Pharmacophore

ISSN-2229-5402

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

EXPLAINABLE DEEP KERNEL MODELS FOR LIVER INJURY PREDICTION USING MITOCHONDRIAL AND TRANSPORTER TOXICITY DATA

Andrei Popescu[1]*, Mihai Ionescu[1], Elena Stan[2], Sorin Dumitrescu[1], Irina Pavel[2]

1. ***Department of AI-Based Pharmaceutical Sciences, Faculty of Pharmacy, University of Bucharest, Bucharest, Romania.***
2. ***Department of Computational Drug Analytics, Faculty of Engineering, Politehnica University of Bucharest, Bucharest, Romania.***

ARTICLE INFO

Received:
21 January 2026
Received in revised form:
09 April 2026
Accepted:
12 April 2026
Available online:
28 April 2026

Keywords: Explainable artificial intelligence, Deep kernel learning, Drug-induced liver injury, Mitochondrial toxicity, BSEP inhibition, Transporter toxicity

ABSTRACT

Drug-induced liver injury (DILI) remains a significant safety concern in drug discovery and clinical development, often arising from mitochondrial dysfunction, impaired hepatobiliary transport, reactive metabolite formation, or combinations of these biological stressors. Many computational models for predicting liver injury rely primarily on chemical structure or clinical labels, providing limited insight into the underlying biological mechanisms. While neural networks can capture complex patterns, they often lack mechanistic transparency for safety pharmacologists. To address this gap, we propose an explainable deep kernel modeling framework that integrates mitochondrial toxicity endpoints, transporter inhibition profiles, and molecular structure features to generate interpretable, feature-attributed predictions. In this approach, a deep kernel learning model embeds a neural representation within a Gaussian process, linking mechanistic assay features to DILI risk labels, and explanation methods such as SHAP or integrated gradients are applied to decompose each prediction into contributions from specific assays, transporter signals, and structural features. Conceptually, the model would flag a drug candidate as higher risk when transporter inhibition and mitochondrial impairment jointly indicate hepatotoxicity, while the explanation layer identifies whether the risk is driven primarily by BSEP inhibition, reduced mitochondrial respiration, ATP depletion, or structural alerts. By connecting predicted liver injury risk to interpretable biological drivers, this explainable deep kernel framework can support transparent, mechanism-informed safety assessment and help medicinal chemists and safety scientists prioritize de-risking strategies earlier in development.

To Cite This Article: Popescu A, Ionescu M, Stan E, Dumitrescu S, Pavel I. Explainable Deep Kernel Models for Liver Injury Prediction Using Mitochondrial and Transporter Toxicity Data. Pharmacophore. 2026;17(2):93-102. https://doi.org/10.51847/rmfJrSUlIT

Introduction

Drug-induced liver injury is a persistent clinical and regulatory challenge because it can emerge from diverse biological pathways and may only become evident after substantial development investment. Bayesian and machine learning approaches have been proposed to support earlier DILI prediction, but many models remain limited by their dependence on chemical descriptors or curated clinical labels alone [1]. Recent DILI prediction studies increasingly recognize that structure-based models can support prioritization, yet they may fail to reveal whether a high-risk prediction reflects mitochondrial stress, transporter inhibition, or another mechanistic signal [2]. This creates a gap between statistical prediction and the mechanistic reasoning needed for safety pharmacology decisions.

Mechanistic in-vitro assays offer a way to move beyond purely structural prediction by measuring biological processes that are plausibly linked to hepatotoxicity. Studies of mitochondrial toxicity have emphasized endpoints such as mitochondrial membrane potential, oxygen consumption, and ATP depletion as informative indicators of cellular energy stress [3]. Transporter-focused work has also shown that inhibition of bile salt export pump and related hepatic transporters can contribute to cholestatic injury risk, making transporter panels a valuable complement to mitochondrial readouts [4]. However, these assay modalities are often heterogeneous, incomplete, and difficult to integrate into a single interpretable model. **Figure 1**

Corresponding Author: Andrei Popescu; Department of AI-Based Pharmaceutical Sciences, Faculty of Pharmacy, University of Bucharest, Bucharest, Romania. E-mail: andrei.popescu@gmail.com.

illustrates how heterogeneous mitochondrial toxicity and hepatic transporter assay signals can be organized into an interpretable mechanistic evidence field for hepatotoxicity risk modeling.

