Enhancing Drug-Induced Liver Injury Prediction via Multi-Representation Molecular Images with Grad-CAM Explainability and Functional Group Attribution

Hoang Phi Yen Duong, *Graduate Student Member, IEEE*, Nghia Trong Vo, *Student Member, IEEE*, Tuan Thanh Nguyen, *Member, IEEE*, and Trung Q. Duong, *Fellow, IEEE*

***Abstract*—Drug-induced liver injury (DILI) constitutes a critical challenge in pharmaceutical development and accounts for over 50% of acute liver failure cases. This study employs deep learning (DL) methodologies for DILI prediction utilizing convolutional neural networks (CNNs) and multi-representation molecular imaging. We employ ResNet (18, 34, 50) and EfficientNet (B0-B3) architectures on three molecular image representations, such as normal images (NM), heatmap 1 (HM1) derived from Crippen logP contributions, and heatmap 2 (HM2) extracted from bioconcentration factor analysis. Experimental outcomes on 475 compounds using 4-fold cross-validation demonstrate that heatmap representations substantially outperform conventional molecular images. ResNet-34 attains optimal performance with HM2, achieving an AUC of 0.8853 and a Recall of 0.8410. This significantly enhances performance compared to conventional images (AUC: 0.8604, Recall: 0.8295). ResNet-34 and EfficientNet-B0 with HM1 exhibits superior functional group identification capabilities with a Recall of 0.8499 and 0.8682. Grad-CAM visualization illustrates that HM1 effectively emphasizes specific functional groups while HM2 facilitates comprehensive molecular analysis. Functional group analysis indicates that nitro groups (93.75%), sulfones (92.31%), and urea derivatives (91.67%) are associated with the highest DILI risk, while sulfur-containing moieties broadly serve as toxicity indicators. In contrast, iodine-containing compounds and ketones exhibit notably lower toxicity rates. Our multi-representation methodology exhibits competitive performance while delivering enhanced interpretability through explainable AI techniques. This framework presents considerable potential for pharmaceutical toxicity assessment and diminishes dependence on animal testing protocols.**

***Index Terms*—Drug discovery, computational toxicology, cheminformatics, machine learning, Grad-CAM**

I. Introduction

Drug-induced liver injury (DILI) is the leading cause of drug failure in late-stage clinical trials and withdrawal from the market [1]. According to a survey in the United States [2], DILI is associated with more than 1,000 drugs and is a significant cause of acute liver failure cases with more than 50%, making it a top priority in pharmaceutical toxicity research. DILI is a serious condition caused by drugs [3]. In recent years, DILI has remained a topic of great interest in drug discovery and highlighted that it is still an emerging research trend [4]–[6]. Traditionally, the research on DILI risk has been tested using in vivo animal studies and in vitro assays [7]. In vitro refers to biological processes conducted in controlled laboratory environments, such as Petri dishes or test tubes, rather than inside living organisms [8]. This approach allows scientists to observe cellular processes outside their natural environment. In contrast, methods that use living organisms such as rodents or primates for research are called in vivo. This method can observe changes in physiological and biological responses in real-time. Thereby, drug treatment interventions can be used, and the complex interactions between the studied entity and its natural environment can be captured. Drug toxicity testing is a complex, time-consuming, and labor-intensive process that requires extensive testing in human and animal cells. However, retrospective analyses of approximately 45% of clinical trials have shown that animal studies often fail to accurately predict the risk of DILI [9]. Specifically, mouse models were only 43% accurate in predicting human toxicity, while non-mouse models were only 63% accurate, and the success rate was even lower when predicting adverse drug reactions in target organs, with a success rate of less than 30% [10]. This demonstrates the limitations in the ability to extrapolate from animal models to humans, leading to increased risks in drug development. In addition, animal testing is ethically controversial because it involves the use of live animals in experiments that may cause pain or injury [11]. To address these limitations and highlight the need for drug development, alternative testing methods to reduce the reliance on animal testing are continually being investigated. This has stimulated further development of new approaches using artificial intelligence (AI) and computational models to improve the ability to predict DILI early in drug development. Toxicity prediction models aid the elimination of potentially toxic compounds, increase efficiency in the drug discovery process, and aid experimental design to reduce the number of unwanted experiments.

In the context of the rapidly developing Internet of Things (IoT) ecosystem in the pharmaceutical field, AI-powered toxicity assessment systems offer extremely practical contributions. This technology can strengthen and directly support continuously connected drug safety assurance processes on

H. P. Y. Duong, N. T. Vo, T. Q. Duong are with Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1C 5S7, Canada (e-mail: {yhpduong, ntvo, tduong}@mun.ca).

T. T. Nguyen is with University of Greenwich, UK (e-mail: tuan.nguyen@greenwich.ac.uk).

This work was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109.

Corresponding author is Trung Q. Duong (tduong@mun.ca).

digital platforms. Its most notable capability is providing very early warning signals about adverse drug reactions in patients, and promptly detecting the risk of liver toxicity even during preclinical trials. Furthermore, in the integration of IoT technology and the pharmaceutical industry, the ability to monitor drug use in real time is also heavily emphasized. These interconnected systems not only help to quickly detect potential adverse events during treatment but also build a solid foundation to support medical teams in making accurate decisions based on real-world data. From these research directions, it is clear that specialized toxicity prediction modules can fully serve as a sophisticated data analysis system. This analysis system will act as a central brain, supporting the entire pharmacovigilance and monitoring process operated through a network of intelligent devices [12]. Machine learning (ML)–driven frameworks are increasingly being applied to accelerate drug discovery and optimize drug delivery systems by enabling predictive modeling of molecular interactions, intelligent analysis of biomedical data, and adaptive optimization of therapeutic formulations and personalized treatments [13]–[17]. Recently, ML models used in toxicity prediction encode the structure of molecules into molecular descriptors to classify candidates based on their toxicity profiles. There are many popular classical ML algorithms, including Random Forests (RF) and Support Vector Machines (SVM), which have shown usefulness in classifying molecules as toxic or non-toxic [18]. However, traditional ML methods are often limited by the need for manual feature selection and model configuration, making it difficult to automate and scale these models for full high-throughput toxicity prediction [19], [20]. A significant challenge in toxicity prediction is the efficient representation of molecular information read by computers [21]. RF and SVM use molecular fingerprints or descriptors as standalone features to study toxicity or other molecular properties, it cannot analyze interactions between structural fragments [19]. Another approach is to use convolutional neural networks (CNNs), which are essential in toxicity prediction because they efficiently handle complex data, especially high-dimensional data such as molecular structures. Unlike traditional ML models, CNNs can automatically learn and extract features from raw data without manual feature selection, thereby reducing the dependence on human knowledge and improving the accuracy of toxicity prediction compared to classical docking methods [22], [23]. Chemical formulas using 1D-CNNs can be represented as text such as Simplified Molecular Input Line Entry System (SMILES), which is a string representation of molecular structures [24], but this representation may lack visualization in terms of input data. By using two-dimensional (2D) images, molecular structures can be represented as images, allowing the model to learn features beyond linear arrangements [25]. 2D-CNNs are particularly useful in molecular image analysis, where the model can recognize complex features and relationships between different structural fragments. However, the use of conventional molecular images sometimes fails to provide intuitive results in toxicity prediction because these images do not necessarily clearly show important locations and features related to the toxicity of the compound.

To overcome these challenges, heatmap images are used in this study. Heatmaps have the ability to highlight important regions in the molecular structure, thus making it easier for the model to identify important features. Additionally, an essential technique for interpreting CNN model results is Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM helps to better understand which parts of the image, and the CNN model focuses on when making predictions. Using heatmaps with Grad-CAM will allow us to see the regions of the molecular structure that the CNN has learned, analyze how the model performs, and show a significant impact on the prediction, potentially improving the model's ability to predict and interpret results. Therefore, moving from SMILES to 2D-CNN and using heatmap images in combination with Grad-CAM not only improves the accuracy of toxicity predictions but also provides a clearer view of the mechanism of action of the model, helping to better understand the focus of predictive models. Moreover, functional group analysis serves as an explainable AI technique that provides interpretation of toxicity predictions obtained from SMILES-based machine learning models. Although black-box models can achieve high predictive accuracy, they often lack transparency in the decision-making processes, limiting their applicability in drug discovery, where understanding the underlying chemistry is critical. By analyzing predicted toxic and non-toxic compounds into their constituent functional groups, this method converts algorithmic outputs into interpretable chemical insights. The method aims to systematically identify which structural motifs are associated with toxicity outcomes in the analyzed dataset, allowing understanding of outcomes not only what the model predicts but also statistical information about the number of functional groups in chemical formulas that are classified as toxic or safe. This analysis bridges the gap between computational predictions and chemical knowledge, providing additional insights into the structure-toxicity relationship. Furthermore, functional group analysis validates the reliability of the model by ensuring that predictions are consistent with established toxicological models and chemical intuition, thus increasing confidence in AI-driven toxicity assessments for pharmaceutical applications.

In this paper, we leverage CNNs to predict molecular toxicity using image-based representations of molecular structures. Our main goal is to use various input molecular images to explore the influence of input data on DILI toxicity prediction, and the process is illustrated in Fig. 1. Our main contributions are as follows:

- First, we collect and obtain a dataset of DILI molecular formulas in SMILES format and apply data processing methods to convert them into standard images and color map images.
- Second, we use CNNs pre-trained models such as ResNet and EfficientNet to build toxicity classification models using various input images and conduct training, testing, and we examine the influence of input data factors on model accuracy, comparing and choosing which model is better.
- Third, we use Grad-CAM aims to identify specific lo-

cations on the molecular structure that the CNN model pays attention to when making toxicity predictions, generate heatmaps showing important regions on molecular images, and test whether the model has actually learned chemical knowledge or just artifacts.
- Finally, functional group analysis is conducted to provide a posteriori interpretation of AI model predictions by systematically examining the distribution and toxicity ratios of chemical functional groups extracted from SMILES representations of both toxic and non-toxic compounds in the dataset. The analysis aims to validate the reliability of the model through chemical knowledge, assess the prediction accuracy for each functional group, identify which groups were most and least accurately predicted by the model, and detect model bias for certain chemical groups.

The subsequent sections of this paper are arranged as follows: Section II surveys existing research relevant to this study. In Section III, we detail the details of the dataset and the overview of the model architecture, and provide the evaluation metrics used to assess the model's performance. Section IV presents experimental results. Section V presents comprehensive analyses, model performance comparisons, and model interpretations with Grad-CAM. Lastly, the paper wraps up in Section VI.

