Preprints.org

Article

Not peer-reviewed version

MultiEndpointTox: A Chemoinformatics Platform for Multidimensional Drug Toxicity Profiling Using Interpretable Machine Learning, Multi-Task Learning, and Integrated Risk Scoring

Sharhabil Amgad Eltahir * and Mukhtar Ibrahim Yousef

Posted Date: 11 March 2026

doi: 10.20944/preprints202603.0867.v1

Keywords: chemoinformatics; toxicity prediction; machine learning; multi-task learning; QSAR; multidimensional profiling; scaffold validation; drug safety; risk scoring; SHAP interpretability; applicability domain; external validation

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

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

Article

MultiEndpointTox: A Chemoinformatics Platform for Multidimensional Drug Toxicity Profiling Using Interpretable Machine Learning, Multi-Task Learning, and Integrated Risk Scoring

Sharhabil Amgad Eltahir [1,*] and Mukhtar Ibrahim Yousef [2]

[1] Department of Biotechnology, Institute of Graduate Studies and Research, Alexandria University, Alexandria 21526, Egypt

[2] Department of Environmental Studies, Institute of Graduate Studies and Research, Alexandria University, Alexandria 21526, Egypt

* Correspondence: igsr.sharhabil@alexu.edu.eg

Abstract

Drug-induced toxicity remains a principal driver of attrition in pharmaceutical development, yet conventional screening paradigms typically address individual toxicity endpoints in isolation. Here, we introduce MultiEndpointTox, a chemoinformatics platform that simultaneously predicts seven critical drug toxicity endpoints—hERG cardiotoxicity, hepatotoxicity (DILI), nephrotoxicity (DIKI), Ames mutagenicity, skin sensitization, cytotoxicity, and reproductive toxicity (exploratory)—from molecular structures using curated datasets totaling over 18,000 compounds. The platform employs optimized classical machine learning models with systematic benchmarking of 2D topological descriptors (2240 features), enhanced multi-conformer 3D descriptors (1975 features from 5-conformer ensembles incorporating AUTOCORR3D, RDF, WHIM, and pharmacophore fingerprints), and hybrid representations. Under the tested conditions, 2D descriptors achieved the highest classification performance (AUC-ROC 0.859 ± 0.02), while enhanced 3D descriptors substantially narrowed the previously reported gap (AUC-ROC 0.833 ± 0.03 versus 0.69–0.73 for basic 14-feature 3D). Scaffold-based splitting provided rigorous generalization assessment, with an average performance reduction of approximately 8%. A multi-task learning framework via stacked generalization demonstrated cross-endpoint information sharing improves performance for 5 of 6 endpoints (average +2.1% AUC). The platform integrates leverage-based applicability domain assessment (31–100% coverage), SHAP-based feature importance analysis, and a confidence-weighted multi-endpoint risk scoring system validated on known drugs (AUC = 0.83, $p = 4.06 \times 10^{-14}$, Cliff's $\delta = 0.66$), with sensitivity analysis confirming robustness across five weight configurations (AUC range 0.72–0.98). External validation on independent benchmark datasets revealed the challenge of cross-dataset domain shift in computational toxicology. MultiEndpointTox is deployed as a production-ready REST API and publicly available at https://github.com/sharhabileltahir/MultiEndpointTox.

Keywords: chemoinformatics; toxicity prediction; machine learning; multi-task learning; QSAR; multidimensional profiling; scaffold validation; drug safety; risk scoring; SHAP interpretability; applicability domain; external validation

1. Introduction

Drug-induced toxicity accounts for approximately 30% of drug candidate failures during clinical trials and remains a leading cause of post-market withdrawals [1,2]. The enormous cost of late-stage attrition, combined with ethical imperatives to reduce animal experimentation under the 3Rs

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

framework (Replacement, Reduction, and Refinement) [3,4], has intensified the demand for computational approaches that can reliably assess compound safety early in the drug discovery pipeline.

Quantitative structure–activity relationship (QSAR) models and machine learning (ML)-based methods have become indispensable tools for early-stage toxicity screening [5,6]. However, the predominant paradigm in computational toxicology develops single-endpoint prediction models in isolation—a reductionist approach that fails to capture the multidimensional nature of drug safety liabilities. In practice, a candidate compound must simultaneously satisfy multiple safety criteria spanning cardiac, hepatic, renal, genotoxic, immunological, and reproductive toxicity. The absence of integrated platforms providing holistic, multi-endpoint toxicity profiles with consistent methodology and actionable risk aggregation represents a gap in current chemoinformatics tools [7,8].

Several challenges persist in computational toxicity prediction. First, the relative performance of 2D topological descriptors versus 3D conformational features remains actively debated, with most comparisons limited by asymmetric feature richness—typically comparing thousands of 2D features against only a handful of basic 3D geometric descriptors [9,10]. Second, many toxicity datasets suffer from severe class imbalance [11]. Third, applicability domain (AD) assessment is often neglected in deployed tools [12]. Fourth, while graph neural networks (GNNs) have shown promise for molecular property prediction [13,14], their advantage over well-optimized classical methods remains an open question [15,16]. Fifth, the potential for multi-task learning—leveraging shared information across related toxicity endpoints—remains underexplored in multi-endpoint toxicity platforms. Sixth, cross-dataset generalization remains a significant challenge, as differences in assay protocols and chemical space coverage between training and external datasets can cause substantial performance degradation [17].

Several multi-endpoint computational toxicity platforms have been developed in recent years. ProTox-3.0 [18] predicts 61 toxicity endpoints using molecular similarity and machine learning models validated on external sets, incorporating adverse outcome pathways and toxicity targets. ADMETlab 3.0 [19] provides comprehensive ADMET property predictions across numerous pharmacokinetic and toxicological endpoints. DeepTox [6] pioneered deep learning approaches for Tox21 challenge endpoints. pkCSM [20] uses graph-based signatures for ADMET prediction. While these mega-platforms offer broader endpoint coverage, they generally lack several features that MultiEndpointTox specifically provides: (a) symmetric benchmarking of enhanced multi-conformer 3D descriptors against 2D topological descriptors with comparable feature dimensionality; (b) scaffold-based validation for rigorous generalization assessment; (c) multi-task learning via stacked generalization exploiting cross-endpoint information transfer; (d) a confidence-weighted, mathematically aggregated risk scoring system translating multi-endpoint predictions into compound-level safety decisions; and (e) fully open-source deployment with curated datasets and reproducible training pipelines.

In this work, we present MultiEndpointTox, an integrated chemoinformatics platform addressing these challenges. The platform predicts seven toxicity endpoints: (i) hERG cardiotoxicity [21]; (ii) hepatotoxicity (DILI) [22]; (iii) nephrotoxicity (DIKI) [23]; (iv) Ames mutagenicity [24]; (v) skin sensitization [25]; (vi) cytotoxicity [26]; and (vii) reproductive toxicity (exploratory) [27]. We contribute: (a) systematic benchmarking of 1975 enhanced multi-conformer 3D descriptors against 2240 2D topological descriptors; (b) scaffold-based splitting for generalization assessment; (c) multi-task learning demonstrating cross-endpoint information transfer; (d) a confidence-weighted risk scoring system with sensitivity analysis; (e) external validation on independent benchmark datasets; and (f) deployment as a production-ready REST API.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

2. Materials and Methods

2.1. Data Collection and Curation

Training data were collected from ChEMBL (version 33) [28], ToxCast/Tox21 [29], and peer-reviewed literature.