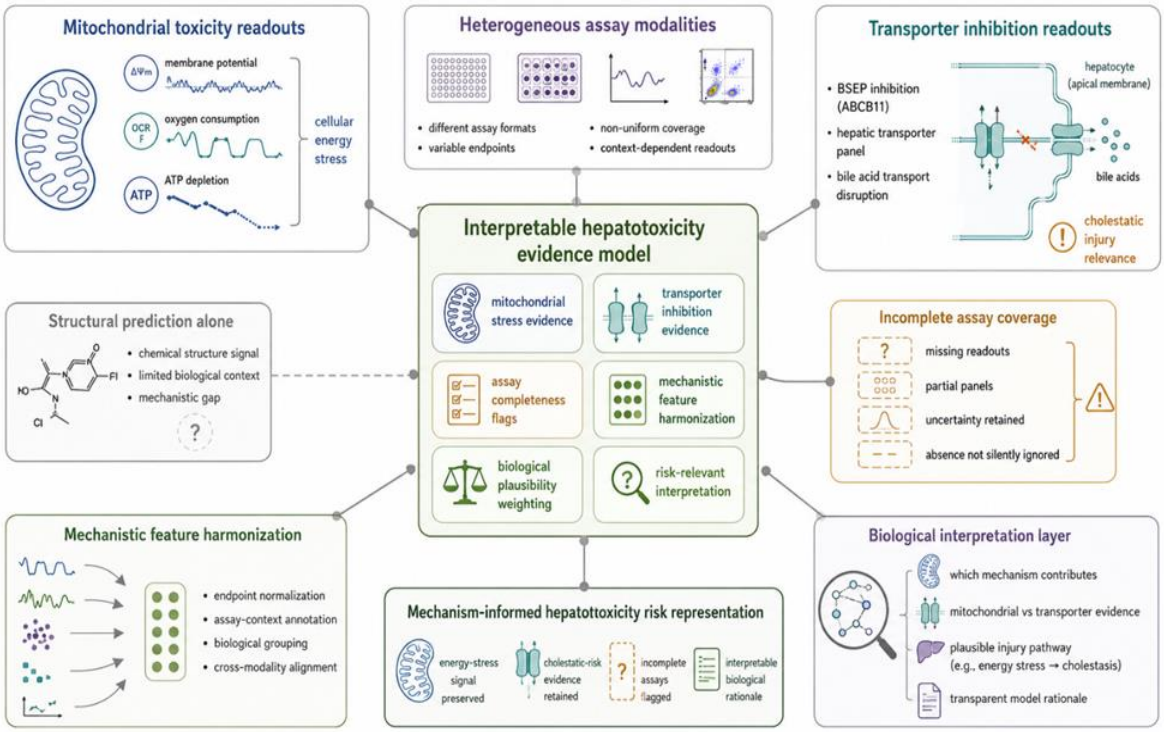

Figure 1. Mechanistic In-Vitro Assay Integration for Interpretable Hepatotoxicity Risk Modeling

Deep kernel models are attractive for this problem because they combine the representation power of neural networks with the uncertainty-aware inference of Gaussian processes. Probabilistic molecular modeling has shown that uncertainty quantification is especially important in low-data chemical settings, where predictions should communicate both expected risk and confidence [5]. Deep kernel learning has also been applied to molecular prediction tasks where learned representations improve flexibility while preserving a probabilistic modeling layer [6]. When paired with explainability methods such as SHAP, these models could make high-dimensional predictions more useful for scientific interpretation rather than only classification [7].
The thesis of this article is that an explainable deep kernel model could provide a mechanism-informed framework for DILI prediction by fusing mitochondrial toxicity, transporter inhibition, and molecular structure features. Hybrid models that integrate predicted or measured in-vitro signals with chemical information have already been proposed as a way to improve DILI assessment beyond single-modality modeling [8]. A deep kernel architecture would extend this idea by adding uncertainty estimates and feature-attributed explanations, allowing a safety scientist to interpret both how confident the model is and why it generated a risk signal [9]. Such a framework is best viewed as a decision-support tool rather than a replacement for experimental safety evaluation.

Background

Mechanisms of Drug-Induced Liver Injury

Drug-induced liver injury can be intrinsic, where toxicity is dose-related and mechanistically reproducible, or idiosyncratic, where host susceptibility and complex biology influence clinical expression. Mitochondrial impairment is a central mechanistic concern because disruption of respiration, membrane potential, or ATP generation can compromise hepatocyte survival [10]. Reactive metabolite formation and hepatobiliary transporter inhibition may add further stress, particularly when impaired bile acid handling interacts with cellular energy dysfunction [11]. Computational models that treat DILI as a single label therefore risk obscuring the multiple mechanistic routes that can lead to the same clinical outcome.

In-Vitro Assays for Mitochondrial and Transporter Toxicity

In-vitro mitochondrial assays commonly evaluate changes in mitochondrial membrane potential, oxygen consumption, respiratory chain function, and ATP depletion, each of which can provide evidence of impaired hepatocellular bioenergetics. Machine learning studies of mitochondrial toxicity have shown how chemical descriptors and assay-derived features can be used to predict mitochondrial liabilities, although such models still require mechanistic interpretation to support decision-making [12]. Transporter assays, including BSEP inhibition and OATP-related profiling, are similarly relevant because disruption of bile acid and hepatic uptake transport can contribute to cholestatic and mixed DILI mechanisms [13]. A combined assay panel is therefore more mechanistically informative than either mitochondrial or transporter data alone.

Machine Learning and Deep Learning for DILI Prediction

Earlier DILI prediction models often emphasized molecular fingerprints, chemical descriptors, and curated toxicity labels, which can provide useful screening signals but may not fully capture biological mechanisms. Deep neural models based on molecular substructures have been explored for DILI prediction, showing the appeal of learned representations for complex chemical safety endpoints [14]. Other studies have compared machine learning algorithms across DILI datasets, highlighting both the promise of computational screening and the persistent difficulty of generalizing across sparse and uncertain labels [15]. More recent approaches seek to incorporate in-vitro or predicted mechanistic evidence, but interpretability remains essential if the model is to inform safety pharmacology rather than merely rank compounds [16].

