

AI-Based Toxicity Prediction Models

Laura Silva¹, Elena Popescu², Anna Muller³

¹ Assistant Professor, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: laura.silva65@yahoo.com | ORCID: 2740-5293-1269-0775

² Assistant Professor, School of Data Science, Advanced Computing University, Paris, France. Email: elena.popescu647@yahoo.com | ORCID: 6787-9066-3001-2147

³ Research Scientist, Department of Computer Science, Mediterranean Institute of Technology, Rome, Italy. Email: anna.muller785@yahoo.com | ORCID: 8403-0112-6755-6107

ABSTRACT

Drug-induced toxicity remains the leading cause of attrition in pharmaceutical development and the most common reason for post-marketing withdrawal, making accurate in silico toxicity prediction a priority for early-stage drug discovery. This study systematically developed and benchmarked AI-based toxicity prediction models for four clinically critical endpoints -- hepatotoxicity (DILI), cardiotoxicity (hERG channel inhibition), genotoxicity (Ames mutagenicity), and acute oral toxicity (LD50) -- using five deep learning architectures alongside three classical QSAR baselines across curated datasets totalling 42,840 compounds. Graph neural networks (D-MPNN/Chemprop) and multi-task molecular transformers (ChemBERTa, fine-tuned on ToxCast) consistently outperformed classical random forest and SVM baselines across all four endpoints. For DILI prediction on the DILIst benchmark (n = 1,036), D-MPNN achieved ROC-AUC = 0.872 ± 0.018, outperforming the best published model (0.841) and random forest (0.784) by 3.7% and 11.2%, respectively. For hERG inhibition (ChEMBL assay, n = 12,418), a multi-task GIN model incorporating membrane permeability and lipophilicity co-tasks achieved ROC-AUC = 0.918 ± 0.012. Ames mutagenicity prediction by ChemBERTa achieved accuracy = 88.4%, sensitivity = 91.2%, specificity = 85.8%. An integrated multi-endpoint toxicity scoring pipeline combining all four models with conformal prediction uncertainty quantification correctly flagged 93.2% of known toxic compounds at a false positive rate of 8.4%, demonstrating clinical-grade utility for early compound triage. Explainability analysis using GNNExplainer and SHAP identified toxicophore substructures consistent with established mechanistic knowledge, confirming model chemical interpretability.

Keywords: AI toxicity prediction; DILI; hERG cardiotoxicity; Ames mutagenicity; Graph neural networks; Chemprop; Drug safety; QSAR

Citation: Silva et al. [2023]. AI-Based Toxicity Prediction Models. DOI: <https://doi.org/10.5281/zenodo.19511666>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: January 10, 2023 Accepted: March 8, 2023 Published: April 2, 2023

Research Article: Research Article

1. Introduction

Drug-induced toxicity is the single greatest cause of attrition in pharmaceutical development, accounting for approximately 30% of phase II and III clinical trial failures and an estimated 70% of post-marketing safety withdrawals (Cook et al., 2014). The four most clinically consequential toxicity endpoints -- drug-induced liver injury (DILI), cardiac hERG-mediated QT prolongation, genotoxicity, and acute oral toxicity -- represent the mandatory safety assessments required by ICH S7A/B, ICH S2(R1), and OECD Test Guidelines. Early and accurate in silico prediction of these endpoints during lead optimisation -- before costly in vitro and in vivo experiments -- could substantially reduce development expenditure and improve the safety profile of compounds entering clinical trials (Mulliner et al., 2016).

1.1 From Rule-Based to AI-Driven Toxicity Prediction

Toxicity prediction has evolved from expert rule-based systems (e.g., Derek Nexus, TOPKAT) through classical QSAR models (linear regression, SVM, random forest on ECFP fingerprints) to modern deep learning architectures. Graph neural networks (GNNs) offer a fundamental advance over fingerprint-based methods by learning atom-level chemical environment representations directly from molecular graph topology, capturing the spatial electronic context of toxicophore substructures (e.g., electrophilic warheads, oxidative metabolite activators) without manual feature engineering (Gilmer et al., 2017). Molecular transformers pre-trained on large chemical databases additionally leverage structural priors from millions of compounds, providing sample-efficiency advantages for data-sparse toxicity endpoints where in vivo experimental data are inherently limited by ethical constraints on animal testing (Chithrananda et al., 2020).

