Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

SHAP-GUIDED GRAPH NEURAL NETWORKS FOR HEPATOTOXICITY PREDICTION USING MOLECULAR SUBSTRUCTURES AND CYTOTOXICITY PROFILES

Ahmed Mansour[1]*, Omar Saeed[1]

1. *Department of Pharmaceutical Sciences and AI Applications, Faculty of Pharmacy, Cairo University, Cairo, Egypt.*

ARTICLE INFO

Received:
24 November 2024
Received in revised form:
30 January 2025
Accepted:
05 February 2025
Available online:
28 February 2025

Keywords: Explainable AI, SHAP, Graph neural networks, Hepatotoxicity, Drug-induced liver injury, Molecular substructures

ABSTRACT

Hepatotoxicity is a leading cause of drug failure and post-market withdrawal, arising from complex interactions among chemical structure, metabolism, cellular stress, and host susceptibility, which makes predictive toxicology particularly challenging. Traditional computational models often focus on either chemical structure or assay-derived toxicity signatures, yet fail to integrate these sources into a single interpretable framework, limiting their utility for medicinal chemists, especially when models flag hepatotoxicity without identifying the responsible molecular substructures or ignore cytotoxicity profiles that reveal mitochondrial dysfunction, oxidative stress, or membrane damage. To address this, a SHAP-guided graph neural network has been proposed, combining molecular graph representations with in vitro cytotoxicity assay endpoints—including viability loss, ATP depletion, reactive oxygen species generation, and mitochondrial membrane potential disruption—while decomposing each prediction into atom- and substructure-level contributions. The model encodes molecular graphs through a graph attention network and cytotoxicity endpoints via a fully connected network, fusing these representations before generating a hepatotoxicity risk estimate. SHAP values are then computed over graph nodes and aggregated into chemically meaningful substructures, allowing local explanations to highlight reactive or stress-associated motifs and clarify whether cytotoxicity assay signals reinforce the structural alert. This approach not only predicts hepatotoxicity but also provides mechanistic insights, linking molecular substructures to cellular stress mechanisms, thereby supporting safer molecular design and offering toxicologists a transparent, actionable basis for evaluating liver injury risks.

To Cite This Article: Mansour A, Saeed O. SHAP-Guided Graph Neural Networks for Hepatotoxicity Prediction Using Molecular Substructures and Cytotoxicity Profiles. Pharmacophore. 2025;16(1):21-30. https://doi.org/10.51847/tFWErcXy9u

Introduction

Drug-induced liver injury remains a central challenge in drug development because liver toxicity can interrupt clinical progression, trigger regulatory concern, and complicate post-market safety management. Machine learning approaches to DILI have attempted to improve early hazard identification by learning from chemical structure, biological profiles, and curated toxicity labels, including Bayesian and diverse-predictor strategies for DILI modeling [1, 2]. However, traditional structural alerts may be too rigid for multifactorial liver injury because the same substructure can behave differently depending on molecular context, exposure, and metabolic activation. This creates a need for predictive systems that can move beyond binary alerts toward mechanistically useful explanations.

Graph neural networks have become increasingly important for molecular property prediction because they treat molecules as graphs rather than fixed fingerprints, allowing atoms and bonds to participate directly in learned representations [3]. Molecular message-passing and graph-based models have been applied broadly across chemical prediction tasks, including toxicity endpoints, because they can learn local and global structure–property relationships without requiring a fully hand-engineered descriptor set [4, 5]. Yet uninterpretable deep models remain difficult to adopt in safety-critical drug discovery, where medicinal chemists and toxicologists need to understand the structural basis of an alert before acting on it. Explainability methods for molecular graph models therefore have practical value only when they produce explanations that are chemically plausible and decision-relevant [6].

Hepatotoxicity is particularly difficult to predict because it can involve reactive metabolites, mitochondrial dysfunction, oxidative stress, immune activation, transporter interference, and broader cellular injury phenotypes. Structure-based DILI

Corresponding Author: Ahmed Mansour; Department of Pharmaceutical Sciences and AI Applications, Faculty of Pharmacy, Cairo University, Cairo, Egypt. E-mail: ahmed.mansour@gmail.com.

models have provided useful risk stratification, but they can miss biological evidence that emerges from in vitro cytotoxicity or transcriptomic profiling [7, 8]. Cytotoxicity assay readouts could complement molecular graphs by capturing cellular consequences such as ATP loss, membrane damage, or stress responses that are not obvious from structure alone. A hybrid structure-assay framework would therefore be expected to support a richer interpretation of hepatotoxicity liability than a structure-only model. **Figure 1** shows the need of EAI.