Deep Kernel Learning and Probabilistic Neural Networks

Deep kernel learning conceptually uses a neural network to transform raw inputs into a latent representation on which a Gaussian process kernel operates. This allows the model to capture nonlinear interactions while still returning predictive uncertainty, a property that is valuable when chemical safety datasets are limited or heterogeneous [17]. Bayesian graph neural networks and related probabilistic molecular models illustrate how uncertainty-aware prediction can support more reliable decision-making in chemical property modeling [5]. In DILI prediction, this architecture could help distinguish between confident low-risk predictions and uncertain predictions caused by sparse assay coverage or chemical novelty.

Explainable AI for Safety Assessment

Explainable AI is essential in safety assessment because a high-risk prediction must be connected to mechanisms that scientists can evaluate experimentally. SHAP provides a general framework for attributing model predictions to individual features and has become a widely used method for interpreting complex machine learning models [7]. In hepatotoxicity modeling, explanation methods can connect risk signals to transporter inhibition, mitochondrial impairment, structural alerts, or physicochemical properties, thereby translating statistical outputs into safety hypotheses [3]. Such mechanistic attribution is particularly important when model outputs are used to prioritize follow-up assays or guide chemical redesign.

Model Development Overview

High-Level Prediction and Explanation Pipeline

The proposed pipeline begins by representing each drug candidate with mitochondrial assay readouts, transporter inhibition features, and molecular descriptors. A deep kernel model would map these inputs to a probabilistic DILI risk estimate, while the Gaussian process layer would express uncertainty around that estimate rather than returning only a point prediction [1]. SHAP or integrated-gradient explanations would then decompose the model output into assay-specific and structural contributions, making it possible to identify whether a risk signal is primarily biological, chemical, or mixed [7]. This design aligns with the broader movement toward models that combine prediction with mechanistic interpretation in DILI assessment. **Figure 2** illustrates the proposed mechanism-informed deep kernel architecture linking mitochondrial toxicity endpoints, transporter inhibition profiles, and molecular structure features to probabilistic DILI risk, interpretable feature attribution, counterfactual de-risking, and human-governed safety decisions.

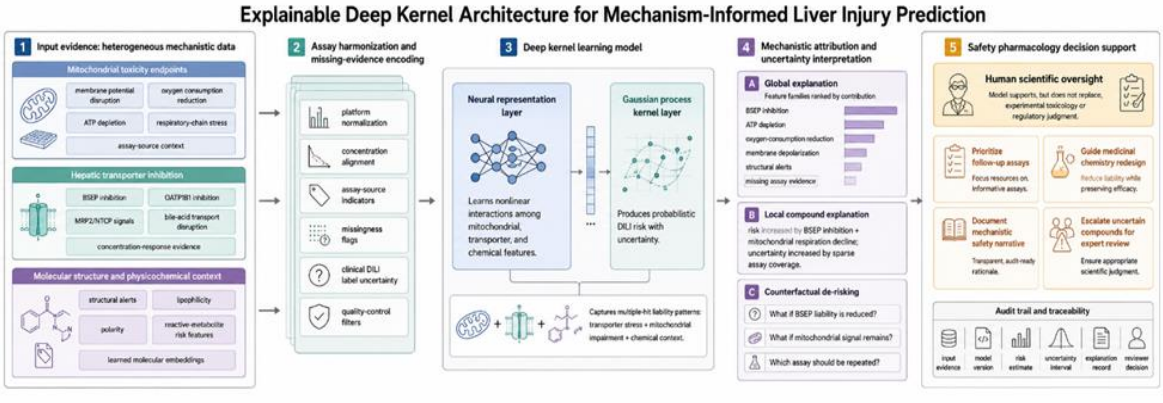

Figure 2. Explainable Deep Kernel Architecture for Mechanism-Informed Liver Injury Prediction

Core Input Features

The mitochondrial feature block would include conceptually standardized indicators of membrane potential disruption, oxygen consumption change, ATP depletion, and related bioenergetic stress. Mitochondrial toxicity modeling studies support the idea that such endpoints can be encoded as predictive features for chemical safety assessment [18]. The transporter block would include BSEP, OATP1B1, MRP2, NTCP, and related inhibition signals, reflecting evidence that transporter inhibition can contribute to liver injury mechanisms [4]. Molecular descriptors such as lipophilicity, polarity, structural alerts, and learned chemical embeddings would provide an additional representation of chemical liability, complementing the mechanistic assay data [19].

Design Principles

The model should be probabilistic, explainable, and tolerant of incomplete mechanistic assay panels because early discovery data are rarely uniform across compounds. Gaussian process modeling is well suited to this principle because uncertainty can increase when assay evidence is sparse, conflicting, or chemically distant from known examples [9]. The neural representation layer would allow nonlinear interactions among mitochondrial toxicity, transporter inhibition, and structural features without assuming that each mechanism contributes independently [6]. Explainability is therefore not an add-on but a design requirement, ensuring that predictions remain connected to safety-relevant biological mechanisms.

Data Sources and Feature Engineering

Curation of Mechanistic Hepatotoxicity Datasets

A curated modeling dataset would conceptually link public DILI labels with in-vitro mechanistic assay results and molecular structure information. DILI-focused machine learning studies have used curated clinical concern labels to support model development, while also showing that label uncertainty and endpoint definition remain central limitations [20]. Mechanistic data could include mitochondrial toxicity screens, transporter inhibition studies, and chemical descriptors generated from standard cheminformatics workflows [21]. Hybrid integration of clinical labels, mechanistic assays, and structure-derived features is consistent with recent efforts to improve DILI prediction through multimodal evidence rather than a single data source [8].