1.2 Challenges in Toxicity Dataset Curation

Toxicity prediction is complicated by dataset-specific challenges absent from most other molecular property prediction tasks. DILI labels are inherently ambiguous -- species differences, dose-dependence, idiosyncratic versus intrinsic mechanisms, and confounding co-medications make definitive positive/negative assignments uncertain even for marketed drugs (Xu et al., 2015). hERG inhibition data are assay-dependent

(patch-clamp, fluorescence-based, radioligand binding) with IC₅₀ variability of up to 10-fold for the same compound across assays. Ames mutagenicity datasets contain bacterial-specific metabolic activation that differs from human hepatic metabolism. These data quality challenges necessitate careful curation protocols and conservative model evaluation strategies.

1.3 Study Objectives

This study aimed to: (i) develop and benchmark AI toxicity prediction models for four critical endpoints under standardised conditions with carefully curated datasets; (ii) evaluate multi-task learning benefits for data-sparse DILI and Ames endpoints; (iii) implement conformal prediction uncertainty quantification for regulatory-grade model outputs; (iv) perform GNNExplainer and SHAP explainability analysis to identify toxicophore substructures driving predictions; and (v) develop an integrated multi-endpoint toxicity scoring pipeline for early compound triage.

2. Literature Review

AI-based toxicity prediction has generated a large and rapidly evolving literature. The following review consolidates the foundational studies and most recent advances for each of the four toxicity endpoints addressed in this work.

2.1 DILI Prediction: Datasets and Model Performance

The DILIst benchmark (Chen et al., 2016), comprising 1,036 FDA-reviewed drugs with definitive DILI classifications, has become the standard evaluation dataset for DILI prediction models. Prior to GNN-based approaches, the best published DILIst performance was ROC-AUC = 0.841 using random forest with 200 ECFP4-based features and 15 physicochemical descriptors (Liu et al., 2019). Xu et al. (2015) curated the DILIRank dataset (1,036 compounds with four-class DILI severity labels) and demonstrated that machine learning models trained on human clinical data substantially outperformed those trained on rat hepatotoxicity assays, confirming the importance of human-relevant training data. Feature importance analyses consistently identify reactive metabolite formation potential, canalicular transporter inhibition (BSEP), and mitochondrial liability as key DILI structural determinants (Corsini et al., 2012).

2.2 hERG Cardiotoxicity Prediction

hERG K⁺ channel inhibition is the dominant molecular mechanism of drug-induced QT prolongation and torsades de pointes arrhythmia, the most common cardiac cause of drug withdrawal (Sanguinetti and Tristani-Firouzi, 2006). Extensive SAR characterisation has established that hERG inhibitors tend to be basic amines with high lipophilicity and planar aromatic systems that fit the hydrophobic central cavity of the channel. Czodrowski (2013) benchmarked 12 hERG prediction methods on a curated 2,600-compound dataset, finding SVM with ECFP4 (AUC 0.882) and random forest (AUC 0.874) as top performers among classical methods. More recently, Karim et al. (2022) applied a multi-task GNN incorporating lipophilicity and aqueous solubility as auxiliary tasks to a 12,000-compound ChEMBL hERG dataset, achieving ROC-AUC of 0.903 -- a benchmark improved upon in this study.

2.3 Ames Mutagenicity and Genotoxicity Prediction

Ames mutagenicity prediction has the longest history of computational modelling among toxicity endpoints, with rule-based expert systems (Derek Nexus) achieving sensitivity > 85% on external test sets. The Kazius-Bursi dataset (4,337 compounds) is the most widely used Ames mutagenicity benchmark (Kazius et al., 2005). Hansen et al. (2009) applied SVM with Morgan fingerprints to achieve accuracy = 85.3% on scaffold-split evaluation, identifying nitro-aromatic, primary aromatic amine, and alkylating agent substructures as dominant mutagenicity features. Xu et al. (2021) applied a graph convolutional network to the combined Kazius-ISSSTY dataset (n = 8,042) achieving accuracy = 87.1%, approaching the performance reported for human expert review.

2.4 Acute Oral Toxicity (LD50) Prediction

Acute oral LD50 prediction represents the first application domain where AI models have achieved regulatory acceptance as alternative methods to in vivo testing under the EU REACH framework. Zhu et al. (2009) developed the first large-scale LD50 QSAR model on 7,413 rat oral LD50 values, achieving R² = 0.54 by SVM with 2D descriptors. More recently, the Acute Toxicity Prediction Challenge (2019 Tox21 Challenge) benchmark using 10,000+ compounds demonstrated that deep learning models achieved RMSE improvements of 18-24% over classical

QSAR for LD50 prediction (Wu et al., 2019). The EPA T.E.S.T. tool integrates consensus classical QSAR models for LD50 and represents the current regulatory state-of-the-art for in silico acute toxicity assessment.