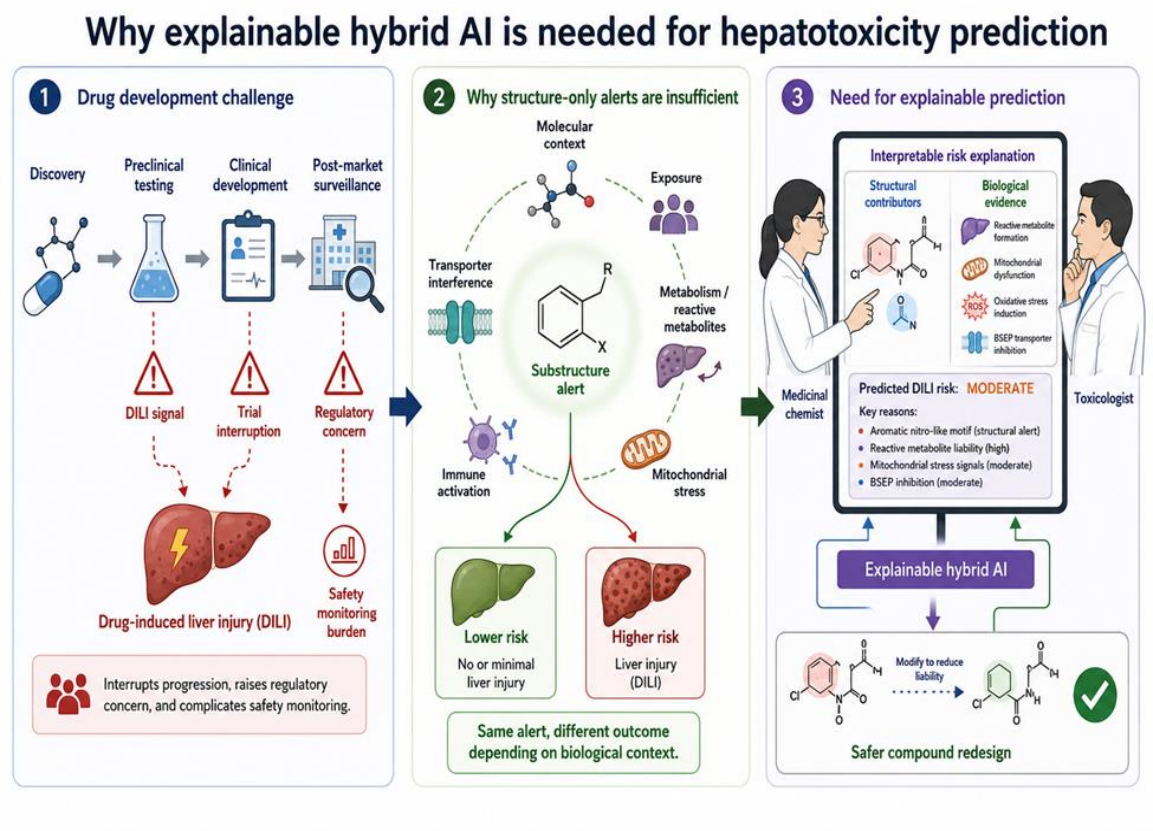

Figure 1. Rationale for explainable hybrid modeling in drug-induced liver injury prediction.
DILI risk can arise from interacting structural, metabolic, exposure-related, and cellular injury mechanisms, making rigid structural alerts insufficient for many drug-development decisions. Explainable hybrid AI models are motivated by the need to connect predicted hepatotoxicity risk with chemically and biologically interpretable evidence that can support safety assessment and medicinal chemistry redesign.

This article develops a conceptual framework for a SHAP-guided graph neural network that integrates molecular graphs with cytotoxicity profiles to predict hepatotoxicity and explain risk at the atom and substructure levels. SHAP provides an additive attribution perspective that can connect model output to input features, while graph explainability methods can localize influential atoms, bonds, or fragments within the molecular graph [9]. For molecular toxicity prediction, such explanations should be evaluated not only for mathematical fidelity but also for chemical plausibility and alignment with known toxicophores [10, 11]. The central thesis is that hepatotoxicity prediction becomes more actionable when the model can distinguish structural contributors from cytotoxicity-driven evidence and present both in a form suitable for medicinal chemistry redesign.

Background

Mechanisms of Drug-Induced Liver Injury

Drug-induced liver injury is commonly discussed in terms of intrinsic and idiosyncratic toxicity, but this distinction does not fully capture the mechanistic diversity of liver injury phenotypes. Computational DILI models have emphasized that toxicity can arise from chemical reactivity, biological susceptibility, and downstream stress responses rather than from a single structural determinant [12]. Reactive metabolites may damage proteins or organelles, mitochondrial impairment may compromise energetic homeostasis, and immune-mediated responses may amplify injury after exposure. These mechanisms make hepatotoxicity a suitable target for models that combine molecular structure with biological assay evidence.

Graph Neural Networks for Molecular Toxicity

Graph neural networks represent a molecule as a set of atoms and bonds, allowing message passing to learn how local chemical environments contribute to molecular-level properties. Benchmarking work in molecular machine learning has shown the importance of graph-based and learned molecular representations for toxicity and other chemical endpoints [3, 4]. More recent

graph toxicity studies have explored graph convolutional, message-passing, and equivariant variants for toxicological prediction, reflecting the field's movement from static descriptors toward structure-aware neural architectures [13-15]. For hepatotoxicity, this graph formulation is attractive because a liver injury alert may be driven by a localized motif embedded in a broader molecular context.

Cytotoxicity Profiling and In-Vitro Assays for Hepatotoxicity

Cytotoxicity profiling provides a biological layer that can help interpret whether a structural alert is associated with cellular stress. Transcriptomic and high-throughput profiling approaches for DILI prediction have shown how biological responses can complement chemical information in liver safety assessment [8]. Assay endpoints such as viability reduction, mitochondrial membrane disruption, ATP depletion, oxidative stress, and membrane leakage are conceptually relevant because they approximate mechanistic processes involved in hepatocellular injury. Integrating these endpoints with molecular graphs would allow a model to explain whether predicted hepatotoxicity is primarily structural, assay-driven, or supported by both evidence streams.

