASLD practice guidance on drug, herbal, and dietary supplement–induced liver injury," *Hepatology*, vol. 77, no. 3, p. 1036, 2023.
2. S. Thakkar, T. Li, Z. Liu, L. Wu, R. Roberts, and W. Tong, "Drug-induced liver injury severity and toxicity (DILIst): binary classification of 1279 drugs by human hepatotoxicity," *Drug Discovery Today*, vol. 25, no. 1, pp. 201–208, 2020.
3. S. Lim, Y. Kim, J. Gu, S. Lee, W. Shin, and S. Kim, "Supervised chemical graph mining improves drug-induced liver injury prediction," *iScience*, vol. 26, no. 1, 2023.

4. W. Wu, J. Qian, C. Liang, J. Yang, G. Ge, Q. Zhou, and X. Guan, "GeoDILI: a robust and interpretable model for drug-induced liver injury prediction using graph neural network-based molecular geometric representation," *Chemical Research in Toxicology*, vol. 36, no. 11, pp. 1717–1730, 2023.
5. T. Xiao, Y. Liu, K. Hu, K. Guo, M. Zhang, T. Wang, W. Lei, W. Wang, S. Zhou, Y. Hu, and R. Su, "Drug-induced liver injury prediction based on graph convolutional networks and toxicogenomics," *PLoS Computational Biology*, vol. 21, no. 9, p. e1013423, 2025.
6. A. O. Olubamiwa, Y. Qu, S. Connor, W. Tong, D. Li, and M. Chen, "DILIrank 2.0: An updated and expanded database for drug-induced liver injury risk based on FDA labeling and a literature review," *Drug Discovery Today*, p. 104485, 2025.
7. Y. Zhao, Z. Zhang, X. Kong, K. Wang, Y. Wang, J. Jia, H. Li, and S. Tian, "Prediction of drug-induced liver injury: From molecular physicochemical properties and scaffold architectures to machine learning approaches," *Chemical Biology & Drug Design*, vol. 104, no. 2, p. e14607, 2024.
8. S. Lee and S. Yoo, "InterDILI: interpretable prediction of drug-induced liver injury through permutation feature importance and attention mechanism," *Journal of Cheminformatics*, vol. 16, no. 1, p. 1, 2024.
9. J. Guan, D. Dong, P. Xie, Z. Zhao, Y. Guo, T.-Y. Lee, L. Yao, and Y.-C. Chiang, "StackDILI: Enhancing drug-induced liver injury prediction through stacking strategy with effective molecular representations," *Journal of Chemical Information and Modeling*, vol. 65, no. 2, pp. 1027–1039, 2025.
10. T. Lee and J. M. Posma, "Improving drug-induced liver injury prediction using graph neural networks with augmented graph features from molecular optimisation," *Journal of Cheminformatics*, vol. 17, no. 1, p. 124, 2025.
11. J. Wang, L. Zhang, J. Sun, X. Yang, W. Wu, W. Chen, and Q. Zhao, "Predicting drug-induced liver injury using graph attention mechanism and molecular fingerprints," *Methods*, vol. 221, pp. 18–26, 2024.
12. H. Zhong, J. Wang, X. Liu, X. Wei, C. Zhou, T. Zou, X. Han, L. Mo, W. Qin, and Y. Zhang, "A multi-dimensional computational framework of drug-induced hepatotoxicity: Integrating molecular structure features with disease pathogenesis," *Briefings in Bioinformatics*, vol. 26, no. 5, p. bbaf456, 2025.
13. Y. Shou, Y. Fang, Y. Ma, Y. Chi, J. Yao, J. Xu, Y. Deng, B. Peng, and M. Fang, "ChemBioHepatox: Multimodal integrating chemical structure and biological fingerprint for robust and interpretable hepatotoxicity prediction," *Environmental Science & Technology*, vol. 59, no. 47, pp. 25200–25213, 2025.
14. T. Liyaqat, T. Ahmad, and M. K. Jamal, "A multimodal deep learning method for predicting drug-induced liver injury using structural and sequential molecular representations," *Journal of Computer-Aided Molecular Design*, vol. 40, no. 1, p. 86, 2026.
15. Y. Liu, K. Hu, J. Geng, Q. Dai, L. Wei, and R. Su, "Integrating multi-type features and knowledge graph for graded prediction of drug-induced liver injury in humans," *PLoS Computational Biology*, vol. 22, no. 7, p. e1013640, 2026.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Publisher and/or the editor(s). Publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.