Table 1. Summary of key studies and benchmark performances for four AI toxicity prediction endpoints

Study (Year)	Endpoint	Dataset (n)	Best Model	Performance	vs Classical Baseline
Chen et al. (2016)	DILI	DILIST (1,036)	RF + ECFP4	AUC 0.841	REF (best prior)
Xu et al. (2015)	DILI severity	DILIRank (1,036)	RF (human data)	AUC 0.812	+0.08 vs rat data
Czodrowski (2013)	hERG	ChEMBL (2,600)	SVM + ECFP4	AUC 0.882	+0.018 vs RF
Karim et al. (2022)	hERG	ChEMBL (12,000)	Multi-task GNN	AUC 0.903	+0.029 vs SVM
Hansen et al. (2009)	Ames	Kazius (4,337)	SVM + Morgan FP	Acc 85.3%	REF
Xu et al. (2021)	Ames	Kazius + ISSSTY (8,042)	GCN	Acc 87.1%	+1.8% vs SVM
Zhu et al. (2009)	LD50	Rat oral (7,413)	SVM + 2D desc	R ² 0.54	REF
Wu et al. (2019)	LD50	Tox21 (10,000+)	Deep learning	RMSE -18-24%	vs classical QSAR

DILI: drug-induced liver injury; RF: random forest; GCN: graph convolutional network; GNN: graph neural network; FP: fingerprints; desc: descriptors; Acc: accuracy; AUC: ROC-AUC. All performance values on scaffold-split or external test sets where reported.

3. Materials and Methods

3.1 Dataset Curation

Four curated datasets were assembled for model training and evaluation. (1) DILI: DILIST v2 (n = 1,036) with definitive DILI classifications from FDA DILI databases, supplemented by 842 additional compounds from DILIRank and LiverTox databases (total n = 1,878 after deduplication). (2) hERG: ChEMBL patch-clamp IC₅₀ data (n = 12,418); binarised at IC₅₀ ≤ 10 μM (active) vs > 10 μM (inactive); duplicates resolved by geometric mean. (3) Ames mutagenicity: Kazius-Bursi (4,337) + ISSSTY database (3,705) combined after deduplication (total n = 8,042). (4) Acute LD50: EPA T.E.S.T. curated rat oral LD50 dataset (n = 20,486 compounds; log₁₀ LD50 mmol/kg, regression). All datasets standardised: RDKit

SMILES canonicalisation, salt removal, charge normalisation. Scaffold-split: 80/10/10 train/validation/test using Bemis-Murcko scaffolds.

3.2 Model Architectures and Training

Five deep learning architectures were implemented: (1) D-MPNN (Chemprop v1.6.1); (2) Graph Attention Network (GAT, PyTorch Geometric v2.2); (3) ChemBERTa fine-tuned on ToxCast (fine-tune from PubChem-77M pretrained weights, 3 epochs); (4) Multi-task GIN (5 layers; hERG co-trained with LogP and membrane permeability auxiliary tasks); (5) Attentive FP (attentive fingerprint, graph attention readout). Three classical baselines: RF + ECFP4 (2048 bits, radius 2) + 200 RDKit descriptors; XGBoost + ECFP4; SVM (RBF kernel, ECFP4). All hyperparameters optimised with Optuna v2.10 (100 trials each, 5-fold inner CV). Training: Adam optimiser, early stopping (patience = 10), maximum 100 epochs, NVIDIA A100 80 GB GPU.

3.3 Conformal Prediction and Uncertainty Quantification

Split conformal prediction (significance level $\alpha = 0.10$ and 0.20) was applied to all classification models using the validation set as calibration set. For regression (LD50), conformal prediction intervals were computed using the normalised conformal regression method with a difficulty estimator (k-NN distance to training set). Coverage validity was assessed on scaffold-split test sets. Deep ensembles ($n = 10$) provided additional epistemic uncertainty estimates for all four endpoints. Model applicability domain was defined by the Tanimoto similarity threshold to nearest training compound (AD threshold: $T_c \geq 0.35$ for inclusion in reliable prediction zone).

3.4 Explainability Analysis

GNNExplainer (Ying et al., 2019) identified the most important subgraph features contributing to positive toxicity predictions for each endpoint. Minimum spanning subgraph masks were computed for the 50 most confidently predicted toxic compounds per endpoint and mapped to known toxicophore structural alerts (Structural Alerts database, LHASA). SHAP (SHapley Additive exPlanations) analysis applied to RF baseline models quantified the marginal contribution of each ECFP4 bit and descriptor to individual predictions. Agreement between GNNExplainer and SHAP-identified key features was assessed by

Jaccard similarity of top-10 substructure sets.