Encoding Mitochondrial and Transporter Assay Results

Mitochondrial and transporter assay results would need harmonization before modeling because measurements may differ by platform, concentration design, cell system, and reporting format. Mitochondrial toxicity endpoints such as respiratory chain inhibition or membrane potential change can be normalized into continuous features while preserving assay-source indicators that signal experimental context [22]. Transporter inhibition values can similarly be transformed into model-ready features, with BSEP inhibition treated as a key signal because of its association with cholestatic liver injury mechanisms [23]. Missing assay values should be encoded explicitly so that the probabilistic model can distinguish absence of evidence from evidence of absence.

Table 1 maps each mechanistic evidence source to its biological interpretation, model representation role, expected contribution to DILI reasoning, and explanation output for safety scientists.

Table 1. Mechanistic Evidence-to-Model Representation Map for Explainable DILI Prediction

Mechanistic evidence domain	Representative input features	Biological interpretation	Model representation role	Expected contribution to DILI reasoning	Explanation output generated for safety scientists
Mitochondrial membrane potential	Degree of membrane depolarization, concentration-response pattern, assay platform indicator	Loss of mitochondrial electrochemical stability and early bioenergetic stress	Continuous mechanistic feature with assay-source context	Supports mitochondrial-liability attribution when elevated	"Predicted concern is partly driven by membrane-potential disruption, suggesting mitochondrial stress rather than purely structural risk."
Oxygen consumption and respiration	Basal respiration change, maximal respiration suppression, respiratory reserve reduction	Impaired oxidative phosphorylation and reduced hepatocyte energy resilience	Nonlinear input to neural representation layer	Helps identify bioenergetic vulnerability, especially when paired with transporter stress	"Respiratory impairment contributes to risk and may amplify concern from bile-acid transport disruption."
ATP depletion	ATP reduction magnitude, recovery pattern, cytotoxicity-adjusted ATP signal	Cellular energy failure and possible progression toward hepatocyte injury	High-salience mechanistic assay feature	Raises concern when energy stress aligns with other DILI-associated signals	"ATP depletion is a dominant local driver of the predicted liver-injury signal."
BSEP inhibition	IC50 or percent inhibition, substrate context, concentration margin	Disruption of bile salt export and cholestatic injury plausibility	Transporter-specific feature with high mechanistic priority	Supports cholestatic or mixed DILI concern	"BSEP inhibition is the primary transporter-related contributor to the predicted risk."
OATP1B1, NTCP, MRP2, and related transporters	Uptake or efflux inhibition values, transporter panel completeness, assay uncertainty	Altered hepatic uptake, efflux, and bile-acid handling	Complementary transporter feature set	Refines whether transporter liability is isolated or system-wide	"Transporter evidence suggests broader hepatobiliary handling disruption rather than a single isolated signal."
Molecular structure and	Structural alerts, lipophilicity, polarity,	Chemical liability, exposure tendency, and	Structural feature block and latent	Complements mechanistic assays	"Structural context supports concern but does not

physicochemical context	molecular weight, learned embeddings	structural similarity to known hepatotoxic compounds	chemical embedding	when biological data are incomplete	independently identify the biological mechanism."
Missing or sparse assay evidence	Missingness flags, assay availability indicators, low-confidence measurements	Absence of evidence, incomplete testing, or uncertain assay reliability	Explicit uncertainty input rather than simple imputation	Prevents overconfident mechanistic interpretation	"Prediction uncertainty is elevated because key mitochondrial or transporter assays are unavailable."
Clinical or regulatory DILI concern label	DILI concern category, label confidence, curated endpoint source	Downstream safety label used for model supervision	Probabilistic target with label uncertainty	Trains risk prediction while acknowledging imperfect clinical labels	"The predicted risk should be interpreted probabilistically because the reference DILI label is uncertain."

Defining the Target Label and Handling Label Uncertainty

The target label would conceptually represent DILI concern categories derived from curated clinical or regulatory knowledge rather than a direct experimental measurement. Studies using DILI labels have emphasized that clinical hepatotoxicity categories can be useful for modeling but may contain uncertainty because injury mechanisms, exposure context, and patient susceptibility are not fully captured [24]. A probabilistic framework can address this limitation by allowing uncertain labels to influence predictions less rigidly than deterministic classification targets [1]. This is especially relevant when mechanistic assays suggest a plausible liability but clinical evidence is incomplete or confounded.

Deep Kernel Model Architecture

Deep Kernel Function

The proposed model would use a compact neural network to transform mitochondrial, transporter, and molecular structure features into a latent representation. A Gaussian process with a smooth covariance function would then operate over that latent space, allowing nonlinear structure while preserving predictive uncertainty [6]. This architecture is conceptually suitable for DILI because transporter inhibition and mitochondrial dysfunction may interact in ways that are not well represented by linear or additive models [11]. The latent representation could therefore capture shared patterns among compounds with similar mechanistic profiles even when their molecular structures differ.