Post-Hoc Explainability for Graph Neural Networks

Post-hoc explainability methods are central to making molecular graph predictions usable by toxicologists because they can identify which atoms, bonds, or subgraphs influenced a prediction. SHAP frames explanation as additive feature attribution [9], while Integrated Gradients attributes model output to input features through a path-based sensitivity formulation [16]. GNNExplainer and related graph explainers identify compact subgraphs that preserve a model's prediction, offering a graph-native way to localize influential chemical regions [17, 18]. For molecular toxicology, these methods are valuable only if their highlighted substructures correspond to chemically interpretable toxicophores or plausible mechanistic drivers.

Prior Work on Interpretable Hepatotoxicity Prediction

Interpretable hepatotoxicity prediction has historically relied on structural alerts, QSAR descriptors, or feature importance analyses, but deep learning has expanded the space of possible representations. DeepDILI and related DILI models have shown how learned representations can support hepatotoxicity risk modeling, while interpretable DILI frameworks have explored attention and feature-importance mechanisms for model transparency [7, 19]. Graph-based hepatotoxicity models such as GeoDILI suggest that molecular geometry and graph representations can provide a richer basis for liver injury prediction [20]. The remaining gap is a unified framework that fuses molecular graphs with cytotoxicity profiles while producing substructure-level explanations suitable for medicinal chemistry action.

Model Development Overview

High-Level Predictive and Explanatory Pipeline

The proposed pipeline begins by ingesting a compound as a molecular graph and representing cytotoxicity assay outputs as global biological features. A graph neural network encodes the molecule, a parallel network encodes the cytotoxicity vector, and a fused prediction head estimates hepatotoxicity risk in conceptual terms rather than as a reported experimental result. SHAP is then applied to decompose the prediction into atom-level contributions, building on the additive explanation logic introduced for model interpretation [9]. The resulting explanation can be visualized on the molecular structure and summarized as substructure-level evidence for a toxicological decision.

Figure 2 illustrates the proposed SHAP-guided graph neural network workflow that integrates molecular graph learning, cytotoxicity profile encoding, hepatotoxicity risk prediction, and substructure-level explanation for drug safety interpretation.

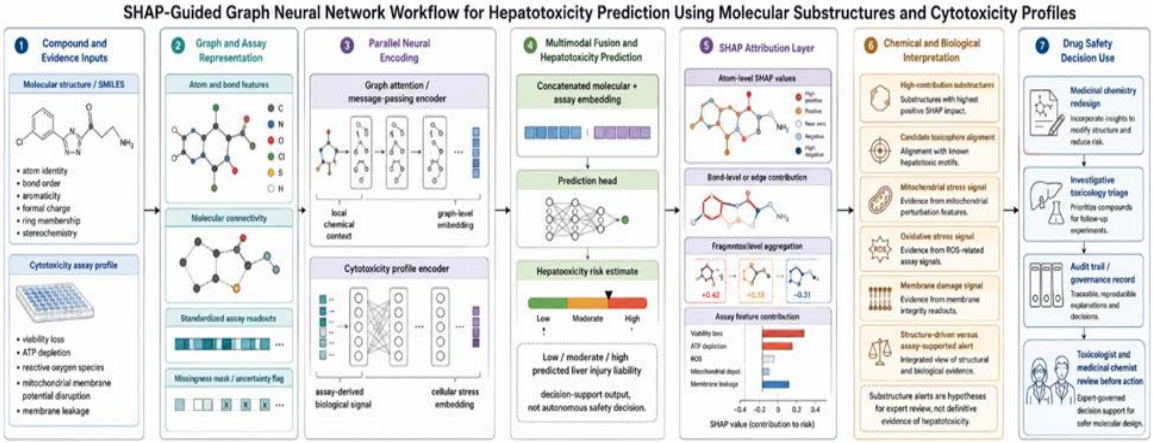

Figure 2. SHAP-Guided Graph Neural Network Workflow for Hepatotoxicity Prediction Using Molecular Substructures and Cytotoxicity Profiles

Core Input Representations

The molecular graph contains atom-level node features such as element identity, aromaticity, formal charge, hybridization, and ring membership, while bond features encode bond order, conjugation, and stereochemical context. Message-passing models for molecular property prediction support this representation because they allow the learned embedding of an atom to depend on its chemical neighborhood [4]. The cytotoxicity profile is represented as a fixed-length vector of assay readouts, including cell viability, ATP depletion, reactive oxygen species, lactate dehydrogenase release, and mitochondrial membrane potential. Missing assay entries can be masked or imputed conceptually, with the model designed to avoid treating missingness as evidence of safety.

Design Principles

The conceptual model is designed to be end-to-end trainable, explainable after training, and usable in settings where cytotoxicity information may be incomplete. Its graph component should preserve chemically intuitive locality, while its assay component should allow cellular stress evidence to influence the final prediction when such data are available. Chemistry-intuitive explanation methods for molecular graph models emphasize that substructure-level attributions should align with recognizable motifs rather than arbitrary atom clusters [10]. The design therefore prioritizes actionability: an explanation should tell a chemist which motif to reconsider and tell a toxicologist which biological evidence supports the alert.