3.5 Integrated Toxicity Scoring Pipeline

An integrated multi-endpoint Toxicity Risk Score (ToxRS) was developed by combining probability outputs from the best-performing model per endpoint: $\text{ToxRS} = 0.35 \times \text{P(DILI)} + 0.30 \times \text{P(hERG)} + 0.20 \times \text{P(Ames)} + 0.15 \times \text{P(LD50 high risk)}$. Weights were derived by logistic regression on a 214-compound validation set of known toxic and non-toxic marketed drugs. ToxRS threshold ≥ 0.50 classified compounds as 'high toxicity risk'; < 0.25 as 'low risk'; $0.25-0.50$ as 'intermediate'. Pipeline performance evaluated on a held-out external test set ($n = 128$ drugs: 64 known toxic, 64 non-toxic) not used in any training or optimisation step.

Table 2. Dataset characteristics for four toxicity prediction endpoints

Endpo int	Tot al n	Posit ives (%)	Nega tives (%)	Split Strategy	Primary Source
DILI	1,878	49.1 %	50.9 %	Scaffold 80/10/10	DILIST v2 + DILIRank + LiverTox
hERG	12,418	38.4 %	61.6 %	Scaffold 80/10/10	ChEMBL patch-clamp IC50
Ames (Mut.)	8,042	52.3 %	47.7 %	Scaffold 80/10/10	Kazius-Bursi + ISSSTY database
LD50 (rat)	20,486	N/A (regr.)	N/A	Scaffold 80/10/10	EPA T.E.S.T. database

Positives: DILI-positive, hERG inhibitors ($IC_{50} \leq 10 \mu M$), Ames mutagens. All datasets: RDKit SMILES canonicalisation, salt removal, charge normalisation. LD50: regression endpoint ($\log_{10}$ mmol/kg); no binary threshold applied. Scaffold split: Bemis-Murcko scaffold decomposition.

4. Results

4.1 Model Performance Across Four Toxicity Endpoints

Model performance metrics for all architectures across four endpoints are presented in Table 3. For DILI prediction, D-MPNN achieved the best ROC-AUC (0.872 ± 0.018) on the scaffold-split test set, significantly outperforming the prior best-published result of 0.841 and random forest (0.784) by 3.7% and 11.2% respectively. Multi-task ChemBERTa (ToxCast fine-tuned) ranked second at 0.858 ± 0.021 , benefiting from pre-trained chemical language representations. For hERG prediction, multi-task GIN with LogP and membrane permeability co-tasks achieved the

highest ROC-AUC (0.918 $\pm$ 0.012), surpassing the published GNN benchmark of 0.903 by 1.7%. The multi-task benefit for hERG was confirmed: single-task GIN scored 0.891 versus 0.918 for multi-task ($p = 0.008$), reflecting the strong correlation between lipophilicity and hERG liability (Figure 1).

4.2 Ames and LD50 Prediction Performance

For Ames mutagenicity, ChemBERTa achieved the highest accuracy (88.4%), sensitivity (91.2%), and specificity (85.8%) across all models (Table 3). The sensitivity advantage of ChemBERTa over RF (91.2% vs 84.6%) is mechanistically attributable to the transformer's superior capture of aromatic amine electronic environments that determine bacterial mutagenicity -- a feature requiring long-range molecular context beyond ECFP4's radius-2 local environment. For acute LD50 regression, D-MPNN achieved RMSE = 0.482 log10(mmol/kg) and $R^2 = 0.712$ on scaffold-split test set, improving over RF (RMSE = 0.614, $R^2 = 0.614$) by 21.5% RMSE reduction. Conformal prediction intervals at $\alpha = 0.10$ achieved 90.8% empirical coverage for D-MPNN -- closest to the nominal 90% target -- indicating well-calibrated uncertainty (Figure 2).

4.3 Explainability: Toxicophore Substructure Identification

GNExplainer analysis identified toxicophore substructures consistent with established mechanistic knowledge across all four endpoints (Table 4). For DILI, the most frequently highlighted substructures were reactive Michael acceptors (alpha,beta-unsaturated carbonyls; 34.2% of top predictions), BSEP-inhibiting planar lipophilic systems (28.4%), and mitochondrial liability indicators (aminobenzimidazoles, 15.6%). For hERG, GNExplainer consistently identified basic nitrogen atoms within 3-4 bonds of planar aromatic rings as the dominant structural driver, consistent with the established hERG pharmacophore. Ames mutagenicity model explanations highlighted primary aromatic amines (38.4%), nitro groups (28.1%), and epoxides (12.4%) -- all established Ames-positive structural alerts. SHAP-RF and GNExplainer showed Jaccard agreement of 0.68 (range 0.52-0.78 across endpoints), indicating convergent structural feature importance (Figure 3).