II. The Intersection of Drug Discovery and Toxicity and Related Work

Drug discovery and development have become essential areas of pharmaceutical research. This resource-intensive pipeline aims to develop new drugs and potential cures for various diseases, including rare conditions [26]. Due to its importance, it attracts significant investment from governments and the research community. Drug discovery and development typically consists of five stages: target discovery and validation, lead discovery and optimization, pre-clinical testing, clinical trials, and FDA approval as shown in Fig. 2 [27]. Traditional drug discovery methods are time-consuming and expensive, often taking over a decade and costing hundreds of millions to billions of dollars to complete and bring to the market [28]. For instance, the average cost of developing a new drug has reached over 2 billion US dollar in recent years [29]. However, this lengthy process is fraught with risks and has a very low success rate due to its complexity [30]. Approximately 96% of drug candidates are rejected during development [31]. Furthermore, the failure rate in late-stage clinical trials is about 90%, emphasizing the need for practical early-stage evaluation, such as lead optimization, to improve the success rate in later stages [32]. The lead optimization is a significant step of this process that shows the intersection of drug discovery and toxicity. Lead is the group of molecular compounds after virtual screening and the primary goal of lead optimization is to eliminate side effects or toxic effects of similar substances [33]. One of the key challenges in this process is the prediction of toxicity, which is a critical application that evaluates the ability of a drug candidate to cause harmful effects on biological systems [34]. The safety issues of drug candidates account for more than 30% of failures, significantly contributing to the high development costs [35]. Accurate toxicity prediction is essential as it enables researchers to exclude potentially unsafe candidates early on, reducing the risk of costly failures in the later stages of clinical trials. Toxicity is determined by measuring the harmful effects of a particular substance on the body or cells and is often the stage between lead optimization and pre-clinical testing. Statistics show that about one-third of drug candidates submitted for marketing are rejected by the FDA due to failure in toxicity assessment [36]. This suggests that toxicity plays a major role in determining whether a drug candidate can be brought to market. Furthermore, failure results in significant drug development costs, and the cycle of candidate optimization and continued pre-clinical and clinical development until safety targets are met results in approximately one-third of drug candidates being rejected. This is a major contributor to drug development costs, especially when toxicity is not detected until late in clinical trials or after marketing.

Due to the importance of toxicity assessment, many computational models have been studied to support more effective prediction, typically AI. The application of AI in this research is practical in terms of reducing the time and cost of conducting many mass experiments. AI can extract relevant information from chemical structures and features to predict toxicity, combining ML and DL, such as different types of neural networks, through which quantitative structure-activity (QSAR) and molecular docking are shown with clear relationships to predict toxicity more accurately. The methods integrating QSAR with AI successfully classified 87% of the evaluated compounds and were confirmed by in vivo experiments with a success rate of 81%, showing positive results when using computational model-based evaluation methods. Showing superiority over conventional evaluation techniques. While AI plays a vital role in drug discovery and has achieved high prediction rates, the importance of testing methods cannot be ignored. The development of AI research will be an essential step in helping to eliminate unnecessary experiments, shorten the time of drug research, and increase patient survival rates.

In silico, toxicity prediction has attracted considerable attention in recent years due to its potential to streamline the drug discovery process by eliminating potentially harmful candidates early in the process. Computational methods for toxicity prediction have been widely developed to address the limitations of traditional in vitro and in vivo assays, which are both costly and resource-intensive. Many studies have developed DILI prediction models using ML in recent years, achieving high performance. One prominent study combined multiple algorithms such as k-nearest neighbors, multilayer perception, random forest, Naïve Bayes, support vector machine (SVM), logistic regression, and Fisher's linear discriminant analysis, achieving an accuracy of up to 84% with 10-fold cross-validation (CV) on 1,075 compounds [37]. Another study utilized consensus models incorporating chemical descriptors from PaDEL, Klekota-Roth, Estate, and PubChem, combined with the hybrid quantum particle swarm optimization algorithm, achieving 80% accuracy, 83.9% sensitivity, and 73.3%

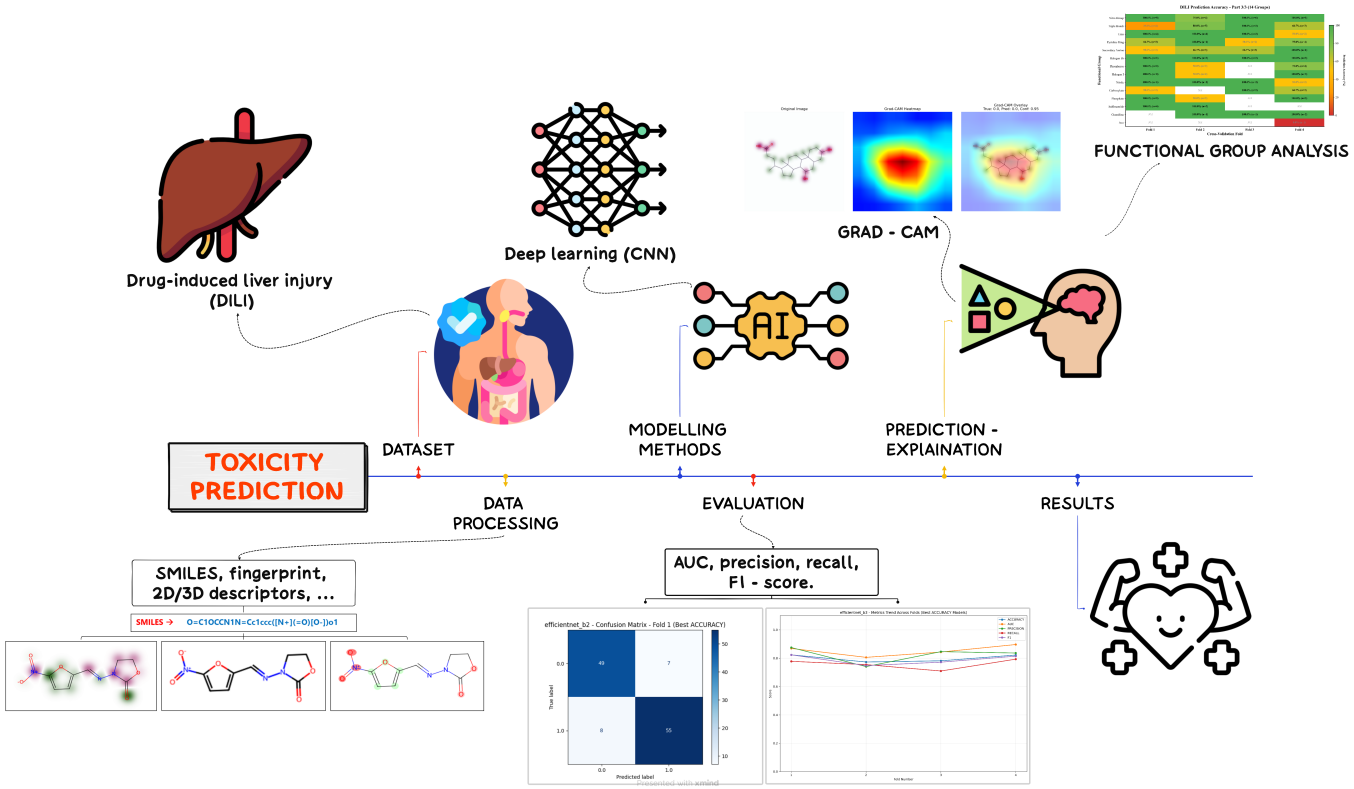

Fig. 1: Toxicity prediction process

specificity when tested on 528 compounds [38], [39]. Additionally, several prediction software tools and graph neural network models have been developed with the primary goal of enhancing DILI prediction, achieving performance exceeding 79% accuracy [40]–[44]. These advances demonstrate the great potential of AI in improving the accuracy and efficiency of hepatotoxicity prediction, contributing to risk reduction in drug development. In addition, some studies have been conducted using DL models to improve toxicity prediction accuracy. The DeepDILI model was developed [45] using deep neural networks (DNNs) to improve the accuracy compared to traditional ML approaches, such as RF and SVM. This study used a dataset with 1,002 drug samples, achieving results showing that ACC reached 68.7%. Several researchers have proposed methods using CNNs to predict DILI, but with different approaches in terms of input features and model performance. One approach [6] utilized CNN combined with NLP by using molecular fingerprint-embedded features, trained the model on 1,597 compounds, and tested it on 322 compounds. This model achieved an accuracy (ACC) of 89% and a ROC-AUC of 96%. Meanwhile, another study [46] introduced ResNet18DNN to evaluate 1,446 compounds and achieved a very high accuracy of 95.8%, which demonstrated superior performance compared to the fingerprint-based CNN method. Both approaches showed excellent results, highlighting the importance of using CNN architectures. Despite much progress, DILI prediction remains a complex problem due to the multifactorial nature of the disease [47]. Structural similarity between DILI-positive and DILI-negative compounds remains a major challenge. The need for larger, higher quality, and more diverse data (combining chemical and biological data) continues to be emphasized to develop more robust and reliable DILI prediction models in the future.

III. Materials and Methods

A. Dataset and data processing

1) Dataset: Currently, the DILI database is quite limited, with only about 1000-2000 molecules and not clearly classified. This study uses the DILI benchmark dataset. This dataset is compiled from the US FDA National Center for Toxicological Research [48], which consists of 475 compounds with the number of toxic and non-toxic samples detailed in Fig. 3. Each molecule is represented by its chemical structure and relevant features associated with toxicity prediction. To evaluate the model performance, K-fold CV is used to ensure that every molecule is used for both training and validation. This approach helps to reduce the risk of overfitting and gives the model a robust generalization capability. The dataset is divided into four subsets, ensuring an equal distribution of toxic and non-toxic molecules within each fold. Each subset is used once as the validation set, while the remaining subsets serve as the training set. This process is repeated four times, ensuring all molecules are utilized in training and validation.

