Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Data Source Heterogeneity, Not Algorithm Choice, Drives Performance Gaps in QSAR for Mutagenicity

Seungmin Yoon
Kyungsung University
Sanghun Kim
Kyungsung University
Hyunpyo Jeon
hpjeon@ks.ac.kr

Kyungsung University

Research Article

Keywords:

Posted Date: July 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10122807/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Data Source Heterogeneity, Not Algorithm Choice, Drives Performance Gaps in QSAR for Mutagenicity

Seungmin Yoon, Sanghun Kim, and Hyunpyo Jeon*

Department of Biosafety, Kyungsung University, Busan, Republic of Korea

E-mail: hpjeon@ks.ac.kr

Abstract

QSAR models for Ames mutagenicity are increasingly used in regulatory risk assessment under EU REACH, K-REACH, and ICH M7, but their reported performance varies widely across evaluation protocols.[1–3] In this study, 225 model configurations—seven algorithms, five feature sets, and three preprocessing scenarios—were systematically evaluated across three genotoxicity datasets: Ames (n = 13,560), *in vitro* chromosome aberration (n = 1,619), and *in vivo* micronucleus (n = 2,153). Using a fixed model configuration (XGBoost, compact features, raw SMILES), Ames MCC dropped from 0.670 under random-split CV to 0.624 under scaffold-split CV. Cross-source leave-domain-out (LDO) evaluation produced markedly larger declines: 0.234 (drug→industrial) and 0.122 (industrial→drug). The source heterogeneity gap ($\Delta = 0.390$) considerably exceeded the split-method gap ($\Delta = 0.046$). A simple source-based classifier reached MCC = 0.608 with no structural information, reflecting a 19.5-fold difference in positive rates between drug-sourced (58.3%) and industrial-sourced (3.0%) compounds. Threshold decomposition showed that prior shift accounted for only part of this gap, with oracle MCC plateauing at 0.27–0.36. Algorithm robustness

under source shift varied considerably: Random Forest maintained MCC = 0.325 in the hardest direction, whereas XGBoost collapsed to 0.122. *In vivo* micronucleus models showed strong ranking performance (ROC-AUC = 0.860, PR-AUC = 0.237) but unreliable classification (MCC 95% CI crossed zero). These results call for a source-aware evaluation cascade as a standard reporting requirement for mutagenicity QSAR.

Scientific contribution

A source-aware evaluation cascade is introduced to disentangle the major drivers of performance variation in genotoxicity QSAR under a single fixed model configuration. Data source heterogeneity far outweighs the effects of split strategy and in-domain algorithm choice, while threshold decomposition shows that prior shift accounts for only part of the cross-source performance loss. Together, these findings establish source-aware evaluation as an essential component of regulatory mutagenicity QSAR validation.

Introduction

Under EU REACH[4] and K-REACH,[5] industrial chemical manufacturers and importers must conduct a chemical safety assessment of the genotoxic potential of their compounds as part of registration. As a New Approach Methodology (NAM), in silico QSAR predictions are accepted as supporting evidence when models are validated and cover the relevant applicability domain.[6] For pharmaceuticals, ICH M7(R2) similarly endorses QSAR alongside rule-based expert systems for evaluating mutagenic impurities.[7] In both contexts, the Ames bacterial reverse mutation assay[8,9] is the primary prediction target, with *in vitro* chromosome aberration (CA) and *in vivo* micronucleus (MN)[10,11] as higher-tier endpoints. Improved predictive models could reduce animal testing under the 3Rs principle.[12]

Reported accuracies (0.80–0.90) and MCC values (0.60–0.80) across published Ames QSAR models[1,13–15] vary widely—but these values come from different datasets, splits, preprocessing pipelines, and evaluation protocols, making direct comparison difficult. At least

three factors can inflate apparent performance.

Splitting strategy. Random splits allow structurally similar compounds (those sharing a Murcko scaffold[16]) to appear in both training and test sets, producing optimistic estimates. Scaffold-based splitting[3,17] provides a more realistic assessment, but the magnitude of the random-vs-scaffold gap has not been measured systematically across genotoxicity endpoints.

Source mixing. Large Ames datasets typically combine compounds from pharmaceutical libraries and industrial chemical registries,[1,13] which differ in chemical composition, positive rates, data quality, and labeling conventions. When such datasets are pooled, models may learn to distinguish compounds by source rather than learning toxicity-related structure–activity relationships—a form of data source heterogeneity that has been noted[18,19] but never systematically quantified. One component of this heterogeneity is prior shift: the large difference in positive rates between sources can miscalibrate the default decision threshold independently of any structural difference. However, threshold decomposition analysis presented here shows that prior shift explains only part of the performance gap; a substantial residual remains even after recalibration.

Preprocessing. Steps such as salt stripping, metal removal, and charge normalization can substantially alter molecular representations,[20,21] but are often undocumented or applied uniformly across endpoints.

Most published studies evaluate a single endpoint with a single split, focus on algorithm comparison,[22,23] rarely test cross-source generalization, and seldom report statistical significance.[24] The relative contributions of split method, source composition, and preprocessing have not been disentangled in a single controlled study.

This study addresses these gaps by:

1. Introducing an evaluation cascade—scaffold-split test, random vs. scaffold CV, and cross-source leave-domain-out—applied consistently under a single fixed configuration.

2. Quantifying the contribution of data source heterogeneity—only partly attributable to prior shift—which produces a scaffold-to-LDO MCC drop of 0.390, considerably larger

than the effect of split method ($\Delta = 0.046$).

3. Comparing in-domain and cross-source algorithm robustness: in-domain differences are small (MCC range = 0.063), whereas cross-source performance varies markedly (RF 0.325 vs. XGB 0.122).