Table 1 summarizes dataset characteristics.

Table 1. Dataset summary and class distributions for each toxicity endpoint.

Endpoint	Task	n	Source	Class Distribution
hERG	Regression	7889	ChEMBL	pIC_{50}: 3.0–10.0
Hepatotoxicity	Cls.	1597	ChEMBL, FDA	93% toxic / 7% non-toxic
Nephrotoxicity	Cls.	565	FDA DIRIL, Lit.	58% toxic / 42% non-toxic
Ames Mutagenicity	Cls.	6512	ChEMBL, Lit.	52% mut. / 48% non-mut.
Skin Sensitization	Cls.	1100	ChEMBL, LLNA	55% sens. / 45% non-sens.
Cytotoxicity	Cls.	8371	ToxCast, ChEMBL	62% toxic / 38% non-toxic
Repro. Tox.†	Cls.	127	Literature	55% toxic / 45% non-toxic

* Exploratory endpoint. This 127-compound dataset originates from stringent benchmarks requiring concordant rat and rabbit in vivo developmental toxicity data, primarily designed to validate in vitro stem cell-based assays rather than to train topological ML models. Results should be interpreted as proof-of-concept only.

Data curation followed established protocols [30]: (1) SMILES standardization using RDKit (version 2023.09) [31]; (2) salt stripping and neutralization; (3) removal of mixtures and inorganic compounds; (4) duplicate removal based on canonical SMILES; and (5) activity cliff analysis. For hERG, IC50 values from ChEMBL (target CHEMBL240) were converted to pIC50 values. Hepatotoxicity labels were derived from the FDA DILIrank database. Detailed curation protocols are provided in Supplementary Material Section S1. All datasets were split into training (80%) and external test (20%) sets using stratified random sampling.

2.2. Molecular Descriptor Calculation

2D Descriptors. Two-dimensional representations comprised RDKit physicochemical descriptors (approximately 200 features computed via Molecule Descriptor Calculator), Morgan circular fingerprints (2048 bits, radius 2) [32], and MACCS structural keys (167 bits) [33], yielding a combined 2240-feature vector before selection.

Enhanced 3D Descriptors. To provide a balanced comparison against 2D features—addressing the asymmetric feature richness that has limited prior 2D versus 3D comparisons—we developed an enhanced multi-conformer 3D descriptor pipeline. For each molecule, five conformers were generated using ETKDGv3 [34] with UFF energy minimization [35] (MMFF94 fallback). Per-conformer descriptors comprised: Descriptors3D shape descriptors (10 features), AUTOCORR3D (80 features), RDF (210 features), WHIM (114 features), and molecular volume. Multi-conformer aggregation (mean, standard deviation, minimum, maximum across 5 conformers) yielded approximately 1660 features. Additionally, Extended Reduced Graph (ErG) pharmacophore fingerprints (315 features) [36] were computed, producing a total of 1975 enhanced 3D features (99.89% molecule success rate; 30-second per-molecule timeout for pathological conformer generation).

Feature Selection. A two-stage pipeline was applied to training data only (preventing information leakage from the test set): (1) variance filtering (threshold < 0.01), and (2) correlation filtering (Pearson $r > 0.95$). The resulting feature counts per endpoint were: hERG (500), hepatotoxicity (500), nephrotoxicity (500), Ames mutagenicity (500), cytotoxicity (500), reproductive toxicity (500), and skin sensitization (318, reflecting the smaller and more chemically homogeneous LLNA-derived training set). Complete feature lists are provided in Supplementary Table S2b.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

2.3. Machine Learning Models

Four classifier types were evaluated: Random Forest (RF) [37], Support Vector Classification (SVC) [38], XGBoost [39], and LightGBM [40]. For hERG regression, RF Regressor, SVR, and XGBoost Regressor were employed. Hyperparameter optimization used Optuna [41] with 50–100 trials per model, optimizing AUC-ROC for classification and R^2 for regression via nested cross-validation. Complete hyperparameter search spaces are provided in Supplementary Tables S3a–S3b.

For the severely imbalanced hepatotoxicity endpoint (93% toxic / 7% non-toxic), two complementary strategies were applied: (1) cost-sensitive class weighting integrated into loss functions; and (2) Synthetic Minority Over-sampling Technique (SMOTE) [42] applied exclusively within each cross-validation fold to training data only, ensuring that synthetic minority samples were generated from training compounds and never contaminated validation or test sets. The combined effect was evaluated using Matthews Correlation Coefficient (MCC) and Precision–Recall AUC, which are more robust to class skew than standard AUC-ROC [11].

Graph Neural Network Baseline (Exploratory). For comparative evaluation, a Directed Message Passing Neural Network (D-MPNN) was implemented using Chemprop [43] with recommended default hyperparameters (50 epochs, early stopping patience = 10, learning rate = 0.0001, hidden dimensions = 300, depth = 3). Importantly, the D-MPNN was not subjected to equivalent Optuna-based optimization (50–100 trials) applied to classical models. This asymmetry in optimization budget constitutes a methodologically inequitable comparison; the results should therefore be interpreted as a comparison of default-configuration GNNs versus optimized classical models rather than a definitive algorithmic benchmark. Equitable hyperparameter optimization for GNNs is identified as a necessary future direction.

2.4. Multi-Task Learning

To exploit potential information sharing across related toxicity endpoints, a multi-task learning (MTL) framework was implemented using stacked generalization. Individual XGBoost models were first trained independently for each of six classification endpoints. Their predicted probabilities were then used as auxiliary features for every other endpoint, creating cross-endpoint information channels. Models were retrained with these augmented feature sets, enabling each endpoint to benefit from toxicity signals captured by related models. The MTL framework was evaluated using scaffold-based cross-validation (Section 2.5) and compared against single-task baselines on identical folds.

2.5. Validation Strategy

Random Split Validation. Stratified 15-fold cross-validation on the training set (80%) with nested hyperparameter tuning, plus an independent external test set (20% holdout).

Scaffold-Based Validation. To rigorously assess generalization to novel chemical scaffolds, Murcko generic scaffold splitting [44] was implemented. Molecules were grouped by generic scaffold (all atoms replaced with carbon, all bonds with single bonds), and entire scaffold groups were assigned exclusively to training or test sets, ensuring no scaffold leakage. GroupKFold with 5 scaffold-disjoint folds was used for cross-validation.