2) Data Processing: In this study, we use SMILES as the primary molecular representation. SMILES is a widely used text-based format that efficiently encodes molecular structures. To ensure data consistency and minimize errors, data preprocessing is performed. First, SMILES strings are

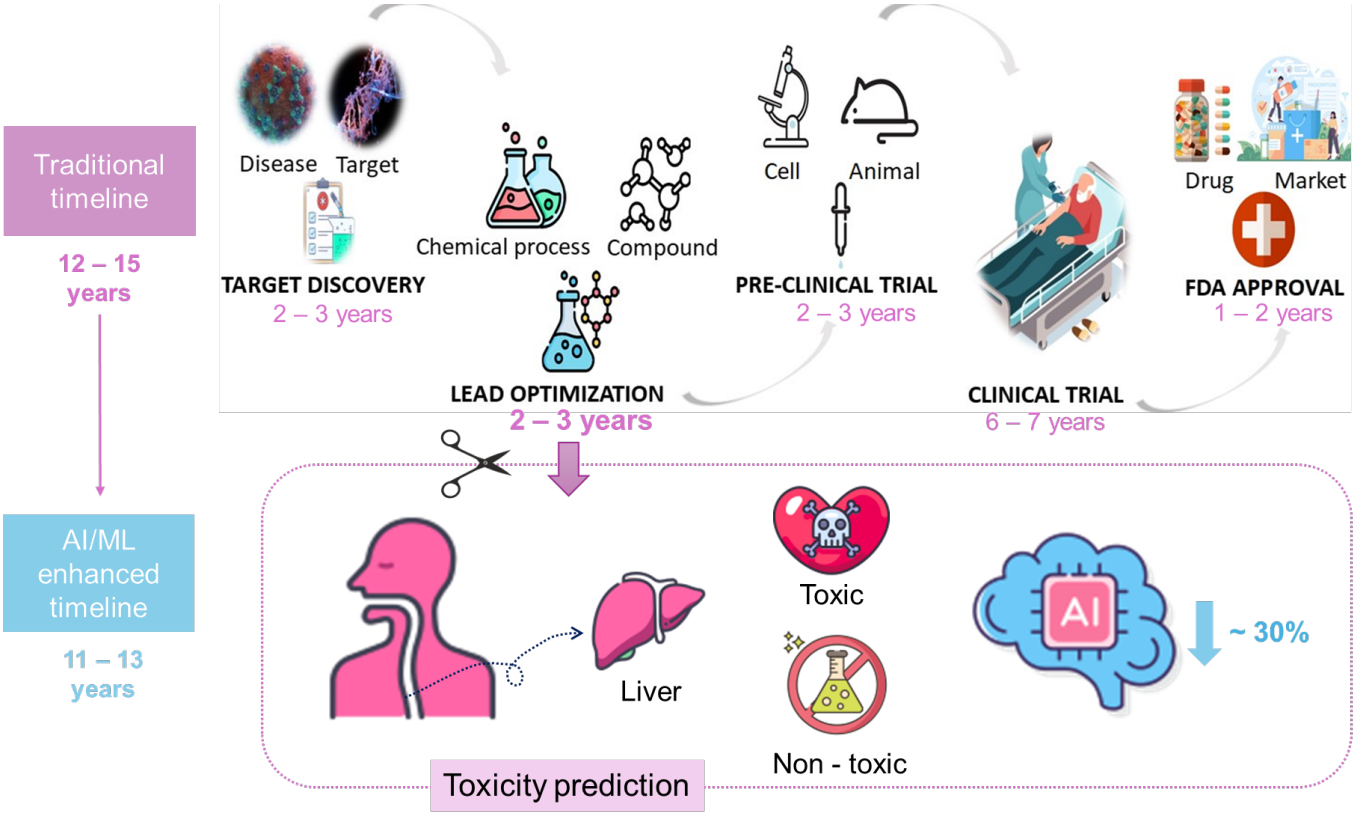

Fig. 2: The intersection of drug discovery and toxicity.

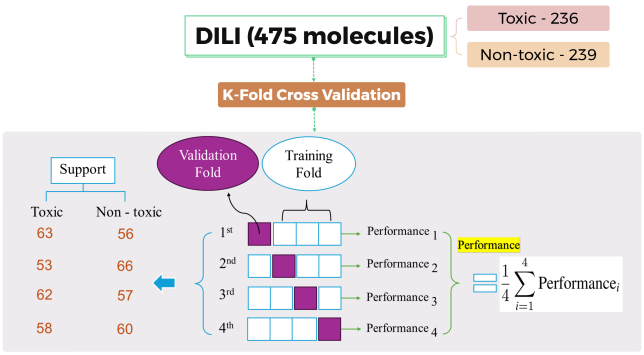

Fig. 3: Drug-induced liver injury (DILI) dataset

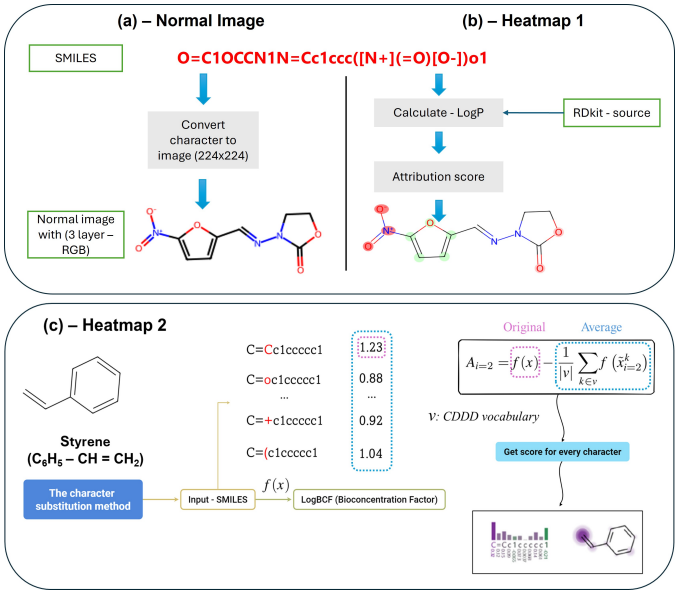

Fig. 4: The method to convert SMILES to images. *(a) Normal Image (b) Heatmap-1 (c) Heatmap-2*

normalized, and any invalid molecular structures are removed. The SMILES strings are then converted into molecular images using RDKit [49], an open-source cheminformatics toolkit. Three different types of input images are utilized in this study: Normal molecular images (NM), heatmap 1 (HM1), and heatmap 2 (HM2).

a) SMILES to NM: SMILES sequences of compounds is converted into 2D molecular color images with a resolution of 224×224 pixels. These images are represented as the 3-channel color model, e.g., red, green, and blue (RGB), to encode atomic structures and bond connectivity. This color representation is designed to enhance feature extraction and facilitate the subsequent processing by CNN-based models for toxicity prediction, as shown in Fig. 4a.

b) SMILES to HM1: Crippen logP [50] is a model that predicts the logP (logarithm of the partition coefficient) value of a molecule based on the contributions of each type of atom in the structure of that molecule, and it is implemented in RDKit. In addition, logP is an important parameter in drug development that measures the ability of a molecule to

dissolve in a hydrophobic solvent (such as oil) in comparison to a hydrophilic solvent (such as water) [51]. It helps to evaluate the solubility and cell membrane penetration ability of a drug molecule. Accurate prediction of logP helps to screen and design potential drug molecules more effectively. In the Crippen logP model, logP was calculated as follows:

$$\text{logP} = \sum_i n_i a_i, \tag{1}$$

where n_i is quantity of a specific type of atom i, a_i represents logP contribution value of that type of atom.

The logP value is predicted based on the Random Forest regression model in the scikit-learn library [52]. Approximately 250,000 molecules are used for training and testing [53]–[55]. For molecular representation, the extended connectivity fingerprint method (ECFP4) is used, where each structural fragment is assigned a random position in a binary vector of length 2048, with the presence and absence of a molecular fragment represented by 1 and 0, respectively. The atomic contributions are visualized as colored contour plots, with the number of contour lines increasing gradually from the minimum to the maximum value, following the formula:

$$N_{\text{contour}} = \frac{ac_{\text{max}} - ac_{\text{min}}}{0.06}, \tag{2}$$

where ac_{max} and ac_{min} represent the highest and lowest atomic contributions of the molecule, respectively. The value of N_{contour} is rounded to the nearest integer. This method is known as atom attribution from fingerprints [56], [57]. After completing the conversion from SMILES to a heatmap, visualize the image and resize it to 224×224 pixels, as depicted in Fig. 4b.

c) SMILES to HM2: An explanation method is used in this study, and logBCF is used to assess the bioaccumulation of chemical compounds in the body. LogBCF is defined as the logarithm of the bioconcentration factor (BCF), a key indicator for evaluating the environmental risk of chemical compounds [58]. In the context of the BCF model, a post-hoc perturbation method [59] is applied to analyze the contribution of each character in the SMILES string to the model prediction and convert it into a heatmap image based on available research. To assess the importance of each character in the SMILES string, they use feature removal and masking techniques, in which a base character replaces a character at position i to observe the changes in the prediction.

The heatmap processing is represented as shown in Fig. 4c, where $f : \mathbb{R}^{v \times n} \rightarrow \mathbb{R}^2$ be a function that maps one-hot encoded SMILES strings (with vocabulary size v and sequence length n) to two target outputs, log BCF and log D. Suppose $\tilde{x}_i \in \mathbb{R}^{v \times n}$ is a perturbed version of x, where the character at position i is replaced by a baseline character, and the contribution (or sensitivity score) A_j^i of that character to target j is given by:

$$A_j^i = f_j(x) - f_j(\tilde{x}_i),\ i \in \{0, ..., n\}, j \in \{0, 1\}, \tag{3}$$

where a positive A_i and a negative A_i indicate that the predicted value decreases and increases when character i is replaced with the baseline character, respectively.

Additionally, a character substitution method is applied to evaluate the sensitivity of each character in the SMILES string by calculating the average model prediction when tall possible characters in the vocabulary replace that character. This method helps quantify the influence of each character on the prediction result, providing more detailed information about how the model learns and makes decisions based on SMILES input data.

$$A_j^i = f_j(x) - \frac{1}{|\nu|} \sum_{k \in \nu} f_j(\tilde{x}_i^k),\ i \in \{0, ..., n\}, j \in \{0, 1\}, \tag{4}$$

where ν is the vocabulary set, and $\tilde{x}_i^k$ is the perturbed SMILES string, where the character at position i is replaced by character k.