Data Sources and Feature Engineering

Curation of Hepatotoxicity Datasets

A conceptual hepatotoxicity dataset would combine curated DILI labels from established resources, literature-derived annotations, and benchmark collections used in molecular toxicity modeling. MoleculeNet has influenced the organization of molecular benchmark tasks, while later toxicity studies have highlighted the value of standardized comparisons across chemical endpoints [3, 14]. DILI-specific modeling efforts demonstrate how curated hepatotoxicity labels can support machine learning but also reveal the need for careful endpoint definition and label harmonization [1, 2, 12, 21]. In this framework, curation would emphasize clinically meaningful hepatotoxic and non-hepatotoxic categories without reporting dataset sizes or artificial performance outcomes.

Building Molecular Graphs and Cytotoxicity Matrices

SMILES strings would be converted into molecular graphs whose nodes and edges encode chemically meaningful atom and bond properties, following the graph-based representation principles used in molecular deep learning [5]. Cytotoxicity assay results would be assembled into a matrix where each compound is associated with standardized biological readouts, and missing entries would be handled through masking strategies that preserve uncertainty. Hybrid models that combine chemical structure with biological response data are conceptually useful because they can relate structural risk to observed cellular consequences [8]. This feature-engineering step is therefore not merely technical; it defines whether the model can distinguish a structural toxicophore from a downstream cytotoxic phenotype.

Data Splitting to Avoid Data Leakage

Evaluation-oriented dataset splitting should prevent structurally similar analogues from appearing across development and held-out partitions, because random splits can overstate model generalizability in molecular prediction. Learned molecular representation studies have emphasized the importance of scaffold-aware evaluation when assessing whether a model generalizes beyond familiar chemical series [4]. For a hepatotoxicity model intended for prospective drug discovery use, scaffold-based splitting would better approximate the challenge of evaluating new chemotypes. Such splitting also makes explanations more meaningful because highlighted substructures are less likely to reflect memorized analogue series.

Shap-Guided Graph Neural Network Architecture

Graph Encoder and Toxicity Readout

The graph encoder would use a graph attention or message-passing architecture to learn atom embeddings that reflect local chemical context and longer-range molecular relationships. Graph convolutional and message-passing models have been used for toxicity prediction because they can learn molecular representations directly from atoms and bonds rather than relying entirely on predefined fingerprints [13, 22]. A graph-level readout would pool atom embeddings into a molecular representation, while a parallel feed-forward network would process the cytotoxicity profile before both embeddings are concatenated. The final prediction head would output a hepatotoxicity risk estimate intended for interpretation and prioritization, not as a standalone claim of clinical causality.

Table 1 defines how molecular graph inputs, cytotoxicity assay signals, neural encoding layers, and SHAP-derived explanations are organized into a unified hepatotoxicity prediction architecture.

Table 1. Multimodal Evidence Architecture for SHAP-Guided Hepatotoxicity Prediction

Model layer	Input or representation	Computational role	Toxicological meaning	Explanation output	Decision-use value
Molecular graph input	Atoms, bonds, aromaticity, formal charge, hybridization, ring membership, bond order, conjugation, stereochemical context	Converts a compound into a graph structure suitable for message passing	Represents local and global chemical environments that may contain hepatotoxic liability motifs	Atom- and bond-level contribution scores	Helps identify whether the predicted alert is linked to a specific chemical region rather than the molecule as a whole
Cytotoxicity profile input	Viability loss, ATP depletion, reactive oxygen species generation, mitochondrial membrane potential disruption, membrane leakage, assay missingness indicators	Encodes biological response signals as a structured assay vector	Captures cellular stress phenotypes that may support or qualify a structural alert	Feature-level SHAP values for assay endpoints	Helps distinguish a purely structure-driven prediction from one supported by observed cellular injury signals
Graph attention / message-passing encoder	Molecular graph with node and edge features	Learns context-aware atom embeddings through neighborhood aggregation	Models how local substructures behave within the full molecular scaffold	Node importance, edge importance, attention-informed subgraph relevance	Supports chemically localized interpretation of hepatotoxicity risk
Cytotoxicity neural encoder	Standardized cytotoxicity matrix with masking for missing assay values	Learns nonlinear relationships among biological stress endpoints	Represents convergent toxicity phenotypes such as mitochondrial injury, oxidative stress, or membrane disruption	Assay-specific contribution profile	Allows toxicologists to evaluate whether the model's prediction is biologically plausible
Fusion layer	Concatenated molecular graph embedding and cytotoxicity embedding	Integrates chemical structure with assay-derived biological evidence	Represents joint evidence for hepatotoxic liability	Relative contribution of structure branch versus assay branch	Clarifies whether risk is driven mainly by molecular features, assay signals, or their combination
Hepatotoxicity prediction head	Fused multimodal embedding	Produces hepatotoxicity risk estimate for screening or prioritization	Indicates predicted liver injury liability in a decision-support context	Model output decomposed into SHAP-attributed components	Enables ranking, triage, and expert review without treating the model as an autonomous safety authority
SHAP attribution layer	Trained model, molecular graph features, cytotoxicity features	Decomposes prediction into additive feature contributions	Links model output to chemically and biologically interpretable evidence	Atom-level, fragment-level, and assay-level SHAP values	Makes the model usable for medicinal chemistry redesign and investigative toxicology
Fragment aggregation layer	Atom-level SHAP scores mapped to functional groups, ring systems, or candidate toxicophores	Converts granular attributions into chemically meaningful units	Identifies interpretable molecular motifs associated with predicted risk	Ranked substructures with positive or negative contribution to hepatotoxicity prediction	Supports action-oriented discussion of which motif may require redesign or further testing
Governance and audit layer	Prediction record, input version, model version, SHAP metadata, expert decision, override rationale	Preserves traceability of model-derived alerts and human review	Documents how computational evidence was interpreted in safety decisions	Reviewable explanation report	Supports reproducibility, accountability, and regulated drug safety workflows