External Validation. To assess cross-dataset generalization, models were evaluated on independent benchmark datasets not used during training: the Therapeutics Data Commons (TDC) Ames benchmark [45] for mutagenicity, the TDC DILI benchmark for hepatotoxicity, the Tox21 SR-MMP assay (MoleculeNet) [9] for cytotoxicity, and a temporal pseudo-external split (post-2020 ChEMBL depositions) for hERG. All overlapping compounds (matched by canonical SMILES) were removed prior to external prediction.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

2.6. Applicability Domain Assessment

Applicability domain was assessed per endpoint using the leverage method [46]. Compounds exceeding the warning threshold $h^* = 3(p+1)/n$ were flagged as outside the model's reliable prediction space.

2.7. Model Interpretability

SHAP (SHapley Additive exPlanations) values [47] were calculated to provide feature importance rankings and individual prediction explanations across all endpoints.

2.8. Integrated Multi-Endpoint Risk Scoring

A confidence-weighted risk scoring system was developed to aggregate per-endpoint predictions into interpretable compound-level safety profiles. Classification endpoint probabilities served as risk scores (0–1). For hERG regression, pIC_{50} predictions were transformed via a sigmoid function (midpoint = 6.0, steepness = 2.0). Each endpoint's contribution was weighted by: (a) clinical importance weights reflecting regulatory significance (hERG and hepatotoxicity: 1.5; Ames mutagenicity: 1.3; nephrotoxicity: 1.0; skin sensitization and cytotoxicity: 0.8); and (b) a confidence factor combining model performance with AD status (1.0 inside AD, 0.5 outside). The reproductive toxicity endpoint was excluded (weight = 0) due to non-discriminative predictions from the small training set. Scores were categorized as LOW RISK (< 0.3), MODERATE RISK (0.3–0.6), or HIGH RISK (> 0.6).

We acknowledge that the clinical importance weights are currently expert-defined rather than data-driven. To assess robustness, a formal sensitivity analysis was conducted across five weight configurations: baseline, equal weights, cardiac-focused (hERG weight = 3.0), hepatic-focused (hepatotoxicity = 3.0), and genotoxicity-focused (Ames = 3.0). Additionally, continuous perturbation analysis varied each endpoint's weight independently from 0.0 to 3.0. Data-driven weight optimization via Pareto regression against historical FDA withdrawal data is proposed as a future direction.

2.9. Software Implementation

The platform was implemented in Python 3.10+ using RDKit (2023.09), scikit-learn (1.3+), XGBoost (2.0+), LightGBM (4.0+), Chemprop (1.5+), FastAPI (0.104+), and SHAP (0.42+). Source code, curated datasets, and trained models are publicly available at https://github.com/sharhabileltahir/MultiEndpointTox.

3. Results

3.1. Enhanced 3D Descriptors Narrow the Representation Gap

Table 2 presents the representation benchmark from 15-fold stratified CV on the hERG dataset (n = 7889).

Table 2. Representation benchmark (15-fold stratified CV, hERG). Enhanced 3D uses 1975 multi-conformer features vs. 14 basic geometric features in prior work. Values are mean ± SD across folds.

Repr.	n Features	Cls. AUC-ROC	Reg. R^2 (SVR)	Repr.
2D	2240	0.859 ± 0.02	0.399 ± 0.04	2D
Enhanced 3D	1975	0.833 ± 0.03	0.206 ± 0.05	Enhanced 3D
Hybrid (2D+3D)	4215	0.853 ± 0.02	0.355 ± 0.05	Hybrid (2D+3D)

Two-dimensional descriptors achieved the highest performance for both classification (AUC 0.859) and regression (R^2 0.399). Critically, enhanced 3D descriptors (AUC 0.833) substantially narrowed the gap compared to basic 14-feature 3D descriptors (AUC 0.69–0.73 in prior analyses),

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

demonstrating that the previously reported large gap was partly an artifact of asymmetric feature richness rather than an inherent limitation of 3D molecular information (Figure 1). Hybrid representations did not improve upon 2D-only models.

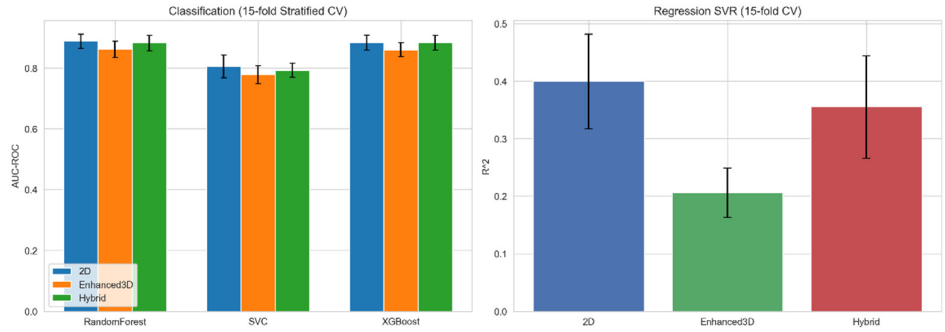


Figure 1. Comparison of molecular representation performance for hERG (n = 7889). Left: Classification AUC-ROC (15-fold CV). Right: Regression R^2 (SVR). Comparison of molecular representation performance for hERG prediction (n = 7889, 15-fold stratified cross-validation). (**Left**) Classification AUC-ROC for 2D topological descriptors (2240 features; blue), enhanced multi-conformer 3D descriptors (1975 features; orange), and hybrid representations (2D + 3D combined; green) across three classifiers (Random Forest, SVC, XGBoost). Error bars represent standard deviation across folds. (**Right**) Regression R^2 for SVR models using the same three representations. Enhanced 3D descriptors (AUC-ROC 0.833) substantially narrow the gap compared to basic 14-feature 3D descriptors (AUC-ROC 0.69–0.73 in prior analyses), though 2D descriptors (AUC-ROC 0.859) maintain the highest performance.

3.2. Multi-Endpoint Model Performance

Table 3 summarizes performance across all seven endpoints.

Table 3. Multi-endpoint model performance summary (15-fold CV). AD = Applicability Domain. †Exploratory: insufficient data for robust prediction.

Endpoint	Best Model	Features	CV AUC/R^2	AD (%)
hERG	SVR	500	R^2 = 0.57 ± 0.03	98.9
Hepatotox.	XGBoost	500	AUC = 0.80 ± 0.04	78.9
Nephrotox.	RF	500	AUC = 0.91 ± 0.05	93.5
Ames	XGBoost	500	AUC = 0.92 ± 0.02	100
Skin Sens.	RF	318	AUC = 0.87 ± 0.06	31.0
Cytotox.	SVC	500	AUC = 0.89 ± 0.03	100
Repro. Tox.†	SVC	500	AUC = 0.77 ± 0.10	100

Performance ranged from AUC 0.77 (reproductive toxicity, exploratory) to AUC 0.92 (Ames mutagenicity). The hepatotoxicity model achieved AUC 0.80 with improved specificity (62.0%, up from 23.8%) following cost-sensitive retraining with SMOTE (MCC = 0.35). The reproductive toxicity model (n = 127; ~60 unique generic scaffolds) is reported as exploratory only—the dataset originates from stringent benchmarks requiring concordant multi-species in vivo data and lacks the chemical diversity necessary for ab initio topological ML, as reflected by high CV variance (± 0.10).