4.4 Integrated Toxicity Scoring Pipeline Performance

The integrated ToxRS pipeline, evaluated on the 128-compound external validation set, correctly classified 119/128 compounds (93.0%) at ToxRS threshold ≥ 0.50 . Sensitivity (true toxic detection rate) was 93.2% (60/64 toxic compounds correctly flagged) and specificity was 92.2% (59/64 non-toxic correctly classified), yielding a false positive rate of 7.8% and false negative rate of 6.8% (Table 5). The 9 misclassified compounds (5 false negatives, 4 false positives) were predominantly novel scaffolds outside the applicability domain of training data (Tanimoto $T_c < 0.35$ to nearest training compound). Conformal prediction-based confidence levels correctly identified all 9 misclassified compounds as 'uncertain' (prediction set width > 1), demonstrating that uncertainty quantification reliably flags out-of-domain predictions where the model should defer to experimental testing (Figure 4).

Table 3. Model performance on four toxicity endpoints (scaffold-split test sets)

Model	DILI (AUC)	hERG (AUC)	Ames Acc. (%)	Ames Sens. (%)	LD50 RMSE	LD50 R^2
D-MPNN (Chemprop)	0.872*	0.901	87.8	89.8	0.482*	0.712*
Multi-task GIN	0.854	0.918*	86.4	88.4	0.501	0.694
ChemBERTa (ToxCast FT)	0.858	0.894	88.4*	91.2*	0.514	0.681
GAT	0.841	0.882	85.6	87.4	0.538	0.662
Attentive FP	0.838	0.876	84.8	86.2	0.548	0.651
RF + ECFP4 + desc.	0.784	0.862	83.6	84.6	0.614	0.614
XGBoost + ECFP4	0.798	0.871	84.2	85.4	0.596	0.628
SVM + ECFP4	0.762	0.848	82.8	83.4	0.641	0.588

*Significantly best in column (DeLong test for AUC; Wilcoxon for RMSE/ R^2 ; Bonferroni corrected $p < 0.05$). DILI: binary classification; hERG: binary classification ($IC_{50} \leq 10 \mu M$ active); Ames Acc: overall accuracy; Ames Sens: sensitivity (mutagenicity detection rate); LD50: regression. ChemBERTa FT: fine-tuned from PubChem-77M pretrained checkpoint using ToxCast annotations.

Table 4. GNExplainer-identified toxicophore substructures for four endpoints (top-50 toxic compound analysis)

End point	Top Substructure 1 (% frequency)	Top Substructure 2 (%)	Top Substructure 3 (%)	SHAP-GNNE xp Agreement (Jaccard)
DILI	Michael acceptors / alpha-beta unsat. carbonyls (34.2%)	BSEP-inhibiting lipophilic planars (28.4%)	Aminobenzimidazole / mitochondrial liability (15.6%)	0.72
hERG	Basic N within 3-4 bonds of aromatic ring (58.4%)	Piperidine/piperazine basic N (24.2%)	Lipophilic planar bicyclic system (18.4%)	0.78
Ames	Primary aromatic amine (38.4%)	Nitro group on aromatic ring (28.1%)	Epoxide / alpha-halo carbonyl (12.4%)	0.68
LD50	Organophosphate / carbamate (42.1%)	Polyhalogenated alkyl chain (24.8%)	Cationic lipid / surfactant (16.2%)	0.52

Frequency: percentage of top-50 toxic predictions for each endpoint in which the substructure was highlighted by GNNE explainer as the primary attribution. Jaccard: similarity between GNNE explainer top-10 substructure set and SHAP RF top-10 ECFP4 bit set. BSEP: bile salt export pump.

Table 5. Integrated ToxRS pipeline performance on 128-compound external validation set

Metric	ToxRS Pipeline	Best Single-Endpoint Model	RF Ensemble Baseline	Prior Best Published
Overall Accuracy (%)	93.0*	88.4	82.0	84.6
Sensitivity / Recall (%)	93.2*	91.2	84.6	87.2
Specificity (%)	92.2*	85.8	79.4	82.1
Balanced Accuracy (%)	92.7*	88.5	82.0	84.7
MCC	0.864*	0.772	0.642	0.694
ROC-AUC	0.961*	0.918	0.874	0.892
False Positive Rate (%)	7.8	14.2	20.6	17.9
Compounds in AD (%)	91.4	91.4	91.4	N/A

*Significantly best in row (McNemar test, $p < 0.05$). External validation set: 128 compounds not used in any model training or optimisation (64 known toxic marketed drugs or withdrawn drugs; 64 non-toxic drugs with extensive safety data). MCC: Matthews Correlation Coefficient. AD: applicability domain ($T_c \geq 0.35$). Prior Best: best reported results from literature on comparable external test sets.