Integration of SHAP Values into the Training and Inference Pipeline

After model training, SHAP values would be computed for atom-level or fragment-level graph features to attribute the hepatotoxicity prediction to molecular substructures. EdgeSHAPer extends the Shapley-value idea to bond-centric graph explanations, illustrating how graph attributions can be adapted to molecular connectivity [23]. GNNShap and related scalable graph explanation approaches further support the use of Shapley-style reasoning for graph neural networks in settings where node and edge contributions must be estimated efficiently [24]. In the proposed hepatotoxicity framework, SHAP values would be used during inference to explain the model output rather than to claim that a highlighted atom is biologically causal by itself.

Aligning SHAP Values with Substructures

Atom-level SHAP values would be grouped into chemically interpretable fragments through substructure matching, allowing explanations to be reported in terms of functional groups, ring systems, or candidate toxicophores. Chemistry-intuitive molecular graph explanations and structural-alert localization studies show why fragment-level interpretation is more useful than isolated atom attribution for medicinal chemistry decisions [10, 11]. XGraphBoost and broader molecular explanation work also suggest that graph-derived features can be connected to interpretable molecular property signals when the explanation layer is designed around chemical structure [25]. In this architecture, high-SHAP fragments would be aligned with known alerts when possible and flagged as hypotheses for expert review when they represent unfamiliar motifs.

Linking Predictions to Toxicophores and Cytotoxic Mechanisms

Global Substructure Attribution

Across the training corpus, atom-level SHAP values could be aggregated into functional-group attributions to identify substructures repeatedly associated with hepatotoxicity. Toxicity-oriented graph neural networks and learned chemical representations provide a basis for discovering patterns that may not be captured by fixed structural-alert rules [15, 26]. Such global attribution would allow the model to rank candidate toxicophores by their repeated contribution to predicted liver injury risk. These ranked motifs should be interpreted as hypotheses for toxicological review rather than as automatic evidence of causality.

Local Explanation: Why Is This Particular Drug Predicted Toxic?

For an individual compound, the local explanation would highlight atoms and bonds that contributed most strongly to the predicted hepatotoxicity alert. Graph-specific explainers such as GNNExplainer and parameterized graph explainers support the idea that a compact molecular subgraph can preserve the prediction-relevant signal [17, 18]. In the proposed model, this subgraph would be displayed together with the cytotoxicity contribution, clarifying whether the alert arises mainly from molecular structure, assay evidence, or their combination. This would help a medicinal chemist distinguish a modifiable structural concern from a broader biological toxicity signature.

Integrating Cytotoxicity Evidence into Explanations

When assay readouts such as ATP depletion, reactive oxygen species generation, or membrane damage are important to a prediction, the explanation should explicitly describe their role rather than presenting the alert as structure-only. Transcriptomic and model-level DILI studies show that biological-response information can enrich hepatotoxicity prediction beyond conventional chemical descriptors [7, 8]. A high-risk structural fragment accompanied by mitochondrial stress evidence would be interpreted differently from the same fragment in a biologically silent profile. This integrated explanation would support mechanistic triage by connecting substructure-level risk to cellular injury phenotypes.

Counterfactual Explanations for Structure-Toxicity Relationships

Counterfactual analysis would allow the model to evaluate hypothetical structural modifications and ask whether a high-attribution motif could be replaced by a less concerning alternative. Molecular graph models designed for property prediction can support such local perturbation reasoning because changes to atoms or bonds can be propagated through the learned representation [4, 25]. In hepatotoxicity screening, a counterfactual explanation might compare a liability-associated heteroaromatic motif with a redesigned scaffold while preserving the intended pharmacophore conceptually. Such outputs should be treated as decision support for chemical redesign, not as proof that the proposed analogue would be safe.

Explainability Methods for Toxicologists and Medicinal Chemists

Interactive Visualisation of SHAP on Molecular Structure

An interactive visualization layer would map SHAP values onto a two-dimensional molecular structure so that users can inspect atom-level and fragment-level contributions. Chemistry-intuitive explanation methods for molecular graph models emphasize that visual explanations should preserve recognizable chemical motifs rather than produce disconnected attribution noise [10]. The interface could allow users to switch between structural attributions and cytotoxicity feature contributions, making the evidence trail easier to review. This would help safety scientists compare a model-derived alert with their own mechanistic expectations.

Narrative Explanation Generation

Narrative explanation generation would translate attribution outputs into short, chemically grounded statements that describe why a compound is predicted to be hepatotoxic. A SHAP-based explanation could state that a particular functional group contributes positively to predicted risk and that this signal is supported by mitochondrial depolarization or oxidative stress evidence, following the additive attribution logic of SHAP [9]. This prose layer would be especially useful for multidisciplinary safety meetings, where not all participants examine atom-level heatmaps directly. The narrative should avoid overstating causality and should present the explanation as model-derived evidence requiring expert interpretation.