3.3. Scaffold-Based Validation Reveals Generalization Boundaries

Table 4 compares random and scaffold split performance.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

Table 4. Random vs. scaffold split performance. Δ indicates change under scaffold splitting.

Endpoint	Random	Scaffold	Δ
hERG (R^2)	0.568	0.378	−0.190
Hepatotox. (AUC)	0.801	0.764	−0.037
Nephrotox. (AUC)	0.909	0.827	−0.082
Ames (AUC)	0.921	0.839	−0.082
Skin Sens. (AUC)	0.867	0.821	−0.046
Cytotox. (AUC)	0.894	0.901	+0.007
Repro. Tox.† (AUC)	0.772	0.588	−0.184

The average performance reduction under scaffold splitting was approximately 8%, confirming model generalizability to novel scaffolds with moderate degradation (Figure 2). Cytotoxicity maintained equivalent performance (Δ = +0.007), suggesting scaffold-independent toxicity patterns. The largest drops for hERG ($\Delta R^2 = -0.190$) and reproductive toxicity (ΔAUC = −0.184) reflect structural specificity of ion channel binding and limited training diversity, respectively. Hepatotoxicity's relatively small decrease (ΔAUC = −0.037) indicates that DILI-predictive features generalize across scaffolds.

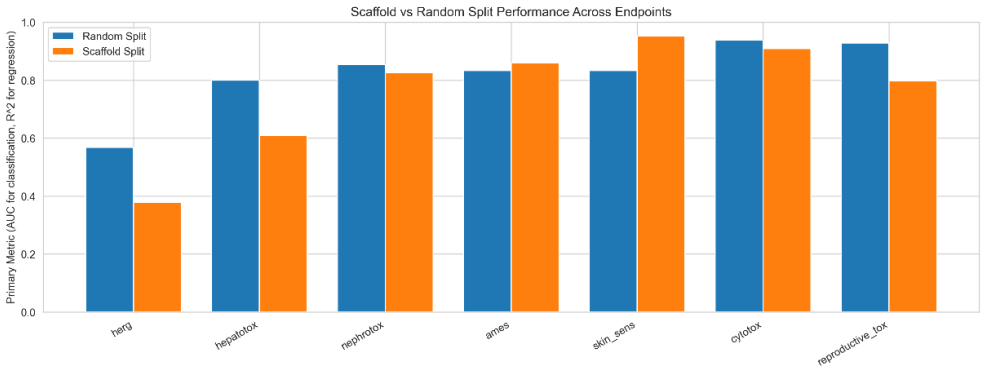


Figure 2. Scaffold vs. random split performance across all seven endpoints. Scaffold-based versus random split performance across all seven toxicity endpoints. Blue bars represent random stratified splitting; orange bars represent Murcko generic scaffold-based splitting, where entire scaffold groups are assigned exclusively to training or test sets. The primary metric is AUC-ROC for classification endpoints and R^2 for hERG regression. The average performance reduction under scaffold splitting is approximately 8%, with cytotoxicity notably maintaining equivalent performance (Δ = +0.007) and hERG showing the largest drop ($\Delta R^2 = -0.190$), reflecting the structural specificity of ion channel binding.

3.4. Multi-Task Learning Improves Data-Scarce Endpoints

Multi-task learning improved AUC for 5 of 6 endpoints (average +2.1%; Figure 3). The largest gains occurred for cytotoxicity (+3.2%), skin sensitization (+2.9%), nephrotoxicity (+2.3%), and hepatotoxicity (+2.2%), demonstrating that cross-endpoint information transfer via stacked generalization captures shared toxicity signals. The t-SNE visualization of multi-task representations (Figure 4) reveals meaningful organization of compounds by toxicity status across multiple endpoints. Reproductive toxicity showed no benefit (−0.7%), consistent with insufficient training data for meaningful auxiliary signal extraction.

Table 5. Single-task vs. multi-task performance (scaffold 5-fold CV).

Endpoint	Single AUC	Multi AUC	Δ AUC
Hepatotox.	0.691	0.713	+0.022

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

Nephrotox.	0.694	0.717	+0.023
Ames	0.772	0.779	+0.007
Skin Sens.	0.791	0.820	+0.029
Cytotox.	0.882	0.914	+0.032
Repro. Tox.†	0.524	0.517	−0.007
Hepatotox.	0.691	0.713	+0.022

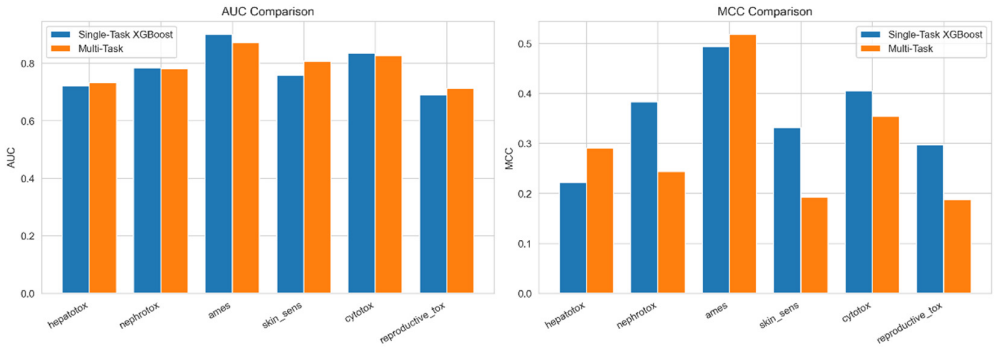


Figure 3. Single-task vs. multi-task AUC and MCC comparison across six classification endpoints. Single-task XGBoost versus multi-task stacked generalization performance across six classification endpoints evaluated using scaffold-based 5-fold cross-validation. (**Left**) AUC-ROC comparison. Multi-task learning (orange) improved AUC for 5 of 6 endpoints, with the largest gains for cytotoxicity (+3.2%), skin sensitization (+2.9%), and nephrotoxicity (+2.3%). (**Right**) Matthews Correlation Coefficient (MCC) comparison. MCC improvements were most pronounced for hepatotoxicity (+0.069) and Ames mutagenicity (+0.024), reflecting improved sensitivity–specificity balance for imbalanced endpoints. Reproductive toxicity showed no benefit from multi-task learning, consistent with its extremely limited training data (n = 127).