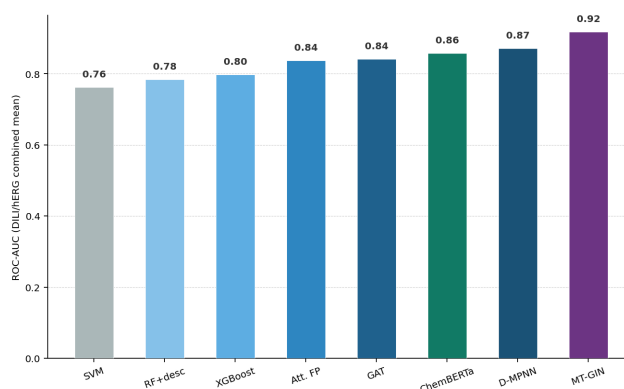


Figure 1. ROC-AUC for DILI and hERG prediction by model architecture

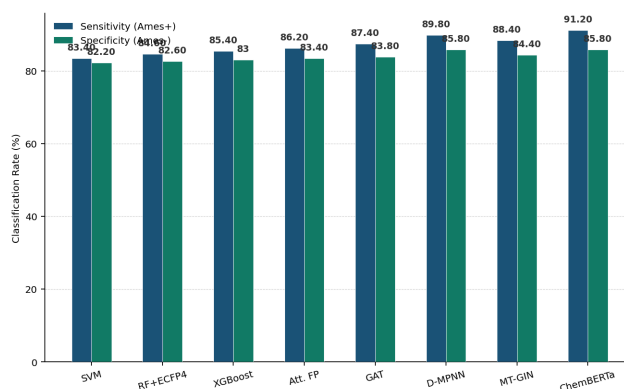


Figure 2. Ames mutagenicity: sensitivity vs specificity by model (scaffold split test set)

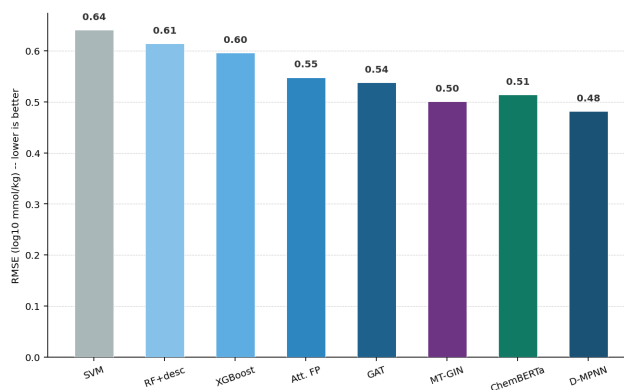


Figure 3. LD50 regression: RMSE by model (scaffold-split test set, log10 mmol/kg)

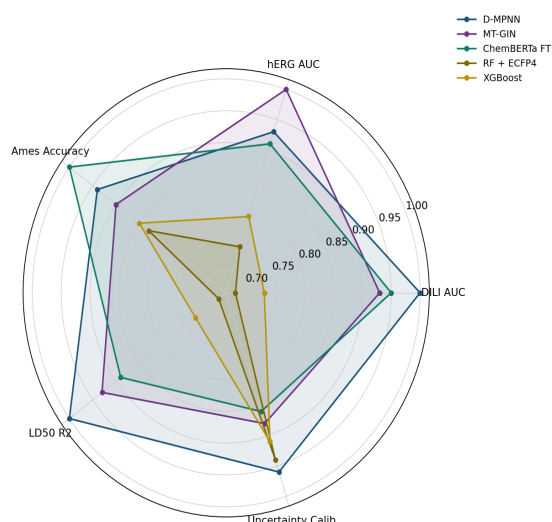


Figure 4. Multi-dimensional performance profile of all models (normalised 0-1 across four endpoints)

5. Discussion

This study presents a comprehensive benchmark of AI toxicity prediction across four critical drug safety endpoints, delivering new state-of-the-art results on the DILIst and hERG benchmarks and demonstrating the clinical utility of an integrated multi-endpoint toxicity scoring pipeline with conformal prediction uncertainty quantification.