B. Data Augmentation

All input molecular images are resized before being fed into the CNN model. Specifically, each image is resized to $1.15\times$ the target input resolution prior to random cropping. A unified strong augmentation pipeline is applied consistently across all convolutional architectures (ResNet and EfficientNet families) during training to mitigate overfitting on the limited dataset. Random rotation up to $\pm 90°$ is applied since molecular structure images have no fixed orientation. To handle variations in image rendering and visualization style across different sources, *ColorJitter* is used with brightness, contrast, and saturation adjustments of ± 0.3 and a hue shift of ± 0.05. Random affine transformations with translation of $\pm 10\%$ and shear of $10°$ are included to add geometric diversity. Gaussian blur with kernel size 3 and σ ranging from 0.1 to 1.0 is applied to reduce sensitivity to high-frequency noise. All images are normalized using per-fold mean and standard deviation computed solely from the training subset to avoid data leakage. Finally, *RandomErasing* is applied with probability $p = 0.3$ over small rectangular regions (2–10% of image area), preventing the model from over-relying on any specific local region of the molecular image. During validation, only center cropping and channel-wise normalization are applied to preserve image fidelity. The target input resolution is set to 224×224 pixels for ResNet variants and EfficientNet-B0, and 300×300 pixels for EfficientNet-B1 through B3, following the recommended specifications for each architecture. The augmentation pipeline is summarized in Table I.

C. Model Architecture

This study uses two well-known DL models: ResNet [60] and EfficientNet [61] to classify molecular images and predict their toxicity.

1) EfficientNet models: We use EfficientNet models, including EfficientNet-B0 through B3, to classify molecular images. EfficientNet uses a compound scaling method that balances network depth, width, and resolution to achieve better accuracy with fewer parameters. Each model is pre-trained on ImageNet, and we fine-tune them by replacing the final classification layer to support binary output.

TABLE I: Data augmentation strategies applied during training

Phase / Model	Augmentation Techniques
Training (ResNet & EfficientNet)	- Resize to 1.15× target resolution - RandomResizedCrop (scale = 0.8–1.2) - RandomRotation (±90°) - ColorJitter (brightness= 0.3, contrast= 0.3, saturation= 0.3, hue= 0.05) - RandomAffine (translate= (0.1, 0.1), shear= 10°) - GaussianBlur (kernel= 3, $\sigma \in [0.1, 1.0]$) - RandomErasing (p = 0.3, area ratio 0.02–0.10) - Normalize (per-fold mean & std)
Validation (All models)	Resize to 1.15× target resolution; CenterCrop to target resolution; Normalize (per-fold mean and std).
Input resolution	ResNet and EfficientNet-B0: 224 × 224 px; EfficientNet-B1/B2/B3: 300 × 300 px.

2) ResNet models: : The ResNet model employs a residual architecture, using skip connections within residual blocks to mitigate the vanishing gradient problem and facilitate training. We select ResNet-18, ResNet-34, and ResNet-50 as our base models. Each model is pre-trained on ImageNet and modified for binary classification by replacing the final fully connected layer.

All experiments are conducted on a workstation equipped with an NVIDIA GeForce RTX 5060 Ti GPU (16 GB VRAM) running CUDA, using Python with PyTorch and TorchVision as the DL framework. All models are trained using the AdamW optimizer. Leveraging ImageNet-pretrained weights as initialization, all models converged well before the maximum epoch limit due to early stopping, demonstrating that transfer learning substantially reduces the computational burden for molecular image classification tasks of this scale. We apply early stopping to prevent overfitting and use 4-fold CV to validate model performance. The main evaluation metrics include AUC, Recall, Precision, and F1-score. The input image is resized, and a dropout layer is added to reduce overfitting. We optimize these modified hyper-parameters to maximize classification performance on the DILI toxicity prediction task based on the analysis of model responses, as shown in Table II.

TABLE II: Hyperparameter configuration for ResNet and EfficientNet

Hyper-parameter	ResNet	EfficientNet
Batch size	16	32
Learning rate (LR)	5×10^{-5}	1×10^{-4}
Weight decay	1×10^{-3}	1×10^{-4}
Optimizer	AdamW	AdamW
LR scheduler	ReduceLROn Plateau	Cosine Annealing
Early stopping	Yes	Yes

D. Evaluation metric

The performance of the model is evaluated based on four main criteria: the area under the receiver operating characteristic (ROC) curve (AUC), Recall, Precision, and F1-score as [62]

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

where TN is true negatives, FN is false negatives, TP represents true positives, and FP represents false positives.

The AUC measures the ability of the model to distinguish between two classes. The Recall represents the proportion of correctly predicted positive samples compared to the total number of positive samples. Meanwhile, Precision measures the proportion of correctly predicted positive samples compared to the total number of samples predicted as positive. F1-score is used as a balance measure between Precision and Recall.

E. Grad-Cam Visualization

Grad-CAM [63] is a visualization technique in deep learning, computed from the final layer of a CNN or similar model, that aims to help explain the model's decisions by highlighting important regions in the input data (images) that the model relies on to make predictions. In the field of toxicity prediction, Grad-CAM is used to explain predictions, check the model's reasonableness, and improve the model. The main formula of Grad-CAM is as follows:

$$L^{\text{class}}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k A^k\right), \tag{8}$$

where A^k represents the feature maps of the last convolutional layer, $\alpha_k = \frac{\partial y^{\text{class}}}{\partial A^k}$ are the gradients pooled over the spatial dimensions, y^{class} is the score for the target class.

The application of Grad-CAM provides interpretable heatmaps, ensuring that the model's predictions are based on biologically or chemically relevant features in the input data.

IV. Results

Table III presents the performance of seven DL models (ResNet-18, ResNet-34, ResNet-50, EfficientNet-B0, B1, B2, and B3) on three types of molecular image representations, including NM, and two heatmap-enhanced versions (HM1 and HM2). The evaluation uses four metrics, including AUC, Recall, Precision, and F1-score, with Recall being prioritized to emphasize the model's ability to correctly identify toxic compounds (DILI-positive), which is a critical factor in drug safety applications.

The results indicate that heatmap images, particularly HM1 and HM2, generally improve model performance compared to conventional images. For example, ResNet-34 achieves the highest Recall (0.8410; 0.8499) and AUC (0.8853; 0.8594) on HM2 and HM1, compared to Recall of 0.8295 and AUC

of 0.8604 on NM, indicating a clear improvement. ResNet-50 also improves its Recall from 0.8061 on NM to 0.8222 on HM2, while maintaining a high AUC of 0.8697. ResNet-18 shows relatively consistent performance across all three image types but performs lower than ResNet-34 on heatmap images in terms of both Recall and AUC.

Among the EfficientNet models, EfficientNet-B2 achieves the highest Recall on NM (0.8374) and HM1 (0.8624), demonstrating strong sensitivity in detecting DILI-positive compounds. However, its AUC is lower than that of the ResNet models, and its lower Precision suggests a trade-off between sensitivity and specificity. EfficientNet-B3 shows the lowest performance across all inputs, with Recall values of 0.8125, 0.8104, and 0.7772 on NM, HM1, and HM2, respectively, and the lowest AUC among all models.

In terms of overall performance balance, both ResNet-34 and ResNet-18 achieve high and stable F1 scores. This observation suggests that heatmap images not only improve Recall but also maintain classification accuracy. These findings confirm that heatmap-enhanced inputs are advantageous in improving the model's ability to detect toxic compounds compared to conventional molecular images. Among the evaluated models, ResNet-34 provides the best and most stable performance.

Based on Fig. 5, the performance comparison between ResNet-34 and EfficientNet-B0 of heatmap image is shown. Fig. 5A presents the mean training and validation loss and accuracy curves across 4-fold CV for ResNet-34 and EfficientNet-B0, with the shaded band representing ±1 standard deviation across folds. For ResNet-34, both training and validation losses decrease steadily from approximately 0.83 and 0.66 at epoch 1, respectively, converging to 0.43 and 0.56 by epoch 24. The validation loss reaches its minimum at epoch 20 and remains stable thereafter, with a negligible drift of only +0.007 above the minimum, indicating no divergence between training and validation performance. The training accuracy increases from 54% to 86%, while the validation accuracy stabilizes above 79% from epoch 15 onward, yielding a controlled training-validation gap of 6.6%. The narrow shaded band throughout training further confirms consistent behavior across all folds. In comparison, EfficientNet-B0 exhibits a similar convergence trend within the stable training (23 epochs), with training and validation losses decreasing to 0.31 and 0.47, respectively. However, the validation loss begins to plateau after epoch 18 while training loss continues to decline, resulting in a slightly wider training-validation accuracy gap of 8.4% and broader standard deviation bands, suggesting greater sensitivity to data partitioning.

Fig. 5B illustrates the per-fold validation accuracy for both models under 4-fold CV. ResNet-34 achieves consistent performance across all folds, with individual fold accuracies of 84.0%, 82.3%, 83.2%, and 82.2%, yielding a mean accuracy of 82.94% and a standard deviation of only 0.73%. This narrow variance demonstrates that ResNet-34 generalizes reliably regardless of data partition, with even the lowest-performing fold remaining within 0.8% of the overall mean. EfficientNet-B0, while achieving a higher peak accuracy of 86.6% in Fold 1, exhibits substantially greater variability across folds, with accuracies of 86.6%, 84.0%, 84.0%, and 74.6%, resulting in a comparable mean of 82.30% but a standard deviation more than six times higher than that of ResNet-34. The notably lower accuracy observed in Fold 4 indicates that EfficientNet-B0's performance is more sensitive to the composition of the training set, which limits its reliability in clinical deployment where consistent prediction across diverse compound datasets is a critical requirement.

Fig. 5C shows the distribution of prediction confidence scores aggregated across all validation folds for each model. ResNet-34 produces a right-skewed distribution concentrated toward higher confidence values, with a mean of 0.827 and 40.8% of all predictions exceeding a confidence threshold of 0.9, indicating that the model assigns high-certainty outputs to the majority of test samples. EfficientNet-B0 yields a broader and more uniform distribution with a lower mean confidence of 0.796, where only 31.6% of predictions surpass the same threshold, representing a reduction of 9.2 percentage points relative to ResNet-34. The wider spread of confidence scores observed in EfficientNet-B0 reflects greater prediction uncertainty, which is consistent with the higher CV variance reported in Fig. 5B.

Overall, ResNet-34 is chosen as the preferred backbone network due to its more stable generalization properties, making this model more reliable for predicting DILI toxicity.

V. Analysis and Discussion

A. Gradient-weighted Class Activation Mapping

Based on the highest value of AUC and Recall of these investigated models, ResNet-34 of HM2 is chosen as the best model to use Grad-CAM to visualize and evaluate.

Fig. 6 (example 1), 7 (example 2), and 8 (example 3) indicate the input types of NM and two types of heatmaps, HM1 and HM2, using Grad-CAM images, which help clarify their role in visualizing the important zone of the model when predicting drug toxicity. The circled areas in the three examples depict the difference in the focusing of the model into 3 types of images. In the original of the NM, Grad-CAM shows that the model is uncertain where to focus due to the absence of highlighted areas for the functional groups with similar shapes.