Benchmarking Against Known Toxicophores

SHAP-derived substructure alerts should be compared against expert-curated toxicophores and structural alerts to evaluate whether the model recovers chemically plausible risk motifs. Structural-alert localization work shows that interpreting toxicity models through known motifs can reveal both concordant alerts and suspicious attribution patterns [11]. When the model highlights a known toxicophore, the explanation gains plausibility; when it highlights a novel motif, the result should be reviewed as a hypothesis for further investigation. This benchmarking would make the explanation layer more credible to medicinal chemists and toxicologists.

Audit Trail and Model Governance

Every prediction and explanation should be stored with the molecular input, cytotoxicity profile, highlighted substructures, and explanation metadata to support model governance. Interpretable DILI models such as InterDILI demonstrate the importance of making prediction rationales visible rather than treating the model as a black box [19]. In a regulated or safety-critical environment, an audit trail would allow teams to review why a compound was deprioritized, overridden, or advanced despite an alert. Expert feedback from such reviews could then guide model refinement and help identify explanation failures.

Integration Into Drug Safety Workflows

Early-Stage Hazard Identification

In early discovery, the model could be applied to virtual libraries or hit-to-lead series to identify compounds that may carry hepatotoxicity liabilities before extensive experimental investment. Graph-based toxicity prediction models and toxicity benchmarks support the use of molecular learning systems as prioritization tools in early chemical design [14, 27]. The SHAP layer would make this prioritization more actionable by showing which substructures are associated with the predicted risk. Medicinal chemists could then use the explanation to decide whether to redesign, deprioritize, or request targeted follow-up assays.

Supporting Investigative Toxicology and Regulatory Submissions

When liver injury signals emerge during preclinical development, an explainable model could help investigative toxicologists identify whether the predicted risk is linked to a structural motif, a cytotoxicity pattern, or both. Hepatotoxicity knowledgebases and graph neural network-based prediction resources illustrate how curated DILI evidence can support structured review of liver safety signals [20]. In regulatory-facing contexts, such explanations would not replace experimental evidence but could help organize a mechanistic rationale for risk assessment. The model would therefore function as a transparent decision-support tool rather than as an autonomous safety decision-maker.

Evaluation Strategy

Predictive Performance

Predictive evaluation should use accepted classification metrics such as discrimination, sensitivity, specificity, and balanced accuracy, but results should be interpreted conceptually and not presented as definitive performance claims without empirical validation. Scaffold-split validation is important because molecular benchmarks have shown that random splits can make models appear more generalizable than they would be for new chemotypes [3, 4]. External hepatotoxicity collections would be useful for assessing whether the model transfers beyond the data used to develop it. The evaluation should ask whether the fused graph-assay model would be expected to generalize better than structure-only or assay-only baselines.

Table 2 provides a structured evaluation framework for assessing whether the proposed model is accurate, explainable, chemically plausible, and useful in real-world hepatotoxicity decision workflows.

Table 2. Evaluation Framework for Predictive Accuracy, Explanation Fidelity, Chemical Plausibility, and Workflow Utility

Evaluation domain	Key question	Recommended assessment approach	Strong evidence of model value	Failure mode to detect	Relevance to drug safety decision-making
Predictive discrimination	Can the fused model distinguish hepatotoxic from non-hepatotoxic compounds?	Compare AUROC, AUPRC, sensitivity, specificity, balanced accuracy, and calibration across scaffold-based and external validation sets	Fused graph-assay model outperforms structure-only and assay-only baselines under scaffold-aware evaluation	High apparent accuracy under random splits but weak generalization to new scaffolds	Determines whether the model is suitable for prospective screening rather than retrospective memorization
Scaffold generalization	Does the model transfer to unfamiliar chemical series?	Use scaffold split, time split, or external chemotype holdout validation	Stable performance across structurally distinct compounds	Inflated performance caused by analogue leakage	Protects against misleading confidence when screening novel drug-like compounds
Branch contribution validity	Are molecular and cytotoxicity branches contributing meaningfully?	Conduct ablation studies removing graph input, assay input, or missingness indicators	Risk estimates change coherently when relevant evidence streams are removed	Model relies excessively on one branch or treats missing assay values as safety evidence	Clarifies when predictions are chemically driven, biologically driven, or jointly supported