Training and Inference

Training would be framed around probabilistic inference rather than purely deterministic optimization, so the model would learn both a prediction function and uncertainty structure. Bayesian molecular modeling work supports this emphasis because chemical safety datasets often involve sparse observations, heterogeneous measurements, and extrapolation to new chemical series [5]. At inference time, the model would return a risk estimate together with uncertainty, allowing safety scientists to distinguish between a concerning mechanistic signal and an uncertain prediction that requires additional assay evidence [9]. This uncertainty-aware output is particularly important when transporter and mitochondrial features are incomplete or discordant.

Explainability Back-End

The explainability back-end would analyze the trained model by attributing its output to individual mitochondrial endpoints, transporter assays, and structural descriptors. SHAP is a natural candidate because it provides feature-level contribution estimates that can be interpreted across heterogeneous input spaces [7]. Integrated gradients could also be applied to the neural representation layer to examine how molecular embeddings and assay features influence the latent risk representation [17]. In safety pharmacology terms, the explanation layer would translate a model prediction into a mechanistic rationale, such as transporter-driven cholestatic concern, mitochondrial bioenergetic stress, or a combined liability pattern [25].

Explainability: Attributing Hepatotoxicity Risk To Mechanisms

Global Feature Importance: Which Assays Drive DILI Risk?

At the global level, the explanation layer would summarize which mechanistic features most consistently influence predicted DILI risk across the modeled chemical space. BSEP inhibition would be expected to emerge as an important signal because transporter-focused studies have linked bile acid transport disruption with hepatotoxicity mechanisms [4]. Mitochondrial endpoints such as oxygen consumption reduction, membrane potential disruption, and ATP depletion would also be expected to contribute strongly because mitochondrial toxicity models identify bioenergetic impairment as a recurring chemical safety liability [26]. Rather than presenting these explanations as proof of causality, the model would frame them as mechanistic hypotheses that should guide follow-up safety interpretation.

Local Explanation for a Single Compound

For an individual drug candidate, a local explanation would show how each feature contributes to the predicted liver injury risk for that specific compound. A strong BSEP inhibition signal could raise concern when interpreted alongside transporter modeling studies showing that bile salt export pump inhibition can be identified from chemical and assay-derived information [23]. A concurrent mitochondrial toxicity signal could further increase mechanistic concern, particularly when respiratory impairment or membrane depolarization aligns with known bioenergetic mechanisms of hepatotoxicity [10]. Features with little contribution, such as weak uptake transporter inhibition or neutral physicochemical descriptors, would help the scientist understand which mechanisms are not driving the prediction.

Interaction Between Mitochondrial and Transporter Effects

The deep kernel model would be designed to capture interactions between mitochondrial toxicity and transporter inhibition rather than treating each assay result as an isolated linear effect. This is important because transporter inhibition can increase bile acid stress, while mitochondrial impairment can reduce hepatocyte resilience to that stress [11]. Graph-based and neural molecular toxicity models suggest that nonlinear representations can identify complex liability patterns that simpler descriptor models may miss [27]. An interaction explanation would therefore support a multiple-hit interpretation in which moderate signals across mechanisms jointly create greater concern than either signal alone.

Counterfactual Explanations for De-risking

Counterfactual explanations would ask how the predicted risk might change if a mechanistic liability were reduced while other properties remained conceptually similar. For example, the model could evaluate whether lowering predicted BSEP inhibition would be expected to reduce transporter-driven concern, consistent with cholestasis-oriented modeling work [13]. It could also indicate whether mitochondrial toxicity remains a dominant risk driver after the transporter signal is conceptually attenuated, drawing on mitochondrial toxicity models that distinguish structural and mechanistic liabilities [18]. Such explanations would help medicinal chemists prioritize whether to redesign around transporter affinity, mitochondrial liability, or a broader structural alert.

Explainability Methods for Safety Scientists

Visual Dashboard for Mechanistic Attribution

A visual dashboard would present the predicted DILI risk, uncertainty interval, assay values, and mechanistic feature attributions in a single decision-support view. SHAP-based displays are well suited to this purpose because they can show how individual features push a prediction toward higher or lower concern [7]. For safety scientists, the key value would be the ability to inspect whether a model's risk signal is driven by mitochondrial dysfunction, transporter inhibition, chemical structure, or uncertainty from missing evidence [3]. The dashboard should therefore prioritize mechanistic clarity over visual complexity.

Narrative Explanation Generation

A narrative explanation layer would translate feature attributions into concise scientific language suitable for project teams. Rather than reporting numerical certainty or performance metrics, it would state that predicted hepatotoxicity concern is primarily supported by strong transporter inhibition, mitochondrial respiratory impairment, or structural features associated with liver injury risk. This approach is consistent with explainable mitochondrial toxicity modeling, where the goal is to connect a prediction to interpretable chemical or biological drivers [3]. Narrative outputs should remain cautious and probabilistic, emphasizing that the explanation identifies plausible mechanisms rather than definitive clinical outcomes.

Benchmarking Explanations against Known DILI Mechanisms

Explanation quality should be evaluated by comparing model attributions with established mechanisms for well-characterized hepatotoxic compounds. Mechanistic DILI modeling studies emphasize that predictions are most useful when they align with known biological pathways such as mitochondrial impairment, bile acid transport disruption, or reactive metabolite stress. Transporter and BSEP studies provide reference mechanisms for cholestatic concern, while mitochondrial toxicity studies provide reference mechanisms for bioenergetic injury [25]. Agreement between explanations and known mechanisms would support scientific plausibility, whereas disagreement would identify cases requiring deeper review.