The HM1 model can improve its ability to focus more activation in the central region. The HM1 images highlight the atomic positions and functional groups on the chemical formulas, with the side chain functional groups on the right marked in red and the benzene group marked in light green. Grad-CAM indicates that the model has captured the specific important points that the model pays attention to through the marked positions and focuses more on learning these formula regions than the conventional images. However, in examples 2 and 3, highlighting light green or colors similar to the original image makes it difficult for the model to distinguish which positions are important, especially the positions marked in light green. The model can only focus on the positions marked in red. This suggests that the choice of marker color is also an important factor in helping the model focus on the necessary positions.

TABLE III: Performance of DL-based prediction models with different inputs

Model	K-Fold	AUC			Recall			Precision			F1-score		
		NM	HM1	HM2	NM	HM1	HM2	NM	HM1	HM2	NM	HM1	HM2
ResNet-18	Mean	0.8765	0.8669	0.8604	0.8347	0.8336	0.8398	0.8109	0.8254	0.8050	0.8212	0.8278	0.8213
	Std (±)	0.0129	0.0102	0.0027	0.0454	0.0541	0.0243	0.0338	0.0287	0.0176	0.0219	0.0233	0.0091
ResNet-34	Mean	0.8604	0.8594	0.8853	0.8295	0.8499	0.8410	0.8048	0.8144	0.8318	0.8162	0.8307	0.8356
	Std (±)	0.0139	0.0106	0.0283	0.0367	0.0461	0.0656	0.0174	0.0259	0.0273	0.0127	0.0201	0.0436
ResNet-50	Mean	0.8682	0.8711	0.8697	0.8061	0.8194	0.8222	0.8567	0.8368	0.7950	0.8272	0.8262	0.8083
	Std (±)	0.0164	0.0099	0.0254	0.0868	0.0303	0.0126	0.0256	0.0542	0.0147	0.0421	0.0203	0.0111
EfficientNet-B0	Mean	0.8643	0.8679	0.8186	0.8273	0.8682	0.7863	0.8186	0.7942	0.7590	0.8215	0.8293	0.7707
	Std (±)	0.0194	0.0303	0.0255	0.0554	0.0355	0.0585	0.0251	0.0534	0.0575	0.0254	0.0441	0.0452
EfficientNet-B1	Mean	0.8021	0.8167	0.8295	0.7623	0.7976	0.8053	0.8243	0.8235	0.8459	0.7914	0.8101	0.8238
	Std (±)	0.0231	0.0318	0.0159	0.0421	0.0514	0.0318	0.0333	0.0363	0.0416	0.0317	0.0423	0.0174
EfficientNet-B2	Mean	0.8550	0.8442	0.8676	0.8374	0.8624	0.7697	0.8077	0.7779	0.8335	0.8207	0.8165	0.7922
	Std (±)	0.0201	0.0605	0.0187	0.0633	0.0331	0.1183	0.0386	0.0713	0.0358	0.0360	0.0465	0.0571
EfficientNet-B3	Mean	0.8492	0.8597	0.8290	0.8125	0.8104	0.7772	0.7920	0.8346	0.7720	0.8011	0.8205	0.7739
	Std (±)	0.0126	0.0273	0.0295	0.0349	0.0453	0.0605	0.0232	0.0462	0.0343	0.0068	0.0257	0.0420

Most notable is the HM2 input, with its large and strong activation region, demonstrating the model's high confidence in recognizing this class. The HM2 result is particularly impressive as it shows the ability to focus on the widest range of all three cases. This can be interpreted as the model learning to strongly and decisively recognize the characteristic features of the HM2 class. A large activation region is not necessarily a negative thing, but may indicate that the model is capturing many complex and correlated features in the molecular structure. This high confidence in the HM2 classification, shown by the strong heatmap, may reflect that this class has clearer and more distinguishable features than the other classes. This result shows that the model has developed effective recognition capabilities, especially good for the HM2 class.

B. Comparison with existing methods

The performance of the proposed ResNet-34 model using HM1 and HM2 representations is compared with state-of-the-art models in Table IV. The goal of this comparison is to evaluate the model performance in the context of DILI prediction.

Based on the results, the ResNet-34 model with NM, HM1, and HM2 achieves an AUC of 0.8604, 0.8594, and 0.8853, respectively. All versions outperform traditional ML models such as Random Forest (0.710), Mixed Learning (0.762), and Naïve Bayes (0.740). In terms of Recall, ResNet-34 achieves 0.8295, 0.8499, and 0.8410, which is higher than Deep Graph Learning (0.768), DeepDILI (0.805), and many other ML models. In terms of F1-score, ResNet-34 with HM1 achieved 0.8307, which is superior to Deep Graph Learning (0.753) and DeepDILI (0.755).

However, ResNet18DNN and CNN-NLP report higher AUC values of 0.958 and 0.960, respectively. This discrepancy is largely attributable to differences in evaluation methodology and dataset characteristics rather than model capability alone. There are many reasons why the ResNet18DNN and CNN models have higher performance than ResNet and EfficientNet. Although the ResNet-34 model has good performance compared to traditional methods by transfer learning using pre-trained models, it still has a few limitations in recognizing chemical molecular formula images, indicating the need for certain architectures with algorithms specifically designed for this type of image.

Different evaluation methods between studies also affect the results. Our model uses K-Fold CV, which ensures that the model is tested on many different datasets, resulting in better generalization. Meanwhile, the ResNet18DNN used a fixed test set, which may result in a higher AUC but does not guarantee generalization when applied to new data. Therefore, although the ResNet18DNN has a higher AUC, the ResNet-34 model has higher stability thanks to K-Fold. In addition, the ResNet18DNN model's data has a serious imbalance between the number of positive (1189) and negative (257) samples. This can lead to model bias, especially when the number of Negative samples in the test set is only 18 samples. When the number of samples of two classes is too different, the model tends to prioritize classifying the dominant group, in this case, DILI positive. From a computational perspective, ResNet-34 provides an efficient balance between model capacity (21.8M parameters) and classification performance, outperforming both the lighter ResNet-18 and the heavier ResNet-50 in our experiments. This suggests that intermediate-

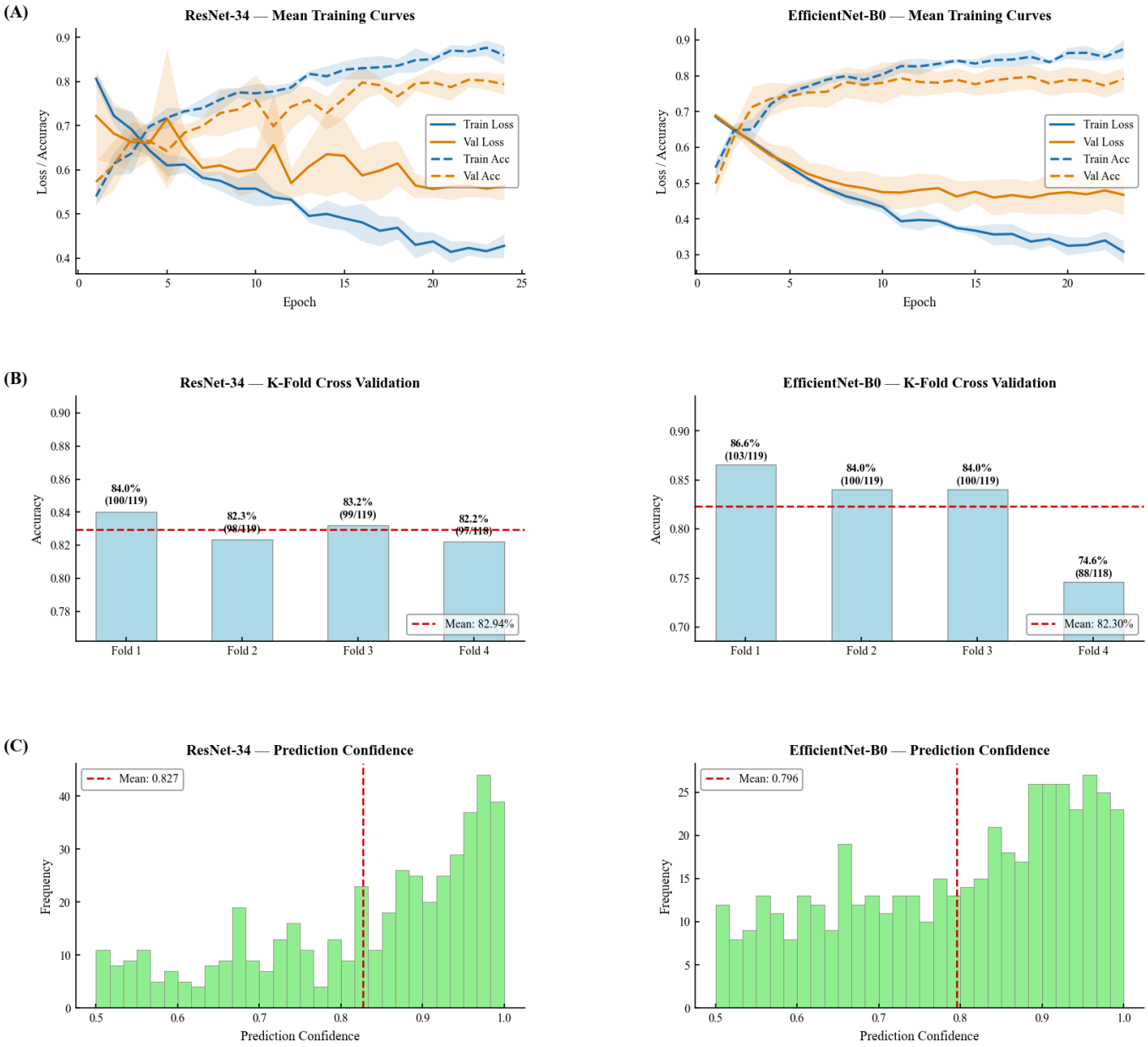

Fig. 5: Distribution of accuracy and prediction confidence of ResNet-34 and EfficientNet-B0 in HM1

depth architectures are better suited for molecular datasets of this scale, where deeper models may overfit and lighter models may underfit. The integration of Stochastic Weight Averaging (SWA) [64] further improved generalization with negligible additional overhead. Compared to existing approaches that relied on fixed test sets and larger proprietary datasets, our framework achieves competitive performance under a more rigorous 4-fold CV protocol, demonstrating both predictive reliability and computational practicality.