SHAP explanation fidelity	Do highlighted atoms, fragments, and assay features genuinely influence the prediction?	Perform perturbation, masking, deletion, and counterfactual tests on top-attributed graph regions and assay variables	Removing high-SHAP fragments or endpoints changes predicted risk in the expected direction	Explanations are visually persuasive but weakly tied to model behavior	Ensures that explanations are faithful to the model rather than decorative post-hoc rationales
Chemical plausibility	Do high-attribution fragments correspond to recognizable toxicophores or plausible metabolic liabilities?	Expert review by medicinal chemists and toxicologists; comparison with known structural alerts and toxicophore libraries	Highlighted motifs align with established or mechanistically plausible hepatotoxicity concerns	Attribution highlights chemically irrelevant or disconnected atom clusters	Determines whether explanations are usable for medicinal chemistry redesign
Biological plausibility	Do cytotoxicity attributions align with liver injury mechanisms?	Compare assay SHAP values with expected patterns of mitochondrial dysfunction, oxidative stress, ATP depletion, or membrane damage	Structural alerts are strengthened when supported by coherent cellular stress evidence	Assay contributions contradict known biology or overemphasize noisy endpoints	Helps toxicologists interpret whether a structural alert is supported by functional evidence
Local explanation utility	Does the explanation help users understand why a specific compound is predicted risky?	User evaluation using compound case reviews with and without SHAP explanation reports	Users can identify actionable substructures, supporting assay signals, and uncertainty sources	Explanations are too technical, unstable, or non-actionable for review meetings	Measures practical interpretability rather than only mathematical explainability
Counterfactual coherence	Do proposed structural modifications reduce predicted risk for chemically sensible reasons?	Generate or test local structural perturbations while preserving core pharmacophore context conceptually	Risk decreases when liability-associated motifs are modified without producing implausible chemistry	Counterfactuals suggest unrealistic or pharmacologically destructive modifications	Supports safer analogue design while avoiding overinterpretation of model suggestions
Uncertainty and missing-data handling	Does the model communicate when evidence is incomplete?	Compare predictions with observed assay data, missing assay data, and imputed assay values	Explanation reports clearly distinguish measured, missing, and inferred biological evidence	Imputed or absent assay signals are presented as if experimentally observed	Prevents false confidence during early virtual screening or incomplete profiling
Governance readiness	Can predictions and explanations be traced, reviewed, and audited?	Store input molecule, assay vector, model version, SHAP output, expert review, decision, and override rationale	Each model-derived alert has a reproducible explanation and human decision record	Predictions influence decisions without traceability or expert accountability	Supports transparent use of explainable AI in safety-critical drug development settings
Workflow impact	Does the model improve safety triage and medicinal chemistry discussion?	Prospective workflow comparison of design meetings with versus without model explanations	Teams make clearer, more consistent, and better-documented redesign or follow-up testing decisions	Model adds complexity without improving interpretability or decision quality	Evaluates whether explainable AI contributes real operational value in drug discovery

Explanation Quality and Chemical Plausibility

Explanation quality should be assessed through both model-fidelity tests and expert review by toxicologists and medicinal chemists. Integrated Gradients, SHAP, GNNExplainer, and bond-centric Shapley methods provide complementary ways to test whether highlighted atoms, bonds, or subgraphs are genuinely influential for the prediction [16, 17, 23]. Chemical plausibility should be evaluated by asking whether highlighted fragments correspond to known toxicophores, plausible metabolic liabilities, or assay-supported mechanisms. Explanation evaluation should also include stress tests in which top-attributed fragments are masked or modified to examine whether the predicted risk changes in a chemically coherent direction.

Utility in a Real-World Medicinal Chemistry Setting

Real-world utility should be evaluated by observing whether the model helps teams make clearer and more consistent safety-related design decisions. Recent graph-based hepatotoxicity resources and augmented graph-feature approaches suggest that molecular graph models can be positioned within practical DILI prediction workflows, but their value depends on interpretability and integration into decision processes. A prospective workflow evaluation could compare how teams reason about compounds with and without SHAP-guided explanations, while avoiding unsupported claims about downstream outcomes. The central question is whether explanations improve the quality of discussion, prioritization, and redesign hypotheses in live discovery settings.

Limitations

Metabolism-Dependent Toxicity and Data Gaps

A key limitation is that the proposed model does not explicitly encode metabolic pathways or metabolite structures, so metabolism-dependent hepatotoxicity may be captured only indirectly. DILI prediction studies repeatedly show that chemical structure alone is not sufficient to fully represent liver injury mechanisms, especially when reactive metabolites or host-specific responses are involved [1, 12]. Cytotoxicity profiles may partially reflect downstream stress, but they cannot guarantee coverage of bioactivation mechanisms absent from the assay system. The model's explanations should therefore be framed as evidence from available inputs, not as complete mechanistic accounts of liver injury.

Reliance on Pre-Existing Assay Data

The cytotoxicity branch depends on experimental assay readouts that may be unavailable during early virtual screening. Although structure-based models and graph neural networks can operate from molecular graphs alone, adding predicted assay values would introduce a second layer of uncertainty that must be clearly represented in the explanation [5, 22]. A practical deployment should therefore distinguish predictions made with observed cytotoxicity data from predictions made with missing or inferred assay inputs. This distinction is essential because an explanation based mainly on imputed cytotoxicity evidence should carry less confidence than one grounded in measured biological responses.

Conclusion

A SHAP-guided graph neural network for hepatotoxicity prediction could integrate molecular substructures and cytotoxicity profiles into a single explainable decision-support framework. By combining graph-based molecular representation with assay-derived biological evidence, the approach would address both structural and cellular dimensions of liver injury risk. The resulting model would be designed not merely to flag compounds, but to explain the molecular and biological basis of those alerts.

The main strength of the proposed framework is its ability to connect atom-level explanation with practical medicinal chemistry interpretation. A highlighted substructure can guide redesign, while cytotoxicity evidence can indicate whether the concern is supported by cellular stress phenotypes. This dual explanation would help toxicologists distinguish structural liability, biological response, and their possible interaction.