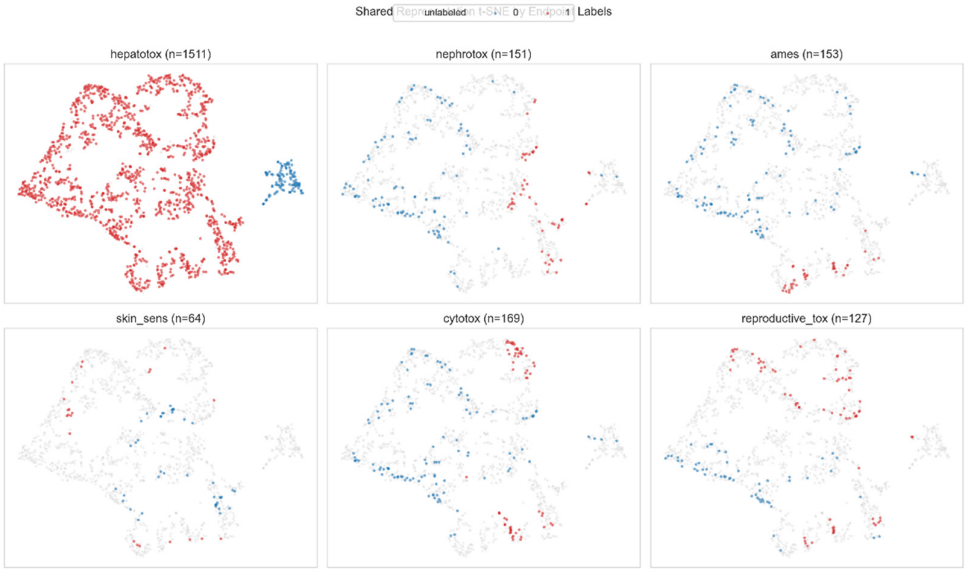


Figure 4. t-SNE visualization of the shared representation, colored by endpoint-specific labels. t-SNE visualization of the shared representation learned by the multi-task stacked generalization framework, colored by endpoint-specific toxicity labels. Each panel displays the same 2D embedding of the unified compound space from the perspective of a different toxicity endpoint. Red points indicate toxic/positive compounds; blue points indicate non-toxic/negative compounds; gray points represent compounds without labels for that specific endpoint. The hepatotoxicity panel (n = 1511) shows the dominant contribution of DILI compounds to the shared chemical space. Partial separation of toxic and non-toxic clusters across multiple endpoints simultaneously

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

suggests that the multi-task framework captures shared toxicity-relevant chemical features. Sample sizes per endpoint reflect the number of labeled compounds available for each task.

3.5. Integrated Risk Scoring with Sensitivity Analysis

Table 6. Risk score validation. High-risk: DILI most-concern, Ames-positive, hERG $pIC_{50} > 6$, cytotoxic. Safe: DILI no-concern, Ames-negative, hERG $pIC_{50} < 5$, non-cytotoxic.

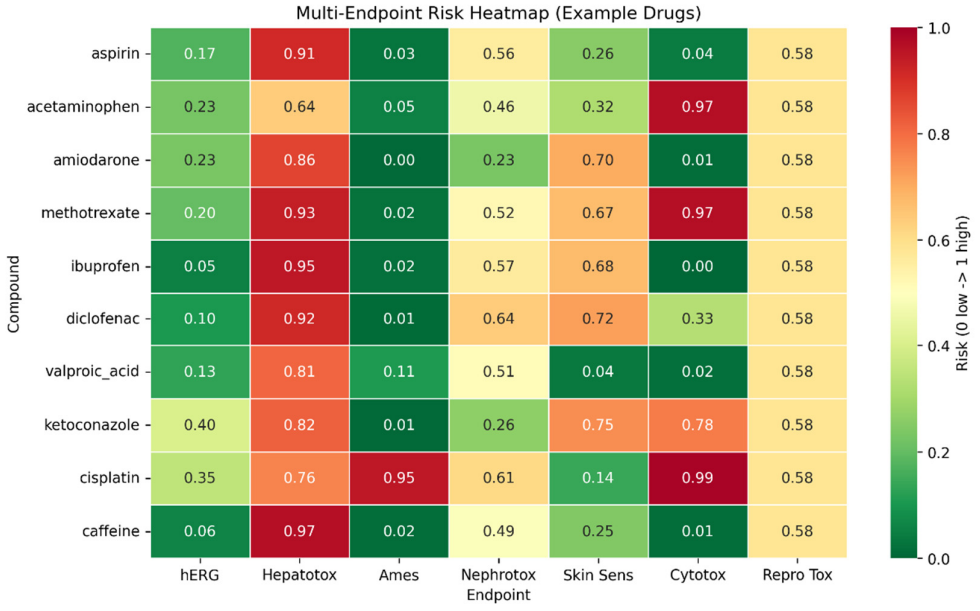
Metric	Value
Mann-Whitney U	8,057
p-value	4.06×10^{-14}
Cliff's δ	0.66 (large effect)
Risk Score AUC	0.83
n (high-risk / safe)	100 / 100

The risk scoring system achieved highly significant discrimination ($p = 4.06 \times 10^{-14}$, AUC = 0.83). Figure 5 illustrates endpoint-specific risk profiles for 10 drugs, demonstrating pharmacologically plausible predictions: cisplatin shows high Ames mutagenicity (0.95) and cytotoxicity (0.99); amiodarone exhibits elevated hepatotoxicity (0.86) and skin sensitization (0.70); caffeine shows low risk across most endpoints.

Sensitivity analysis (Table 7) confirmed robustness across five weight configurations, with AUC ranging from 0.72 (equal weights) to 0.98 (cardiac-focused). All scenarios maintained highly significant discrimination ($p < 5 \times 10^{-8}$). Continuous perturbation analysis (Supplementary Figure S5) revealed that AUC is most sensitive to hERG weight, consistent with the critical regulatory importance of cardiac safety.

Table 7. Risk score sensitivity analysis across weight configurations.

Scenario	AUC	p-value	Cliff's δ
Baseline	0.83	4.4×10^{-16}	0.66
Equal weights	0.72	5.4×10^{-8}	0.44
Cardiac-focused	0.98	8.5×10^{-32}	0.96
Hepatic-focused	0.78	4.2×10^{-12}	0.56
Genotox-focused	0.79	5.1×10^{-13}	0.58



© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

Figure 5. Multi-endpoint risk heatmap for 10 example drugs. Color scale: green (low) to red (high risk). Multi-endpoint risk heatmap for 10 well-characterized drugs demonstrating the platform's ability to generate pharmacologically plausible endpoint-specific toxicity profiles. Each cell displays the predicted risk probability (0 = low risk, green; 1 = high risk, red) for the given compound–endpoint combination. Notable predictions consistent with known pharmacology include: cisplatin showing high Ames mutagenicity risk (0.95) and cytotoxicity (0.99), consistent with its DNA-crosslinking mechanism of action; amiodarone exhibiting elevated hepatotoxicity (0.86) and skin sensitization (0.70), reflecting its documented multi-organ toxicity profile; and ketoconazole displaying high cytotoxicity (0.78) and skin sensitization (0.75), consistent with its CYP450 inhibition and known hepatotoxic potential. The reproductive toxicity column shows constant predictions (0.58) across all compounds due to the model's non-discriminative performance on this endpoint (n = 127), which is excluded from the aggregate risk score.