C. Functional group distribution analysis

Although CNN models can achieve high predictive accuracy, they inherently lack transparency in their decision-making process, making it difficult to understand which chemical features drive toxicity predictions. Functional group analysis addresses this limitation by examining the distribution and toxicity ratios of structural motifs extracted from SMILES representations, bridging the gap between computational predictions and established chemical knowledge. Our study analyzed molecular descriptors consisting of 40 groups, with the main goal of helping to clarify the basic chemical features in the DILI dataset, as shown in Fig. 9 and Fig. 10. The results showed that the majority of compounds in the dataset had the same basic structure. Heavy atoms appeared in 100% of the compounds, while carbon atoms and aromatic atoms appeared in 99.8%, indicating that the chemical formula of the drug often has a structural framework based on carbon and aromatic rings.

A clear trend emerged when observing the frequency, structural features such as branches (96.8%) and rings (93.3%) were

TABLE IV: Comparison of different prediction models and testing methods for DILI

Model	Data	Test-Method	AUC	Recall	Precision	F1-score
ResNet18DNN [46]	1,446	Test set	0.958	0.935	0.947	0.926
DeepDILI [65]	–	Test set	0.659	0.805	–	0.755
Binary Bayesian [66]	1,036	Test set	0.833	0.720	–	–
Deep Graph Learning [67]	1,300	Validation	0.882	0.768	–	0.753
SVM [60]	–	Test set	0.552	0.852	–	–
SVM [68]	1,317	Test set	0.65	–	–	–
Random Forest [69]	–	Test set	0.710	0.720	0.750	–
Mixed Learning [47]	–	Test set	0.762	0.681	–	–
Bayesian [70]	295	10-fold CV	0.620	–	–	–
Naïve Bayes (NB) [71]	420	Test set	0.740	–	–	–
Decision Forest [72]	197	10-fold CV	–	–	–	–
Stacking Ensemble (NB) [39]	1,087	5-fold CV	0.740	–	–	–
Ensemble-Top5 [73]	1,241	5-fold CV	0.760	–	–	–
CNN - NLP [6]	1,919	Test set	0.960	–	–	–

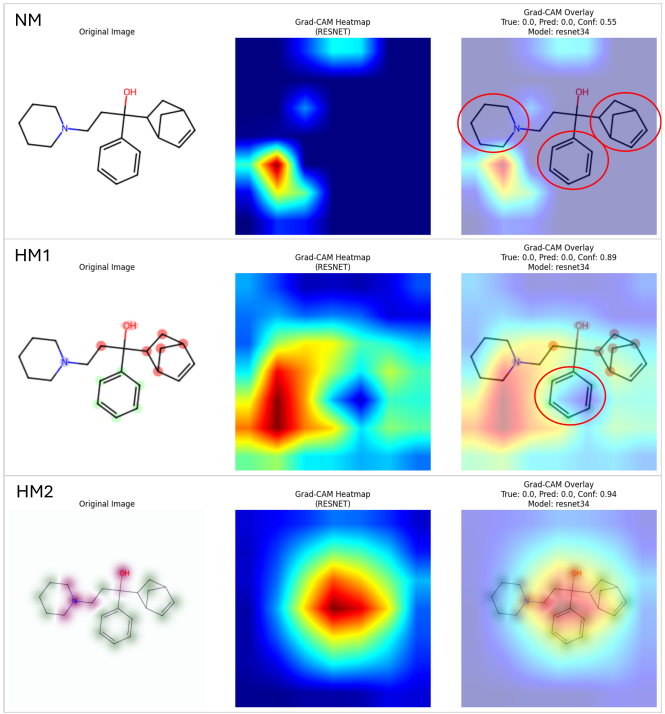

Fig. 6: Grad-CAM result for example 1.

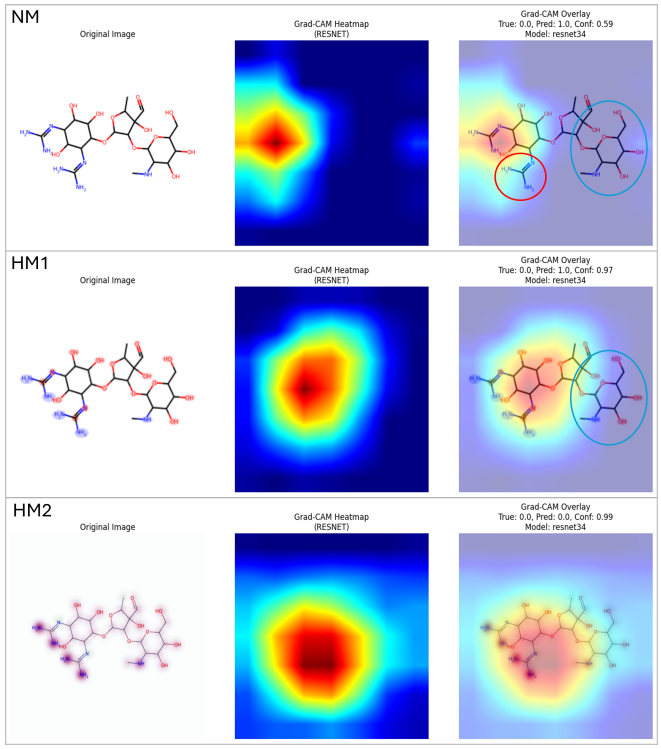

Fig. 7: Grad-CAM result for example 2.

very common, while heteroatoms such as oxygen (89.0%) and nitrogen (85.9%) were also very common. This shows that the chemical formula only includes a certain number of elements, and the appearance of strange elements is very rare, making identification difficult. However, special types of bonds such as triple bonds are very rare (less than 5%), and this low frequency and specificity makes the model easier to remember, supporting more accurate toxicity classification.

Analysis of functional groups and DILI toxicity shows some remarkable patterns. The groups with the highest toxicity levels (over 90%) include: Nitro group (93.75% toxicity, 100% prediction accuracy), Sulfone (92.31%) and Urea (91.67%). These are strong indicators of toxicity, especially the Nitro group is very important because it is predicted accurately by the model. Highly toxic groups (80–90%) include Sul-

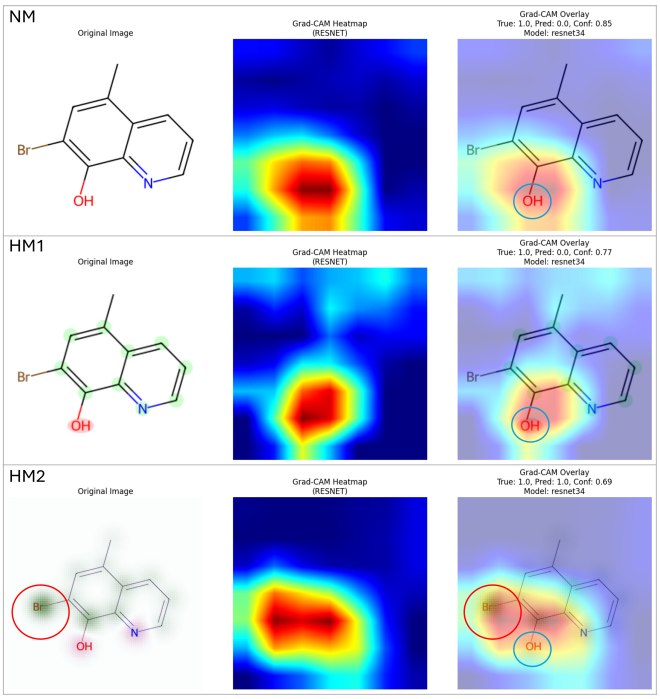

Fig. 8: Grad-CAM result for example 3.

foxides (88.89% toxicity and accuracy), Sulfur (81.82%) and Hydrazine (80.95%, and marked as high risk). In contrast, some relatively safe groups include Halogen I (0% toxicity, 7 compounds), Azo (0%, 1 compound), Ketone (29.41%) and Phosphate (28.57%).

In terms of frequency, the most common features include Carbon and Aromatic Atoms (99.79%), Branches (96.84%) and Rings (93.26%). However, rare but important groups such as Nitro (3.37%) and Sulfone (5.47%) deserve attention due to their strong association with toxicity.

In terms of model performance, the best predicted groups included Nitro (100%), Hydrazine (95.24%), Sulfone (92.31%), and Urea (91.67%), while the less predictable groups included Azo (0% correct, only one sample), Ketone (70.59%), and Carboxylate (71.43%). The main findings point out that sulfur-containing compounds, Sulfoxides, Sulfones, and Sulfonamides are generally highly toxic, making sulfur derivatives potential indicators. The toxicity of the halogen group varies significantly, such as Chlorine, Fluorine are more toxic (58.46% and 67.5%), while Iodine is completely safe. The dataset is generally well balanced, with the most common functional groups having toxicity levels around 50%. These insights demonstrate that the toxicity potential depends on the functional group distribution and location, as well as the frequency of occurrence during drug formulation, helping to guide safer drug development.

VI. Conclusion and Future Works

In conclusion, our multi-representation deep learning approach with explainable AI capabilities advances the state-of-the-art in computational toxicology. The experimental results reveal that ResNet-34 with HM2 input achieves optimal performance, attaining an AUC of 0.8853 and a Recall of 0.8410,