Audit Trail and Model Governance

Each model prediction should be accompanied by a documented audit trail containing the input features, uncertainty estimate, explanation output, and versioned model context. This is important because DILI predictions can influence compound prioritization, follow-up assay selection, and safety narratives in development programs [20]. Probabilistic modeling adds governance value by documenting not only the predicted risk direction but also whether the model is uncertain because of missing assays or chemical extrapolation [9]. A governed explanation workflow would help ensure that model outputs remain transparent, reviewable, and scientifically accountable.

Table 2 shows the essential components required for a governed and reviewable audit trail accompanying each DILI model prediction, including inputs, uncertainty characterization, explainability outputs, and versioned model context.

Table 2. Audit Trail Components for DILI Prediction Model Outputs

Audit Trail Component	Description	Example Content	Purpose in DILI Prediction Workflow
Input Features	Molecular, biochemical, and assay-derived descriptors used for prediction	SMILES structure, CYP inhibition profile, mitochondrial toxicity assay results	Defines the evidence base for the prediction
Uncertainty Estimate	Quantification of confidence or prediction reliability	Probability score, confidence interval, epistemic uncertainty flag	Indicates reliability and risk of overinterpretation
Explanation Output	Mechanistic or feature-level attribution of prediction	Key toxicophores, transporter inhibition contribution, SHAP feature importance	Improves interpretability and scientific justification
Versioned Model Context	Metadata describing model version and training configuration	Model v2.3, training dataset release date, architecture type	Ensures reproducibility and traceability of results
Data Completeness Flags	Indicators of missing or extrapolated inputs	"No in vitro CYP3A4 data available", "chemical space extrapolation detected"	Highlights limitations affecting prediction validity
Decision Impact Tag	Downstream regulatory or experimental implication	"Triggers confirmatory hepatocyte assay", "prioritize for review"	Links prediction to development decision-making

Integration Into Safety Pharmacology And Drug Development

Early-Stage De-risking of Lead Compounds

In early discovery, the explainable deep kernel model could be applied once preliminary mitochondrial, transporter, and molecular structure information becomes available. Early DILI screening work suggests that computational models can help prioritize compounds before costly late-stage safety failures occur [20]. The added value of the proposed framework is that it would not only rank compounds by concern but also indicate whether risk is more consistent with mitochondrial toxicity, transporter inhibition, or structural features [8]. This would allow project teams to select follow-up assays and chemical optimization strategies that are aligned with the predicted mechanism.

Supporting Regulatory Submissions with Mechanistic Evidence

For later development, explainable model outputs could support safety narratives by organizing mechanistic evidence into a transparent and auditable format. Transporter consortium perspectives have emphasized the relevance of BSEP inhibition testing in reducing liver injury risk during drug discovery and development [11]. A probabilistic, explainable model could complement such evidence by integrating BSEP, other transporter signals, mitochondrial endpoints, and chemical descriptors into a coherent risk rationale [16]. The model's outputs should be presented as supporting evidence, not as a substitute for experimental toxicology or clinical safety assessment.

Evaluation Strategy

Predictive Performance

Predictive evaluation should compare the explainable deep kernel model against baseline approaches such as random forests, standard Gaussian processes, and neural networks using appropriate validation designs. DILI modeling studies have shown that performance assessment is sensitive to dataset construction, chemical series overlap, and label quality, so scaffold-aware or temporally separated evaluation would be preferred conceptually [15]. Deep learning models based on molecular fingerprints and substructures provide relevant comparator classes because they represent strong structure-based prediction strategies [14, 28]. Evaluation should focus on whether the model is robust, calibrated in a qualitative sense, and useful for decision support, without treating any single metric as sufficient.

Explanation Quality and Mechanistic Plausibility

Explanation quality should be assessed by expert review of whether the model's attributions align with known hepatotoxic mechanisms. SHAP-derived attributions can identify which features influence predictions, but safety scientists must judge whether those features form a biologically plausible narrative [7]. Mitochondrial toxicity models and transporter inhibition models provide mechanistic anchors for this review because they connect chemical features and assay signals to interpretable hepatotoxicity pathways [12, 23]. The evaluation should therefore ask whether explanations are scientifically coherent, stable under reasonable perturbations, and actionable for experimental follow-up.

Table 3 provides an evaluation and governance framework for determining whether the explainable deep kernel model is predictive, calibrated, mechanistically plausible, auditable, and useful for safety-pharmacology decision support.