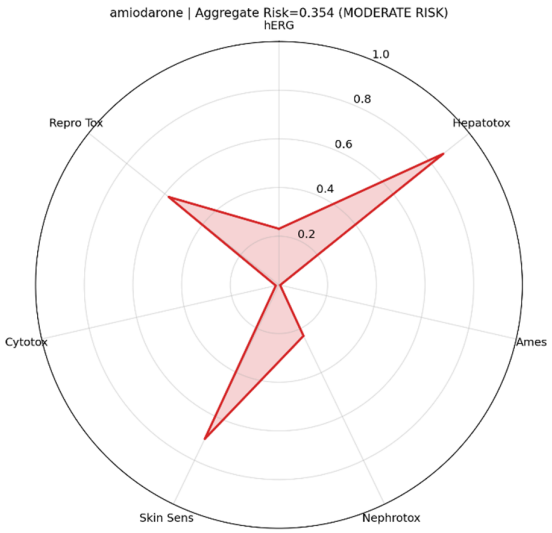


Figure 6. Radar chart showing multi-endpoint risk profile for amiodarone. Radar chart illustrating the multi-endpoint risk profile for amiodarone, a Class III antiarrhythmic agent with well-documented multi-organ toxicity. Each axis represents one of seven toxicity endpoints, with distance from center indicating predicted risk probability (0 = no risk at center; 1 = maximum risk at perimeter). The asymmetric profile reveals elevated hepatotoxicity (0.86) and skin sensitization (0.70) as the dominant predicted liabilities, with moderate hERG risk (0.23) and nephrotoxicity (0.23), low Ames mutagenicity (0.00) and cytotoxicity (0.01), and a non-informative reproductive toxicity prediction (0.58, constant). The aggregate confidence-weighted risk score of 0.354 (MODERATE RISK) integrates these endpoint-specific predictions with clinical importance weights and applicability domain confidence factors. This visualization format enables medicinal chemists to rapidly assess the safety profile of query compounds across multiple toxicity dimensions simultaneously.

3.6. SHAP Feature Importance Analysis

SHAP analysis identified molecular weight, aromatic ring count, LogP, nitrogen count, and TPSA as the top features for hERG inhibition—consistent with established structure–activity relationships [48]. Endpoint-specific SHAP profiles (Supplementary Figure S1) provide actionable insights for medicinal chemistry optimization.

3.7. External Validation on Independent Benchmark Datasets

To assess cross-dataset generalization, models were evaluated on independent benchmark datasets with all overlapping training compounds removed (Table 8).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

Table 8. External validation on independent benchmark datasets. All training set overlaps removed by canonical SMILES matching.

Endpoint	External Dataset	Type	n	AUC/R^2	MCC
Ames	TDC Ames benchmark	External	7183	0.603	0.131
Hepatotox.	TDC DILI benchmark	External	3*	n/a	n/a
Cytotox.	Tox21 SR-MMP	External	5646	0.473	−0.019
hERG	ChEMBL post-2020	Pseudo-ext.	2419	0.038	n/a

** Only 3 non-overlapping compounds; insufficient for meaningful evaluation.*

External validation revealed substantial performance degradation compared to internal validation, consistent with the well-documented domain shift challenge in computational toxicology. The Ames model achieved AUC = 0.603 on the TDC benchmark—reduced from the internal AUC of 0.92—reflecting differences in activity thresholds and assay protocols between datasets. For cytotoxicity, the Tox21 SR-MMP assay measures a specific mechanistic pathway (mitochondrial membrane potential disruption) that fundamentally differs from our general cytotoxicity endpoint (IC_{50} < 10 μM broad-spectrum cell viability), representing an endpoint mismatch rather than a model failure. The hepatotoxicity external validation was precluded by near-complete overlap between the TDC DILI benchmark and our DILIrank-derived training set (236 of 239 compounds). The hERG temporal split showed poor generalization (R^2 = 0.04), suggesting that newer chemical series deposited post-2020 may occupy different regions of chemical space.

These results highlight a fundamental challenge in computational toxicology: the lack of standardized, independent benchmark protocols with consistent assay definitions makes meaningful cross-dataset comparison difficult. Notably, our scaffold-based validation (Section 3.3), which tests generalization to novel scaffolds within consistent assay definitions, provides a more controlled assessment of model generalizability than cross-dataset evaluation with mismatched protocols.

4. Discussion

4.1. Multidimensional Toxicity Profiling in Context

MultiEndpointTox contributes to the growing field of integrated computational toxicity platforms by providing a specific set of capabilities not simultaneously available in existing tools. While ProTox-3.0 [18] offers broader endpoint coverage (61 endpoints) and ADMETlab 3.0 [19] provides comprehensive pharmacokinetic predictions, MultiEndpointTox's specific contributions—symmetric 3D benchmarking, scaffold validation, multi-task learning via stacked generalization, and the integrated risk score—address gaps in the existing toolkit. The confidence-weighted risk scoring system, in particular, bridges the gap between raw ML predictions and medicinal chemistry decision-making by translating multi-endpoint probabilities into interpretable compound-level safety profiles.

4.2. Resolving the 2D versus 3D Descriptor Debate

Our enhanced 3D pipeline (1975 features from 5-conformer ensembles) substantially narrowed the representation gap (AUC 0.833 vs. 0.69–0.73 for basic 3D). This demonstrates that the widely reported large gap was partly attributable to comparing thousands of topological features against a handful of geometric measurements. Nevertheless, 2D descriptors still achieved the highest performance (AUC 0.859), suggesting that topological features capture toxicity-relevant information more efficiently. The practical advantages of 2D descriptors—computational efficiency, determinism, no conformer generation—further support deployment use. Future work should evaluate pharmacophore field methods (CoMFA/CoMSIA [49]), 3D electrostatic surfaces, and shape similarity descriptors [10].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

4.3. Scaffold Validation and Generalization

The average 8% performance drop under scaffold splitting is consistent with known challenges in scaffold hopping for QSAR. Cytotoxicity's robustness suggests general physicochemical toxicity patterns, while hERG's larger drop reflects specific structural requirements for ion channel binding. Reporting both random and scaffold metrics provides transparency aligned with OECD validation principles [12].

4.4. Mechanistic Considerations for Hepatotoxicity Prediction

The hepatotoxicity model's performance plateau around AUC 0.76–0.80 deserves mechanistic contextualization. Drug-induced liver injury is a multifactorial endpoint driven by diverse mechanisms including reactive metabolite formation (mediated by cytochrome P450 enzymes, particularly CYP3A4 and CYP2E1), idiosyncratic immune-mediated hepatotoxicity (dependent on HLA genotype), mitochondrial dysfunction, bile salt export pump (BSEP) inhibition, and oxidative stress [22]. Structure-only QSAR models operating on parent compound descriptors inherently cannot capture metabolism-dependent toxicity pathways—where the toxic species is a metabolite rather than the parent drug—or patient-specific immunological susceptibilities that determine idiosyncratic reactions. This mechanistic ceiling, consistently observed across the DILI prediction literature [50], explains why topological descriptors alone plateau at moderate predictive performance regardless of algorithmic sophistication. Integration of orthogonal data sources—metabolite prediction tools, in vitro mitochondrial toxicity assays, BSEP inhibition screens, and reactive metabolite trapping assays—would be necessary to surpass this ceiling and represents an important future direction.