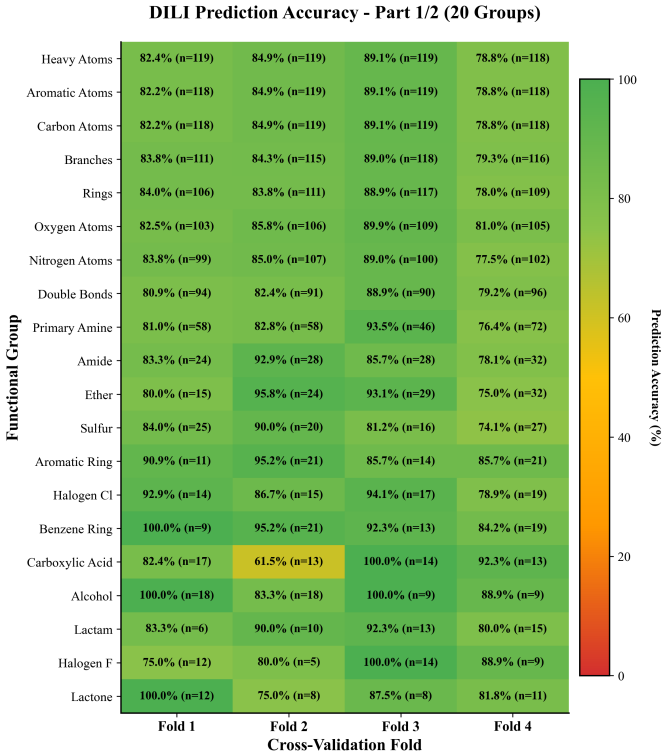

Fig. 9: DILI prediction accuracy - Part 1

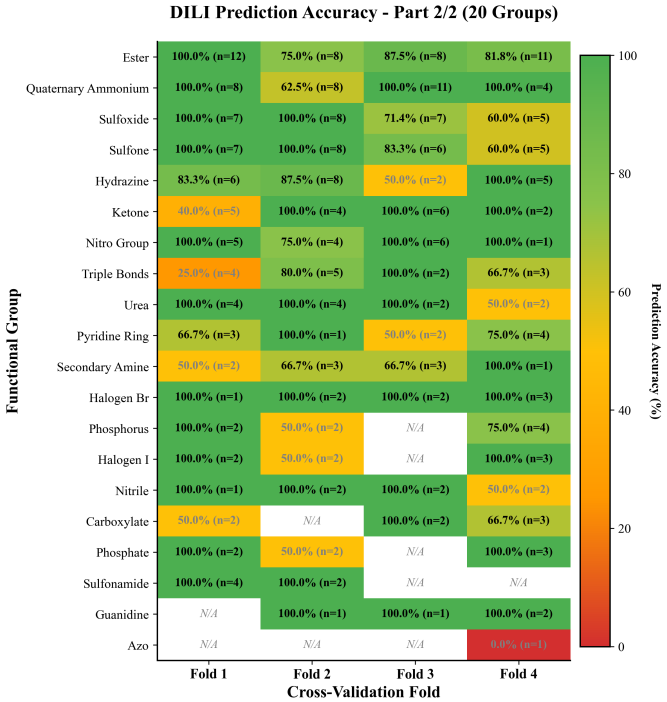

Fig. 10: DILI prediction accuracy - Part 2

representing substantial improvements over normal molecular images. ResNet-34 and EfficientNet-B0 with HM1 exhibits superior functional group identification capabilities, achieving a Recall of 0.8499 and 0.8682. These findings confirm that incorporating chemical knowledge through heatmap visualizations enhances model performance by highlighting toxicologically relevant molecular features. Therefore, using heatmaps will help the model to better localize the focus. While HM1 emphasizes important functional groups, HM2 can help the model analyze the whole molecule, thereby improving the generality in toxicity prediction. However, several limitations warrant consideration. The dataset size of 475 compounds, while standard for DILI research, remains relatively small for DL applications. Future work should investigate larger, more diverse datasets to improve model robustness. Additionally, the study focuses primarily on structural features; incorporating pharmacokinetic and metabolic data could further enhance prediction accuracy.

In the future, the combination of HM1 and HM2 could be a potential approach that can help the model focus on important regions while having an overall view of the molecular structure. This may improve model performance in predicting drug toxicity. The integration of multiple molecular representations with explainable AI techniques offers promising potential for pharmaceutical toxicity assessment. This framework reduces dependence on costly animal testing protocols while providing interpretable predictions that align with established toxicological knowledge. The ability to identify specific functional groups associated with hepatotoxicity enables medicinal chemists to make informed decisions during drug design and optimization processes. Future research could integrate this toxicity prediction framework into IoT-enabled pharmaceutical monitoring systems, where multi-representation molecular images are combined with complementary pharmaceutical data such as molecular descriptors and pharmacokinetic profiles to enable more accurate real-time DILI risk screening during early-stage drug development.

REFERENCES

[1] D. Cook, D. Brown, R. Alexander, R. March, P. Morgan, G. Satterthwaite, and M. N. Pangalos, "Lessons learned from the fate of astrazeneca's drug pipeline: A five-dimensional framework," *Nature Reviews Drug Discovery*, vol. 13, no. 6, pp. 419–431, 2014.

[2] G. Ostapowicz, R. J. Fontana, F. V. Schiødt, A. Larson, T. J. Davern, S. H. Han, T. M. McCashland, A. O. Shakil, J. E. Hay, L. Hynan *et al.*, "Results of a prospective study of acute liver failure at 17 tertiary care centers in the United States," *Annals of Internal Medicine*, vol. 137, no. 12, pp. 947–954, 2002.

[3] J. R. Senior, "Drug hepatotoxicity from a regulatory perspective," *Clinics in Liver Disease*, vol. 11, no. 3, pp. 507–524, 2007.

[4] G. A. Kullak-Ublick, R. J. Andrade, M. Merz, P. End, A. Benesic, A. L. Gerbes, and G. P. Aithal, "Drug-induced liver injury: Recent advances in diagnosis and risk assessment," *Gut*, vol. 66, no. 6, pp. 1154–1164, 2017.

[5] L. Kuna, I. Bozic, T. Kizivat, K. Bojanic, M. Mrso, E. Kralj, R. Smolic, G. Y. Wu, and M. Smolic, "Models of drug induced liver injury (DILI)–current issues and future perspectives," *Current Drug Metabolism*, vol. 19, no. 10, pp. 830–838, 2018.

[6] T.-H. Nguyen-Vo, L. Nguyen, N. Do, P. H. Le, T.-N. Nguyen, B. P. Nguyen, and L. Le, "Predicting drug-induced liver injury using convolutional neural network and molecular fingerprint-embedded features," *ACS Omega*, vol. 5, no. 39, pp. 25 432–25 439, 2020.

[7] S. U. Vorrink, Y. Zhou, M. Ingelman-Sundberg, and V. M. Lauschke, "Prediction of drug-induced hepatotoxicity using long-term stable primary hepatic 3D spheroid cultures in chemically defined conditions," *Toxicological Sciences*, vol. 163, no. 2, pp. 655–665, 2018.

[8] L. V. Allen Jr, "Dosage form design and development," *Clinical Therapeutics*, vol. 30, no. 11, pp. 2102–2111, 2008.

[9] L. A. Vernetti, A. Vogt, A. Gough, and D. L. Taylor, "Evolution of experimental models of the liver to predict human drug hepatotoxicity and efficacy," *Clinics in Liver Disease*, vol. 21, no. 1, pp. 197–214, 2017.

[10] M. B. Bracken, "Why animal studies are often poor predictors of human reactions to exposure," *Journal of the Royal Society of Medicine*, vol. 102, no. 3, pp. 120–122, 2009.

[11] E. Pérez Santín, R. Rodríguez Solana, M. González García, M. D. M. García Suárez, G. D. Blanco Díaz, M. D. Cima Cabal, J. M. Moreno Rojas, and J. I. López Sánchez, "Toxicity prediction based on artificial intelligence: A multidisciplinary overview," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 11, no. 5, p. e1516, 2021.

[12] V. Özdemir, "Phenomics 2.0: Real-world real-time patient outcomes measured by the internet of pharmaceutical things," *OMICS: A Journal of Integrative Biology*, vol. 24, no. 3, pp. 119–121, 2020.

[13] A. K. Bashir, N. Victor, S. Bhattacharya, T. Huynh-The, R. Chengoden, G. Yenduri, P. K. R. Maddikunta, Q.-V. Pham, T. R. Gadekallu, and M. Liyanage, "Federated learning for the healthcare metaverse: Concepts, applications, challenges, and future directions," *IEEE Internet of Things Journal*, vol. 10, no. 24, pp. 21 873–21 891, 2023.

[14] Y. Sun, H. Tan, and Y. Chen, "A Bio-Nano systems interconnection hierarchical network model for targeted drug delivery," *IEEE Internet of Things Journal*, vol. 12, no. 21, pp. 44 867–44 881, 2025.

[15] P. K. Bulasara and S. R. Sahoo, "A robust and secure drug delivery with single transmitter and dual symmetrical receivers in an internet of Bio-Nano things," *IEEE Internet of Things Journal*, vol. 11, no. 14, pp. 25 074–25 087, 2024.

[16] S. Liu, Z. Wang, and C.-M. Chen, "A blockchain-assisted drug management and authentication scheme in IoMT," *IEEE Internet of Things Journal*, vol. 12, no. 16, pp. 34 283–34 296, 2025.

[17] W. Wang, X. Jing, S. Jiang, and Y. Han, "Pill-MSVAE: A multi-scale variational autoencoder based on the internet of things for accurate automatic drug identification and monitoring," *IEEE Internet of Things Journal*, pp. 1–1, 2026.

[18] C. Lou, H. Yang, H. Deng, M. Huang, W. Li, G. Liu, P. W. Lee, and Y. Tang, "Chemical rules for optimization of chemical mutagenicity via matched molecular pairs analysis and machine learning methods," *Journal of Cheminformatics*, vol. 15, no. 1, p. 35, 2023.

[19] D. Fan, K. Xue, R. Zhang, W. Zhu, H. Zhang, J. Qi, Z. Zhu, Y. Wang, and P. Cui, "Application of interpretable machine learning models to improve the prediction performance of ionic liquids toxicity," *Science of The Total Environment*, vol. 908, p. 168168, 2024.

[20] H. P. Y. Duong, B. D. E. McNiven, O. A. Dobre, S. M. A. Rizvi, H. Shin, T. T. Nguyen, and T. Q. Duong, "Quantum machine learning for drug discovery: Taxonomy, research challenges, and the road ahead," *ACM Computing Surveys*, vol. 58, no. 8, pp. 1–36, 2026.

[21] J. Born, G. Markert, N. Janakarajan, T. B. Kimber, A. Volkamer, M. R. Martínez, and M. Manica, "Chemical representation learning for toxicity prediction," *Digital Discovery*, vol. 2, no. 3, pp. 674–691, 2023.

[22] J. C. Pereira, E. R. Caffarena, and C. N. Dos Santos, "Boosting docking-based virtual screening with deep learning," *Journal of Chemical Information and Modeling*, vol. 56, no. 12, pp. 2495–2506, 2016.

[23] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, and D. R. Koes, "Protein–ligand scoring with convolutional neural networks," *Journal of Chemical Information and Modeling*, vol. 57, no. 4, pp. 942–957, 2017.

[24] M. Hirohara, Y. Saito, Y. Koda, K. Sato, and Y. Sakakibara, "Convolutional neural network based on SMILES representation of compounds for detecting chemical motif," *BMC Bioinformatics*, vol. 19, pp. 83–94, 2018.

[25] T. Shi, Y. Yang, S. Huang, L. Chen, Z. Kuang, Y. Heng, and H. Mei, "Molecular image-based convolutional neural network for the prediction of ADMET properties," *Chemometrics and Intelligent Laboratory Systems*, vol. 194, p. 103853, 2019.