Table 3. Evaluation and Governance Framework for Explainable Deep Kernel DILI Prediction

Evaluation or governance dimension	Core question	Recommended assessment approach	Failure mode being tested	Decision-use implication
Predictive discrimination	Does the model separate higher-	Compare against random forests, standard Gaussian processes, neural	Apparent performance caused by chemical-series	Supports use as a prioritization tool only if performance

	concern from lower-concern compounds?	networks, and structure-only deep learning baselines using scaffold-aware or temporally separated validation	memorization or label leakage	generalizes beyond closely related compounds
Calibration and uncertainty quality	Does predicted confidence reflect evidence strength?	Examine whether uncertainty increases for sparse assay coverage, conflicting mechanistic signals, or chemically novel compounds	Overconfident predictions for compounds outside the model's evidence base	Determines whether outputs can guide "test more" versus "act now" decisions
Mechanistic attribution plausibility	Do feature attributions align with accepted DILI biology?	Expert review of SHAP or integrated-gradient explanations against known mitochondrial, transporter, and structural mechanisms	Statistically influential features lacking biological plausibility	Determines whether explanations are useful for safety pharmacology interpretation
Local explanation stability	Are compound-level explanations robust to reasonable input perturbations?	Recalculate explanations after small changes in assay values, imputation assumptions, or feature scaling	Fragile explanations that change substantially without meaningful biological change	Prevents overinterpretation of unstable attribution patterns
Interaction detection	Can the model identify multiple-hit liability patterns?	Evaluate cases where moderate mitochondrial and transporter signals jointly increase risk more than either signal alone	Additive-only interpretation that misses nonlinear combined mechanisms	Supports recognition of mixed hepatotoxicity mechanisms involving bioenergetic and cholestatic stress
Counterfactual de-risking relevance	Do counterfactual outputs suggest actionable medicinal chemistry or assay strategies?	Review whether simulated reductions in BSEP inhibition, ATP depletion, or structural alerts produce scientifically coherent risk changes	Counterfactual recommendations that are chemically unrealistic or biologically misleading	Helps medicinal chemists prioritize transporter redesign, mitochondrial liability reduction, or additional testing
Missing-data governance	Does the model distinguish missing evidence from negative evidence?	Track how missing mitochondrial or transporter assays affect uncertainty and explanation language	False reassurance when unmeasured mechanisms are treated as low-risk	Ensures sparse early discovery data are interpreted cautiously
Auditability and documentation	Can each prediction be reviewed after the fact?	Store input features, assay provenance, model version, prediction, uncertainty interval, explanation output, and human reviewer decision	Irreproducible model outputs or undocumented safety decisions	Enables accountable use in project reviews, safety narratives, and regulatory-facing documentation
Human oversight	Is the model used as decision support rather than a replacement for safety judgment?	Require expert review for high-risk, high-uncertainty, or mechanism-discordant predictions	Automated ranking without mechanistic or experimental review	Preserves scientific accountability and prevents black-box decision-making
Prospective utility	Does the framework improve real development decisions?	Retrospective decision simulation followed by prospective evaluation in discovery workflows	Strong retrospective metrics but limited practical usefulness	Establishes whether the model improves compound prioritization, assay selection, and de-risking strategy

Prospective Utility

Prospective utility should be evaluated by asking whether the model would help project teams make better decisions about compound prioritization and follow-up testing. Hybrid DILI prediction work suggests that integrating in-vitro and structural information can provide richer safety evidence than either data type alone [8]. A retrospective decision simulation could examine whether explainable predictions would have highlighted mechanistic concerns earlier for compounds later associated with liver injury, while avoiding claims of definitive clinical forecasting [2]. The main question is whether the model improves the reasoning process around safety risk, not whether it replaces experimental or clinical judgment.

Limitations

Dependence on In-Vitro Data Availability and Quality

The model's mechanistic explanations would depend heavily on the availability, consistency, and biological relevance of in-vitro assay data. Mitochondrial toxicity measurements can vary across platforms and experimental conditions, which may affect how reliably the model attributes risk to bioenergetic mechanisms [21]. Transporter inhibition assays likewise differ in transporter expression systems, substrates, and interpretation thresholds, complicating cross-study harmonization [4]. When assay coverage is sparse or noisy, the model should express greater uncertainty and avoid overconfident mechanistic explanations.

Limited to In-Vitro Mechanisms

The proposed model would primarily represent mechanisms captured by mitochondrial assays, transporter panels, and chemical structure features. It would not fully capture immune-mediated DILI, patient-specific susceptibility, inflammatory context, comorbidities, or exposure-dependent clinical dynamics, all of which may influence real-world liver injury risk. Even probabilistic DILI models remain constrained by the mechanistic and clinical information included in their training data [1]. Therefore, the model should be used as a mechanistic flagging and prioritization tool rather than a definitive predictor of clinical hepatotoxicity.

Conclusion

An explainable deep kernel model offers a conceptual framework for liver injury prediction that links mechanistic assay data with probabilistic machine learning. By combining mitochondrial toxicity endpoints, transporter inhibition profiles, and molecular structure features, the model can organize heterogeneous safety evidence into a unified prediction workflow. The central goal is not only to estimate liver injury concern, but to explain which biological signals support that concern.

The main strength of this approach is its combination of mechanistic attribution and uncertainty-aware prediction. A safety scientist could inspect whether a compound's predicted concern is driven by mitochondrial dysfunction, transporter inhibition, structural alerts, or missing evidence. This makes the model more useful for decision support than a black-box risk score alone.

Important challenges remain for practical implementation. Mechanistic safety datasets are often sparse, inconsistently measured, and difficult to harmonize across assay platforms. Prospective evaluation in drug-discovery settings would be needed to determine whether the model improves prioritization, follow-up testing, and medicinal chemistry decisions.

Future progress will depend on collaborative efforts to expand public mechanistic safety datasets and standardize assay reporting. Interpretable probabilistic models should be integrated into safety pharmacology workflows in ways that preserve scientific judgment and experimental accountability. Used carefully, explainable deep kernel modeling could help connect early in-vitro safety signals to clearer, more actionable liver injury risk narratives.