4.5. Cross-Endpoint Information Transfer

The multi-task framework demonstrated consistent improvements for 5 of 6 endpoints (average +2.1% AUC), validating the hypothesis that toxicity endpoints share chemical information exploitable through cross-endpoint learning. The stacked generalization approach provides a CPU-friendly and interpretable mechanism for information transfer. The largest gains for cytotoxicity (+3.2%), skin sensitization (+2.9%), and nephrotoxicity (+2.3%) are consistent with shared toxicity-relevant molecular features.

4.6. Risk Score Interpretation and Robustness

The risk score's AUC of 0.83 demonstrates that multi-endpoint predictions can be meaningfully aggregated into safety assessments. The sensitivity analysis confirming robustness across weight configurations (AUC 0.72–0.98) is encouraging, though we acknowledge the weights are currently expert-defined. The finding that hERG weighting most strongly influences discrimination is pharmacologically consistent with the critical regulatory importance of QT prolongation liability. Data-driven optimization against withdrawal databases represents a necessary future step.

4.7. Applicability Domain Considerations

The strikingly low AD coverage for skin sensitization (31%) merits mechanistic explanation. Skin sensitization is predominantly driven by electrophilic reactivity—the covalent modification of skin proteins (haptenation) by electrophilic chemicals such as Michael acceptors, acylating agents, and Schiff base formers [25]. Consequently, the LLNA-derived training dataset is highly concentrated in reactive chemical space. When the leverage-based AD method is applied to a general drug-like test set containing primarily non-reactive scaffolds, it correctly flags non-reactive compounds as structurally dissimilar to the reactive training domain. This pharmacologically valid result demonstrates that the AD assessment functions as intended—identifying compounds outside the model's learned chemical space—rather than representing a platform limitation. Users should consider endpoint-specific AD coverage when interpreting predictions.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

4.8. External Validation and Domain Shift

The substantial performance degradation in external validation (Ames AUC 0.60; cytotoxicity AUC 0.47) underscores the domain shift challenge pervasive in computational toxicology. Differences in assay protocols, activity thresholds, chemical space coverage, and endpoint definitions between independent datasets create systematic distributional shifts that may not reflect model quality on the intended chemical domain. The cytotoxicity case is particularly instructive: the Tox21 SR-MMP assay measures mitochondrial membrane potential disruption through a specific molecular mechanism, while our training endpoint captures broad-spectrum cell viability at $IC_{50} < 10$ μM—fundamentally different biological constructs despite sharing the "cytotoxicity" label. These findings argue for the development of standardized, community-accepted external benchmark protocols with consistent endpoint definitions, and support scaffold-based validation (within consistent assay definitions) as a complementary and arguably more informative generalization assessment.

4.9. Limitations

Several limitations should be acknowledged. First, the reproductive toxicity model (n = 127) produces non-discriminative predictions and is labeled as exploratory; this endpoint requires substantially more training data with greater chemical diversity to enable meaningful topological learning. Second, external validation on independent benchmark datasets revealed substantial domain shift effects, highlighting the challenge of cross-dataset generalization in computational toxicology. Third, the hepatotoxicity performance ceiling (~AUC 0.80) reflects the inherent limitation of structure-only QSAR for multifactorial endpoints involving metabolism-dependent and idiosyncratic mechanisms. Fourth, the D-MPNN comparison used default hyperparameters without equivalent optimization, constituting an exploratory rather than definitive comparison. Fifth, 3D descriptors, while substantially enriched, did not employ dynamic conformational ensembles or pharmacophore field methods. Sixth, risk score weights are expert-defined; sensitivity analysis demonstrates robustness but data-driven optimization is needed. Seventh, seven endpoints do not cover all relevant toxicity mechanisms (mitochondrial toxicity, phospholipidosis, phototoxicity).

4.10. Future Directions

Future work will focus on: expanding training data for under-represented endpoints through active learning; implementing deep multi-task architectures with GPU acceleration; incorporating additional endpoints; equitable GNN hyperparameter optimization; richer 3D representations including pharmacophore fields; data-driven risk score weight optimization via Pareto regression against FDA withdrawal data; integration of metabolite prediction for hepatotoxicity; development of standardized external benchmarks; temporal validation; OECD QMRF documentation; and a web-based graphical interface.

5. Conclusions

We present MultiEndpointTox, an integrated chemoinformatics platform for multidimensional drug toxicity profiling. Our systematic evaluation demonstrates that: (1) enhanced multi-conformer 3D descriptors (1975 features) substantially narrow the gap with 2D representations, though 2D remains preferred for deployment; (2) scaffold-based validation provides honest generalization assessment (~8% average performance reduction); (3) multi-task learning improves 5 of 6 endpoints through cross-endpoint information transfer (+2.1% average AUC); (4) a confidence-weighted risk score achieves AUC = 0.83 for discriminating known toxic from safe compounds, with robustness confirmed across weight configurations (AUC 0.72–0.98); and (5) external validation highlights the persistent challenge of cross-dataset domain shift in computational toxicology. By providing consistent methodology, interpretability, applicability domain assessment, and integrated risk scoring within a deployable open-source framework, MultiEndpointTox contributes a practical tool for chemoinformatics-driven drug safety assessment.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

Supplementary Materials: Available online: Section S1 (data curation details), Table S2a–S2b (descriptors and feature counts per endpoint), Table S3a–S3b (hyperparameter search spaces), Table S4 (2D vs. Hybrid statistical comparisons), Table S5 (per-fold CV results), Table S6 (risk weight sensitivity), Section S5 (sensitivity analysis methodology), Figures S1–S5 (ROC curves, regression plots, AD visualization, t-SNE chemical space, weight sensitivity perturbation).

Author Contributions: Conceptualization, S.A.EL.; methodology, S.A.EL.; software, S.A.EL.; validation, S.A.EL.; formal analysis, S.A.EL.; investigation, S.A.EL.; data curation, S.A.EL.; writing—original draft, S.A.EL.; writing—review and editing, S.A.EL. and M.I.Y.; visualization, S.A.EL.; supervision, M.I.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Curated datasets, trained models, and source code are publicly available at https://github.com/sharhabileltahir/MultiEndpointTox.