[26] W. Sun, W. Zheng, and A. Simeonov, "Drug discovery and development for rare genetic disorders," *American Journal of Medical Genetics Part A*, vol. 173, no. 9, pp. 2307–2322, 2017.

[27] R. Qureshi, M. Irfan, T. M. Gondal, S. Khan, J. Wu, M. U. Hadi, J. Heymach, X. Le, H. Yan, and T. Alam, "AI in drug discovery and its clinical relevance," *Heliyon*, vol. 9, no. 7, 2023.

[28] M. Schlander, K. Hernandez-Villafuerte, C.-Y. Cheng, J. Mestre-Ferrandiz, and M. Baumann, "How much does it cost to research and

develop a new drug? A systematic review and assessment," *Pharmacoeconomics*, vol. 39, pp. 1243–1269, 2021.

[29] A. C. Fisher, S. L. Lee, D. P. Harris, L. Buhse, S. Kozlowski, L. Yu, M. Kopcha, and J. Woodcock, "Advancing pharmaceutical quality: An overview of science and research in the US FDA's office of pharmaceutical quality," *International Journal of Pharmaceutics*, vol. 515, no. 1-2, pp. 390–402, 2016.

[30] J. F. Kadow, N. A. Meanwell, K. J. Eastman, K.-S. Yeung, and A. J. DelMonte, "Chemistry in the pharmaceutical industry," *Handbook of Industrial Chemistry and Biotechnology*, pp. 531–579, 2017.

[31] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve R&D productivity: The pharmaceutical industry's grand challenge," *Nature Reviews Drug Discovery*, vol. 9, no. 3, pp. 203–214, 2010.

[32] D. Sun, W. Gao, H. Hu, and S. Zhou, "Why 90% of clinical drug development fails and how to improve it?" *Acta Pharmaceutica Sinica B*, vol. 12, no. 7, pp. 3049–3062, 2022.

[33] S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, and M. J. Ahsan, "Machine learning in drug discovery: A review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 1947–1999, 2022.

[34] R. J. Weaver and J.-P. Valentin, "Today's challenges to de-risk and predict drug safety in human "mind-the-gap"," *Toxicological Sciences*, vol. 167, no. 2, pp. 307–321, 2019.

[35] S. Giri and A. Bader, "A low-cost, high-quality new drug discovery process using patient-derived induced pluripotent stem cells," *Drug Discovery Today*, vol. 20, no. 1, pp. 37–49, 2015.

[36] F. P. Guengerich, "Mechanisms of drug toxicity and relevance to pharmaceutical development," *Drug Metabolism and Pharmacokinetics*, vol. 26, no. 1, pp. 3–14, 2011.

[37] J. R. Mora, Y. Marrero-Ponce, C. R. García-Jacas, and A. Suarez Causado, "Ensemble models based on QuBiLS-MAS features and shallow learning for the prediction of drug-induced liver toxicity: Improving deep learning and traditional approaches," *Chemical Research in Toxicology*, vol. 33, no. 7, pp. 1855–1873, 2020.

[38] Y. Wang and X. Chen, "Joint decision-making model based on consensus modeling technology for the prediction of drug-induced liver injury," *Journal of Chemistry*, vol. 2021, no. 1, p. 2293871, 2021.

[39] C. Y. Liew, Y. C. Lim, and C. W. Yap, "Mixed learning algorithms and features ensemble in hepatotoxicity prediction," *Journal of Computer-aided Molecular Design*, vol. 25, pp. 855–871, 2011.

[40] G. Xiong, Z. Wu, J. Yi, L. Fu, Z. Yang, C. Hsieh, M. Yin, X. Zeng, C. Wu, A. Lu *et al.*, "ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties," *Nucleic acids research*, vol. 49, no. W1, pp. W5–W14, 2021.

[41] Y. Wei, S. Li, Z. Li, Z. Wan, and J. Lin, "Interpretable-ADMET: A web service for ADMET prediction and optimization based on deep neural representation," *Bioinformatics*, vol. 38, no. 10, pp. 2863–2871, 2022.

[42] V. Venkatraman, "FP-ADMET: A compendium of fingerprint-based ADMET prediction models," *Journal of Cheminformatics*, vol. 13, pp. 1–12, 2021.

[43] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[44] D. Hwang, M. Jeon, and J. Kang, "A drug-induced liver injury prediction model using transcriptional response data with graph neural network," in *Proc. 2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Feb. 2020, pp. 323–329.

[45] T. Li, W. Tong, R. Roberts, Z. Liu, and S. Thakkar, "DeepDILI: Deep learning-powered drug-induced liver injury prediction using model-level representation," *Chemical Research in Toxicology*, vol. 34, no. 2, pp. 550–565, 2020.

[46] Z. Chen, Y. Jiang, X. Zhang, R. Zheng, R. Qiu, Y. Sun, C. Zhao, and H. Shang, "ResNet18DNN: Prediction approach of drug-induced liver injury by deep neural network with ResNet18," *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab503, 2022.

[47] P. Banerjee, A. O. Eckert, A. K. Schrey, and R. Preissner, "ProTox-II: A webserver for the prediction of toxicity of chemicals," *Nucleic Acids Research*, vol. 46, no. W1, pp. W257–W263, 2018.

[48] Y. Xu, Z. Dai, F. Chen, S. Gao, J. Pei, and L. Lai, "Deep learning for drug-induced liver injury," *Journal of Chemical Information and Modeling*, vol. 55, no. 10, pp. 2085–2093, 2015.

[49] G. Landrum, "RDKit: Open-source cheminformatics," https://www.rdkit.org, 2016, accessed: March 2025.

[50] S. A. Wildman and G. M. Crippen, "Prediction of physicochemical parameters by atomic contributions," *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 5, pp. 868–873, 1999.

[51] P. Isnard and S. Lambert, "Estimating bioconcentration factors from octanol-water partition coefficient and aqueous solubility," *Chemosphere*, vol. 17, no. 1, pp. 21–34, 1988.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[53] X. Yang, J. Zhang, K. Yoshizoe, K. Terayama, and K. Tsuda, "ChemTS: An efficient python library for de novo molecular generation," *Science and Technology of Advanced Materials*, vol. 18, no. 1, pp. 972–976, 2017.

[54] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Central Science*, vol. 4, no. 2, pp. 268–276, 2018.

[55] J. H. Jensen, "A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space," *Chemical Science*, vol. 10, no. 12, pp. 3567–3572, 2019.

[56] J. Jiménez-Luna, M. Skalic, N. Weskamp, and G. Schneider, "Coloring molecules with explainable artificial intelligence for preclinical relevance assessment," *Journal of Chemical Information and Modeling*, vol. 61, no. 3, pp. 1083–1094, 2021.

[57] S. Riniker and G. A. Landrum, "Similarity maps-a visualization strategy for molecular fingerprints and machine-learning methods," *Journal of Cheminformatics*, vol. 5, pp. 1–7, 2013.

[58] A. V. Weisbrod, L. P. Burkhard, J. Arnot, O. Mekenyan, P. H. Howard, C. Russom, R. Boethling, Y. Sakuratani, T. Traas, T. Bridges *et al.*, "Workgroup report: Review of fish bioaccumulation databases used to identify persistent, bioaccumulative, toxic substances," *Environmental Health Perspectives*, vol. 115, no. 2, pp. 255–261, 2007.

[59] L. Zhao, F. Montanari, H. Heberle, and S. Schmidt, "Modeling bioconcentration factors in fish with explainable deep learning," *Artificial Intelligence in the Life Sciences*, vol. 2, p. 100047, 2022.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[61] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[62] M. Choudhury, M. Tanvir, M. A. Yousuf, N. Islam, and M. Z. Uddin, "Explainable AI-driven scalogram analysis and optimized transfer learning for sleep apnea detection with single-lead electrocardiograms," *Computers in Biology and Medicine*, vol. 187, p. 109769, 2025.

[63] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[64] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," *arXiv preprint arXiv:1803.05407*, 2018.

[65] Z. Chen, Z. WU, X. Shi, B. XU, N. Zhao, and Y. Qiao, "A study on model performance for ethanol precipitation process of lonicera japonica by NIR based on bagging-PLS and boosting-PLS algorithm," *Chinese Journal of Analytical Chemistry*, pp. 1679–1686, 2014.

[66] M. Chen, A. Suzuki, S. Thakkar, K. Yu, C. Hu, and W. Tong, "DILIrank: The largest reference drug list ranked by the risk for developing drug-induced liver injury in humans," *Drug Discovery Today*, vol. 21, no. 4, pp. 648–653, 2016.

[67] A. Quinton, P. Latry, and M. Biour, "Hepatox: Database on hepatotoxic drugs," *Gastroentérologie Clinique et Biologique*, vol. 17, no. 5 Pt 2, pp. H116–20, 1993.

[68] C. Zhang, F. Cheng, W. Li, G. Liu, P. W. Lee, and Y. Tang, "In silico prediction of drug induced liver toxicity using substructure pattern recognition method," *Molecular Informatics*, vol. 35, no. 3-4, pp. 136–144, 2016.

[69] E. Minerali, D. H. Foil, K. M. Zorn, T. R. Lane, and S. Ekins, "Comparing machine learning algorithms for predicting drug-induced liver injury (DILI)," *Molecular Pharmaceutics*, vol. 17, no. 7, pp. 2628–2637, 2020.

[70] S. Ekins, A. J. Williams, and J. J. Xu, "A predictive ligand-based bayesian model for human drug-induced liver injury," *Drug Metabolism and Disposition*, vol. 38, no. 12, pp. 2302–2308, 2010.

[71] H. Zhang, L. Ding, Y. Zou, S.-Q. Hu, H.-G. Huang, W.-B. Kong, and J. Zhang, "Predicting drug-induced liver injury in human with Naïve Bayes classifier approach," *Journal of Computer-aided Molecular Design*, vol. 30, pp. 889–898, 2016.

[72] M. Chen, H. Hong, H. Fang, R. Kelly, G. Zhou, J. Borlak, and W. Tong, "Quantitative structure-activity relationship models for predicting drug-induced liver injury based on FDA-approved drug labeling annotation and using a large collection of drugs," *Toxicological Sciences*, vol. 136, no. 1, pp. 242–249, 2013.

[73] H. Ai, W. Chen, L. Zhang, L. Huang, Z. Yin, H. Hu, Q. Zhao, J. Zhao, and H. Liu, "Predicting drug-induced liver injury using ensemble learning methods and molecular fingerprints," *Toxicological Sciences*, vol. 165, no. 1, pp. 100–107, 2018.