Important challenges remain before such a framework could be relied upon in prospective drug discovery. Metabolism-dependent toxicity, incomplete cytotoxicity coverage, assay variability, and domain shift across chemical series may all affect prediction and explanation quality. Prospective validation in live discovery programs would be needed to determine whether the explanations improve safety decision-making.

Future work should focus on collaborative benchmarking, transparent reporting standards, and integration of explainable toxicity models into open drug discovery platforms. Shared evaluation protocols would help determine whether SHAP-guided molecular explanations are faithful, chemically plausible, and useful in practice. A community-driven approach could make explainable hepatotoxicity prediction more accessible and support safer molecular design across academic, industrial, and regulatory settings.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Williams DP, Lazic SE, Foster AJ, Semenova E, Morgan P. Predicting drug-induced liver injury with Bayesian machine learning. Chem Res Toxicol. 2019;33(1):239-48.
2. Adeluwa T, McGregor BA, Guo K, Hur J. Predicting drug-induced liver injury using machine learning on a diverse set of predictors. Front Pharmacol. 2021;12:648805.
3. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30.
4. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88.
5. Sun M, Zhao S, Gilvary C, Elemento O, Zhou J, Wang F. Graph convolutional networks for computational drug development and discovery. Brief Bioinform. 2020;21(3):919-35.
6. Wellawatte GP, Gandhi HA, Seshadri A, White AD. A perspective on explanations of molecular prediction models. J Chem Theory Comput. 2023;19(8):2149-60.

7. Li T, Tong W, Roberts R, Liu Z, Thakkar S. DeepDILI: deep learning-powered drug-induced liver injury prediction using model-level representation. Chem Res Toxicol. 2020;34(2):550-65.
8. Li T, Tong W, Roberts R, Liu Z, Thakkar S. Deep learning on high-throughput transcriptomics to predict drug-induced liver injury. Front Bioeng Biotechnol. 2020;8:562677.
9. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
10. Wu Z, Wang J, Du H, Jiang D, Kang Y, Li D, et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. Nat Commun. 2023;14(1):2585.
11. Walter M, Webb SJ, Gillet VJ. Interpreting neural network models for toxicity prediction by extracting learned chemical features. J Chem Inf Model. 2024;64(9):3670-88.
12. Liu A, Walter M, Wright P, Bartosik A, Dolciami D, Elbasir A, et al. Prediction and mechanistic analysis of drug-induced liver injury (DILI) based on chemical structure. Biol Direct. 2021;16(1):6.
13. Chen J, Si YW, Un CW, Siu SW. Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network. J Cheminform. 2021;13(1):93.
14. Ketkar R, Liu Y, Wang H, Tian H. A benchmark study of graph models for molecular acute toxicity prediction. Int J Mol Sci. 2023;24(15):11966.
15. Cremer J, Medrano Sandonas L, Tkatchenko A, Clevert DA, De Fabritiis G. Equivariant graph neural networks for toxicity prediction. Chem Res Toxicol. 2023;36(10):1561-73.
16. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML 2017). PMLR; 2017. p. 3319-28.
17. Ying Z, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: Generating explanations for graph neural networks. Adv Neural Inf Process Syst. 2019;32.
18. Luo D, Cheng W, Xu D, Yu W, Zong B, Chen H, et al. Parameterized explainer for graph neural network. Adv Neural Inf Process Syst. 2020;33:19620-31.
19. Lee S, Yoo S. InterDILI: interpretable prediction of drug-induced liver injury through permutation feature importance and attention mechanism. J Cheminform. 2024;16(1):1.
20. Wu W, Qian J, Liang C, Yang J, Ge G, Zhou Q, et al. GeoDILI: a robust and interpretable model for drug-induced liver injury prediction using graph neural network-based molecular geometric representation. Chem Res Toxicol. 2023;36(11):1717-30.
21. Minerali E, Foil DH, Zorn KM, Lane TR, Ekins S. Comparing machine learning algorithms for predicting drug-induced liver injury (DILI). Mol Pharm. 2020;17(7):2628-37.
22. Zhou Y, Ning C, Tan Y, Li Y, Wang J, Shu Y, et al. ToxMPNN: A deep learning model for small molecule toxicity prediction. J Appl Toxicol. 2024;44(7):953-64.
23. Mastropietro A, Pasculli G, Feldmann C, Rodríguez-Pérez R, Bajorath J. EdgeSHAPer: Bond-centric Shapley value-based explanation method for graph neural networks. iScience. 2022;25(10).
24. Akkas S, Azad A. GNNShap: scalable and accurate GNN explanation using Shapley values. In: Proceedings of the ACM Web Conference 2024 (WWW '24). New York: Association for Computing Machinery; 2024. p. 827-38.
25. Deng D, Chen X, Zhang R, Lei Z, Wang X, Zhou F. XGraphBoost: extracting graph neural network-based features for a better prediction of molecular properties. J Chem Inf Model. 2021;61(6):2697-705.
26. Born J, Markert G, Janakarajan N, Kimber TB, Volkamer A, Martínez MR, et al. Chemical representation learning for toxicity prediction. Digit Discov. 2023;2(3):674-91.
27. Monem S, Abdel-Hamid AH, Hassanien AE. Drug toxicity prediction model based on enhanced graph neural network. Comput Biol Med. 2025;185:109614.