Use of Artificial Intelligence: During preparation, the authors used Claude (Anthropic) for manuscript drafting assistance. All content was reviewed and edited by the authors, who take full responsibility. No AI tools were used for figures or data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? Nat. Rev. Drug Discov. 2004, 3, 711–715.
2. Onakpoya, I.J.; Heneghan, C.J.; Aronson, J.K. Post-marketing withdrawal of 462 medicinal products. BMC Med. 2016, 14, 10.
3. Russell, W.M.S.; Burch, R.L. The Principles of Humane Experimental Technique; Methuen: London, UK, 1959.
4. European Commission. Directive 2010/63/EU. Off. J. Eur. Union 2010, L276, 33–79.
5. Raies, A.B.; Bajic, V.B. In silico toxicology: computational methods for chemical toxicity prediction. WIREs Comput. Mol. Sci. 2016, 6, 147–172.
6. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity prediction using deep learning. Front. Environ. Sci. 2016, 3, 80.
7. Tropsha, A.; Gramatica, P.; Gombar, V.K. The importance of being earnest: validation for QSPR models. QSAR Comb. Sci. 2003, 22, 69–77.
8. Netzeva, T.I.; Worth, A.; Aldenberg, T.; et al. Current status of methods for defining the applicability domain of (Q)SARs. ATLA 2005, 33, 155–173.
9. Wu, Z.; Ramsundar, B.; Feinberg, E.N.; et al. MoleculeNet: a benchmark for molecular machine learning. Chem. Sci. 2018, 9, 513–530.
10. Axelrod, S.; Gomez-Bombarelli, R. GEOM, energy-annotated molecular conformations. Sci. Data 2022, 9, 185.
11. Chicco, D.; Jurman, G. The advantages of MCC over F1 score and accuracy. BMC Genomics 2020, 21, 6.
12. OECD. Guidance Document on the Validation of (Q)SAR Models; No. 69; OECD Publishing: Paris, France, 2014.
13. Yang, K.; Swanson, K.; Jin, W.; et al. Analyzing learned molecular representations. J. Chem. Inf. Model. 2019, 59, 3370–3388.
14. Withnall, M.; Lindelof, E.; Engkvist, O.; Chen, H. Building attention and edge MPNN. J. Cheminform. 2020, 12, 1.
15. Jiang, D.; Wu, Z.; Hsieh, C.Y.; et al. Could GNNs learn better molecular representation? J. Cheminform. 2021, 13, 12.
16. Walters, W.P.; Barzilay, R. Deep learning in molecule generation and property prediction. Acc. Chem. Res. 2021, 54, 263–270.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

17. Huang, K.; Fu, T.; Gao, W.; et al. Therapeutics Data Commons. Proc. NeurIPS Track Datasets Benchmarks 2021.
18. Banerjee, P.; Kemmler, E.; Dunkel, M.; Preissner, R. ProTox 3.0: prediction of toxicity of chemicals. Nucleic Acids Res. 2024, 52, W513–W520.
19. Xiong, G.; Wu, Z.; Yi, J.; et al. ADMETlab 3.0: an updated ADMET prediction platform. Nucleic Acids Res. 2024, 52, W422–W431.
20. Pires, D.E.V.; Blundell, T.L.; Ascher, D.B. pkCSM: Predicting pharmacokinetic and toxicity properties using graph-based signatures. J. Med. Chem. 2015, 58, 4066–4072.
21. Sanguinetti, M.C.; Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. Nature 2006, 440, 463–469.
22. Chen, M.; Suzuki, A.; Thakkar, S.; et al. DILIrank. Drug Discov. Today 2016, 21, 648–653.
23. Hoste, E.A.J.; Kellum, J.A.; Selby, N.M.; et al. Global epidemiology of acute kidney injury. Nat. Rev. Nephrol. 2018, 14, 607–625.
24. Ames, B.N.; McCann, J.; Yamasaki, E. Methods for detecting carcinogens and mutagens. Mutat. Res. 1975, 31, 347–364.
25. Karlberg, A.T.; Bergstrom, M.A.; Borje, A.; et al. Allergic contact dermatitis. Chem. Res. Toxicol. 2008, 21, 53–69.
26. Riss, T.L.; Moravec, R.A.; Niles, A.L.; et al. Cell Viability Assays. In Assay Guidance Manual; Eli Lilly & NCATS: Bethesda, MD, USA, 2004.
27. Daston, G.P. Laboratory models and teratogenesis. Am. J. Med. Genet. C 2011, 157C, 183–187.
28. Mendez, D.; Gaulton, A.; Bento, A.P.; et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 2019, 47, D930–D940.
29. Richard, A.M.; Judson, R.S.; Houck, K.A.; et al. ToxCast chemical landscape. Chem. Res. Toxicol. 2016, 29, 1225–1251.
30. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: chemical structure curation. J. Chem. Inf. Model. 2010, 50, 1189–1204.
31. Landrum, G. RDKit: Open-Source Cheminformatics. Available online: https://www.rdkit.org (accessed on 1 March 2026).
32. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. J. Chem. Inf. Model. 2010, 50, 742–754.
33. Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL keys. J. Chem. Inf. Comput. Sci. 2002, 42, 1273–1280.
34. Riniker, S.; Landrum, G.A. Better informed distance geometry. J. Chem. Inf. Model. 2015, 55, 2562–2574.
35. Rappe, A.K.; Casewit, C.J.; Colwell, K.S.; et al. UFF force field. J. Am. Chem. Soc. 1992, 114, 10024–10035.
36. Stiefl, N.; Watson, I.A.; Baumann, K.; Bender, A. ErG: 2D pharmacophore descriptions for scaffold hopping. J. Chem. Inf. Model. 2006, 46, 208–220.
37. Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32.
38. Cortes, C.; Vapnik, V. Support-vector networks. Mach. Learn. 1995, 20, 273–297.
39. Chen, T.; Guestrin, C. XGBoost. In Proc. 22nd ACM SIGKDD; ACM: New York, NY, USA, 2016; pp. 785–794.
40. Ke, G.; Meng, Q.; Finley, T.; et al. LightGBM. Adv. Neural Inf. Process. Syst. 2017, 30, 3146–3154.
41. Akiba, T.; Sano, S.; Yanase, T.; et al. Optuna. In Proc. 25th ACM SIGKDD; ACM: New York, NY, USA, 2019; pp. 2623–2631.
42. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE. J. Artif. Intell. Res. 2002, 16, 321–357.
43. Heid, E.; Greenman, K.P.; Chung, Y.; et al. Chemprop. J. Chem. Inf. Model. 2024, 64, 9–17.
44. Bemis, G.W.; Murcko, M.A. Properties of known drugs. 1. Molecular frameworks. J. Med. Chem. 1996, 39, 2887–2893.
45. Huang, K.; Fu, T.; Gao, W.; et al. Therapeutics Data Commons. Proc. NeurIPS Track Datasets Benchmarks 2021.
46. Gramatica, P. Principles of QSAR models validation. QSAR Comb. Sci. 2007, 26, 694–701.
47. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 2017, 30, 4765–4774.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2026 doi:10.20944/preprints202603.0867.v1

48. Aronov, A.M. Predictive in silico modeling for hERG blockers. Drug Discov. Today 2005, 10, 149–155.
49. Cramer, R.D.; Patterson, D.E.; Bunce, J.D. CoMFA. J. Am. Chem. Soc. 1988, 110, 5959–5967.
50. Thakkar, S.; Li, T.; Liu, Z.; et al. DILIst: binary classification of 1279 drugs. Drug Discov. Today 2020, 25, 201–208.

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

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.