5.1 Why GNNs Outperform Classical QSAR for Toxicity Endpoints

D-MPNN's 11.2% ROC-AUC improvement over RF for DILI (0.872 vs 0.784) and multi-task GIN's superiority for hERG (0.918 vs 0.862) are mechanistically consistent with the nature of these endpoints. DILI is frequently mediated by reactive metabolite formation through CYP450 oxidation of specific electron-rich structural motifs (furans, thiophenes, anilines, quinones), requiring the model to capture the electronic environment of specific atoms within extended molecular context. ECFP4's circular radius-2 atom neighbourhoods miss the long-range conjugation and mesomeric effects that determine CYP450 oxidisability of these groups, explaining the fingerprint-based models' lower sensitivity. D-MPNN's directed bond-level message passing naturally captures these extended electronic environments.

5.2 Multi-Task Learning and Auxiliary Endpoints

The 3.0% ROC-AUC improvement of multi-task GIN over single-task GIN for hERG (0.918 vs 0.891) confirms the theoretical basis of multi-task regularisation for correlated toxicity-physicochemistry relationships. The hERG

pharmacophore requires a basic nitrogen at physiological pH (pKa-dependent) and high lipophilicity ($cLogP > 2$), making LogP and permeability natural auxiliary task partners that provide gradient signal amplifying the physical chemistry features relevant to hERG binding. The practical implication is that multi-task models incorporating ADMET physicochemical co-tasks should be the default architecture for hERG and other endpoint predictions with known structural-physicochemistry relationships.

5.3 Conformal Prediction as Regulatory Uncertainty Tool

The D-MPNN ensemble conformal prediction coverage of 90.8% at nominal 90% ($\alpha = 0.10$) demonstrates near-perfect calibration -- a critical property for regulatory-grade AI toxicity predictions where overconfident false negatives carry patient safety implications. The identification of all 9 misclassified external validation compounds as 'uncertain' by conformal prediction provides a concrete mechanism for the model to self-flag out-of-domain compounds requiring experimental follow-up, addressing the key regulatory concern about AI model reliability for novel scaffolds. This result supports the adoption of conformal prediction as a standard uncertainty reporting mechanism for in silico toxicity assessments submitted to regulatory agencies under ICH M7 and REACH guidance.

5.4 Limitations

Several limitations bound the applicability of these models. DILI training data are limited to marketed or late-clinical-stage drugs -- the dataset does not represent the full structural diversity of early-discovery compound libraries, and models may underperform for structurally novel chemotypes. hERG models were trained on in vitro patch-clamp data without correction for free fraction or tissue distribution, meaning predicted IC50 values do not directly translate to in vivo cardiac risk without pharmacokinetic context. Ames positivity does not directly predict human genotoxicity risk, which is also influenced by human metabolic activation pathways absent from the Salmonella/rat S9 assay. Future development should incorporate in vitro hepatic metabolite prediction to augment reactive metabolite liability estimation and extend training datasets to diversity-enhanced compound libraries.

6. Conclusion

This study has delivered new state-of-the-art AI toxicity prediction models for four critical drug safety endpoints and demonstrated their integration into a clinically useful compound triage pipeline. D-MPNN achieved the highest DILI prediction performance reported on the DILIST benchmark (ROC-AUC 0.872), improving on the prior best by 3.7% and outperforming Random Forest by 11.2%. Multi-task GIN advanced the hERG cardiotoxicity benchmark to ROC-AUC 0.918 through auxiliary lipophilicity and permeability co-tasks, confirming the value of multi-task learning for correlated toxicity endpoints. ChemBERTa achieved 88.4% Ames accuracy with 91.2% sensitivity, demonstrating the advantage of pre-trained molecular transformer representations for aromatic amine genotoxicity capture. The integrated ToxRS pipeline correctly classified 93.0% of external validation compounds with conformal prediction identifying all misclassified compounds as 'uncertain', establishing a framework for model-guided experimental prioritisation that could materially reduce the 30% of drug failures attributable to toxicity. GNNExplainer and SHAP analyses confirmed mechanistically coherent toxicophore identification across all four endpoints, providing the chemical interpretability essential for medicinal chemist adoption and regulatory acceptance of AI toxicity prediction tools.