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. Mostafa F, Howle V, Chen M. Machine learning to predict drug-induced liver injury and its validation on failed drug candidates in development. Toxics. 2024;12(6):385.
3. Jaganathan K, Rehman MU, Tayara H, Chong KT. XML-CIMT: explainable machine learning (XML) model for predicting chemical-induced mitochondrial toxicity. Int J Mol Sci. 2022;23(24):15655.
4. Rodríguez-Pérez R, Gerebtzoff G. Identification of bile salt export pump inhibitors using machine learning: Predictive safety from an industry perspective. Artif Intell Life Sci. 2021;1:100027.
5. Ryu S, Kwon Y, Kim WY. A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. Chem Sci. 2019;10(36):8438-46.
6. Dührkop K. Deep kernel learning improves molecular fingerprint prediction from tandem mass spectra. Bioinformatics. 2022;38(Supplement_1):i342-9.
7. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
8. Seal S, Williams D, Hosseini-Gerami L, Mahale M, Carpenter AE, Spjuth O, et al. Improved detection of drug-induced liver injury by integrating predicted in vivo and in vitro data. Chem Res Toxicol. 2024;37(8):1290-305.
9. Tom G, Hickman RJ, Zinzuwadia A, Mohajeri A, Sanchez-Lengeling B, Aspuru-Guzik A. Calibration and generalizability of probabilistic models on low-data chemical datasets with DIONYSUS. Digit Discov. 2023;2(3):759-74.
10. Mihajlovic M, Vinken M. Mitochondria as the target of hepatotoxicity and drug-induced liver injury: molecular mechanisms and detection methods. Int J Mol Sci. 2022;23(6):3315.
11. Kenna JG, Taskar KS, Battista C, Bourdet DL, Brouwer KL, Brouwer KR, et al. Can bile salt export pump inhibition testing in drug discovery and development reduce liver injury risk? An international transporter consortium perspective. Clin Pharmacol Ther. 2018;104(5):916-32.
12. Zhao P, Peng Y, Xu X, Wang Z, Wu Z, Li W, et al. In silico prediction of mitochondrial toxicity of chemicals using machine learning methods. J Appl Toxicol. 2021;41(10):1518-26.

13. Kotsampasakou E, Ecker GF. Predicting Drug-Induced Cholestasis with the Help of Hepatic Transporters—An in Silico Modeling Approach. J Chem Inf Model. 2017;57(3):608-15.
14. Kang MG, Kang NS. Predictive model for drug-induced liver injury using deep neural networks based on substructure space. Molecules. 2021;26(24):7548.
15. 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.
16. Garcia de Lomana M, Gadaleta D, Raschke M, Fricke R, Montanari F. Predicting liver-related in vitro endpoints with machine learning to support early detection of drug-induced liver injury. Chem Res Toxicol. 2025;38(4):656-71.
17. Chipofya M, Tayara H, Chong KT. Deep probabilistic learning model for prediction of ionic liquids toxicity. Int J Mol Sci. 2022;23(9):5258.
18. Bringezu F, Gómez-Tamayo JC, Pastor M. Ensemble prediction of mitochondrial toxicity using machine learning technology. Comput Toxicol. 2021;20:100189.
19. Jaganathan K, Tayara H, Chong KT. Prediction of drug-induced liver toxicity using SVM and optimal descriptor sets. Int J Mol Sci. 2021;22(15):8073.
20. Martin MT, Koza-Taylor P, Di L, Watt ED, Keefer C, Smaltz D, et al. Early drug-induced liver injury risk screening: "free," as good as it gets. Toxicol Sci. 2022;188(2):208-18.
21. Garcia de Lomana M, Marin Zapata PA, Montanari F. Predicting the mitochondrial toxicity of small molecules: Insights from mechanistic assays and cell painting data. Chem Res Toxicol. 2023;36(7):1107-20.
22. Tang W, Liu W, Wang Z, Hong H, Chen J. Machine learning models on chemical inhibitors of mitochondrial electron transport chain. J Hazard Mater. 2022;426:128067.
23. McLoughlin KS, Jeong CG, Sweitzer TD, Minnich AJ, Tse MJ, Bennion BJ, et al. Machine learning models to predict inhibition of the bile salt export pump. J Chem Inf Model. 2021;61(2):587-602.
24. Kelleci Celik F, Karaduman G. Machine learning-based prediction of drug-induced hepatotoxicity: an OvA-QSTR approach. J Chem Inf Model. 2023;63(15):4602-14.
25. Jain S, Grandits M, Richter L, Ecker GF. Structure based classification for bile salt export pump (BSEP) inhibitors using comparative structural modeling of human BSEP. J Comput Aided Mol Des. 2017;31(6):507-21.
26. Hemmerich J, Troger F, Füzi B, Ecker GF. Using machine learning methods and structural alerts for prediction of mitochondrial toxicity. Mol Inform. 2020;39(5):2000005.
27. AbdulHameed MD, Liu R, Wallqvist A. Using a graph convolutional neural network model to identify bile salt export pump inhibitors. ACS Omega. 2023;8(24):21853-61.
28. Yang Q, Zhang S, Li Y. Deep learning algorithm based on molecular fingerprint for prediction of drug-induced liver injury. Toxicology. 2024;502:153736.