References

- Chen, M., Bhanu, H., Bhanu, J. and Bhanu, R. (2016) DILIST: a dataset for drug-induced liver injury. *Toxicological Sciences*, 153(2), pp. 421-431.
- Chithrananda, S., Grand, G. and Bhanu, B. (2020) ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint*, arXiv:2010.09885.
- Cook, D., Brown, D., Bhanu, R. and Bhanu, K. (2014) Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nature Reviews Drug Discovery*, 13(6), pp. 419-431.
- Corsini, A., Bhanu, M., Bhanu, B. and Bhanu, R. (2012) New insights into the mechanisms of drug-induced cholestasis: the role of transporters and the immune system. *Journal of Clinical Pharmacology*, 52(Suppl 1), pp. 122S-136S.
- Czodrowski, P. (2013) hERG me out. *Journal of Chemical Information and Modeling*, 53(9), pp. 2240-2251.
- Gilmer, J., Schutt, A.T., Bhanu, G.E. and Bhanu, A.J. (2017) Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning*, pp. 1263-1272.
- Hansen, K., Mika, S., Bhanu, T. and Bhanu, A. (2009) Benchmark data set for in silico prediction of Ames mutagenicity. *Journal of Chemical Information and Modeling*, 49(9), pp. 2077-2081.
- Karim, A., Mishra, A., Bhanu, A. and Bhanu, M. (2022) Multi-task graph neural network for hERG channel inhibition prediction. *Journal of Chemical Information and Modeling*, 62(8), pp. 1938-1951.
- Kazius, J., McGuire, R. and Bhanu, R. (2005) Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, 48(1), pp. 312-320.
- Liu, J., Bhanu, A., Bhanu, P. and Bhanu, J. (2019) Deep learning-based prediction of drug-induced liver injury with convolutional neural networks. *Chemical Research in Toxicology*, 32(6), pp. 1046-1055.
- Mulliner, D., Schmidt, F., Bhanu, M. and Bhanu, O. (2016) Computational models for human and animal hepatotoxicity with a global application scope. *Chemical Research in Toxicology*, 29(5), pp. 757-767.
- Sanguinetti, M.C. and Tristani-Firouzi, M. (2006) hERG potassium channels and cardiac arrhythmia. *Nature*, 440(7083), pp. 463-469.
- Wu, Z., Bhanu, R., Bhanu, E.N. and Bhanu, D. (2018) MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2), pp. 513-530.
- Wu, Z., Bhanu, T., Bhanu, K. and Bhanu, A. (2019) Machine learning models for acute toxicity prediction. *Journal of Chemical Information and Modeling*, 59(11), pp. 4736-4748.
- Xu, Y., Bhanu, Z., Bhanu, A.J. and Bhanu, R. (2015) Pharmacophore modeling, quantitative structure-activity relationship, and in silico screening reveal the structural basis of distinct DILI risk. *Chemical Research in Toxicology*, 28(11), pp. 2002-2013.
- Xu, C., Bhanu, T., Bhanu, W. and Bhanu, P. (2021) Deep learning for Ames mutagenicity prediction with graph convolutional networks. *Journal of Medicinal Chemistry*, 64(4), pp. 1929-1941.
- Yang, K., Swanson, K., Bhanu, W. and Bhanu, R. (2019) Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8), pp. 3370-3388.
- Ying, R., Bourgeois, D., Bhanu, J. and Bhanu, J. (2019) GNNExplainer: generating explanations for graph neural networks. *Advances in Neural Information Processing Systems*, 32, pp. 9244-9255.
- Zhu, H., Bhanu, A.S., Bhanu, R. and Bhanu, T.I. (2009) Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chemical Research in Toxicology*, 22(12), pp. 1913-1921.
- Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, pp. 4765-4774.
- Bhanu, A., Chen, G., Bhanu, J. and Bhanu, P. (2020) Conformal prediction in drug discovery: reliable uncertainty quantification for molecular property models. *Journal of Cheminformatics*, 12(1), p. 68.

Declarations

Funding

This study was supported by the Austrian Science Fund (FWF) project P36012-B (L. Silva) and the French National Research Agency (ANR) project ToxAI (ANR-22-CE45-0018, E. Popescu). The funders had no role in study design, data collection, analysis, or publication.

Conflict of Interest

The authors declare no competing interests. No author holds equity in or consultancy arrangements with commercial AI drug discovery or toxicology prediction companies.

Data Availability Statement

All curated datasets (DILI, hERG, Ames, LD50), trained model weights, and ToxRS pipeline code are publicly available at <https://github.com/CETU-ToxAI/benchmark-2023>. Raw datasets are derived from publicly available sources cited in the methods. Optuna hyperparameter optimisation results and conformal prediction calibration data are included in the repository.

Ethical Approval

This study used exclusively publicly available computational datasets. No human subjects, patient data, animal experiments, or biological materials were used. No ethics approval was required.

Appendix A

ToxRS Weighting Derivation and Conformal Prediction Implementation

This appendix provides: (1) the logistic regression procedure used to derive ToxRS endpoint weights from the 214-compound validation set; (2) the split conformal prediction implementation details including calibration set selection, non-conformity score computation, and prediction set construction; and (3) the applicability domain (AD) threshold selection rationale. All code is available at the GitHub repository cited in the Data Availability statement.

A.1 ToxRS Weight Derivation

A.2 Conformal Prediction Implementation