4. Characterizing endpoint-specific preprocessing effects: salt stripping improved the best *in vivo* model by +0.22 MCC while reducing *in vitro* performance.

5. Releasing an open-source pipeline for full reproducibility.

Materials and Methods

Datasets

Three datasets were evaluated (Table 1): Ames bacterial reverse mutation ($n = 13{,}560$), *in vitro* chromosome aberration ($n = 1{,}619$), and *in vivo* micronucleus ($n = 2{,}153$).

The Ames dataset combines pharmaceutical (drug-like) and industrial chemical sources. Compounds carry a source label based on their originating dataset; the term "source" is used here to emphasize that this label reflects data provenance—encompassing potential differences in chemical space, data quality, and labeling conventions—rather than a computed chemical classification. Drug-sourced compounds ($n = 6{,}416$) had a positive rate of 58.3%, while industrial-sourced ($n = 7{,}144$) had 3.0%, a 19.5-fold difference. The *in vitro* and *in vivo* datasets lacked source annotations.

Raw study records for all three endpoints were first mapped to endpoint categories (Ames, *in vitro* CA, *in vivo* MN) using a unified classification table spanning OECD, EPA, and EU Test Guidelines; the full guideline-to-endpoint mapping is provided in the supplementary materials. For industrial chemicals sourced from ECHA dossiers, inclusion was restricted to studies with Klimisch reliability scores of 1 (reliable without restriction) or 2 (reliable with restrictions) where available,[25] excluding records with score 3 or 4. When a single compound

had both positive and negative study records of equal Klimisch score, a conservative hazard-assessment convention consistent with EU REACH and K-REACH guidance[4,5] was applied, and the compound was labeled as positive. Compounds with irreconcilable label conflicts across distinct sources were excluded (169 Ames, 3 *in vitro*, 2 *invivo*).

Cross-endpoint overlap analysis revealed substantial sharing: 97.5% of *in vitro* and 92.1% of *in vivo* compounds also appeared in Ames, with moderate label concordance (Cohen's $\kappa = 0.14$–0.20).

All datasets were split 80/20 using scaffold-based splitting[17] with zero scaffold-group overlap. Murcko scaffolds[16] were generated with RDKit;[26] acyclic compounds were clustered by agglomerative clustering on Morgan fingerprint[27] similarity.

Table 1: Dataset summary. Positive rate refers to the training partition. The 225 total configurations comprise 3 datasets × 3 preprocessing scenarios × (6 tabular models × 4 feature modes + 1 GNN).

Dataset	Assay	n	Train	Test	Pos. rate (%)	Scaffolds
Ames	Bacterial reverse mutation	13,560	10,848	2,712	29.1	1,691
In vitro CA	Chromosome aberration	1,619	1,295	324	10.4	353
In vivo MN	Micronucleus	2,153	1,722	431	3.7 / 2.8[a]	410

[a]Training / test positive rate.

Preprocessing, Features, and Algorithms

Each dataset was processed under three scenarios. **Raw**: original SMILES with no modification. **Salt-stripped**: the largest fragment was retained using RDKit's `LargestFragmentChooser`[26] followed by charge normalization, which modified 45% of Ames, 81% of *in vitro*, and 83% of *in vivo* SMILES. **No-metal**: compounds containing metal atoms were excluded.

Five feature representations were evaluated. **Compact** (138 features) combined 13 physicochemical descriptors (MW, LogP, TPSA, HBD, HBA, rotatable bonds, fraction C_{sp3}, ring counts, heteroatoms, heavy atoms, formal charge, valence electrons, aromatic ring count) with 125 structural alerts based on Benigni–Bossa SA1–SA31,[28,29] Kazius toxicophores,[30]

and ISS Ames alerts. **Broad FP-512/1024/2048** (266/394/522 features) extended compact with 128/256/384 Morgan fingerprint bits[27] selected by mutual information[31] on the training set only (bits with prevalence below 1% or above 99% excluded). **Graph** representations used molecular graphs with 9-dimensional atom features and normalized adjacency matrices, paired exclusively with the GCN.

Seven classifiers were evaluated: XGBoost (XGB),[32] LightGBM (LGBM),[33] Random Forest (RF),[34] SVM with RBF kernel,[35] Logistic Regression (LR),[36] a 128–64 MLP,[37] and a three-layer GCN[38] with LayerNorm and global mean pooling. SVM and LR used standard scaling. Class imbalance was handled via `scale_pos_weight` or `class_weight='balanced'`. Hyperparameters were tuned via `RandomizedSearchCV` with `GroupKFold`.[39]

Evaluation Cascade

To compare evaluation methods under identical conditions, a single fixed model configuration—XGB, compact features, raw SMILES—was applied at all three cascade levels (Figure 1):

Level 1 (in-domain test): the scaffold-split held-out test set, with bootstrap 95% CI (500 iterations) on MCC,[40] ROC-AUC, PR-AUC, balanced accuracy, sensitivity, specificity, and Brier score.[41]

Level 2 (split-method comparison): 5-fold stratified random CV and 5-fold scaffold-grouped CV, each with 5 repeats, on the full dataset.

Level 3 (cross-source LDO): for Ames, models were trained on one source and tested on the other, using the fixed configuration plus LGBM and RF to assess algorithm-specific robustness.

Additional Analyses

Beyond the fixed configuration, all 225 combinations were evaluated on the scaffold-split test set ($3 \times 3 \times [6 \times 4 + 1] = 225$). Applicability domain (AD) was assessed using Tanimoto similarity on 1024-bit Morgan fingerprints[27] with a threshold of ≥ 0.3. SHAP values[42]

(`TreeExplainer`) were computed for tree-based models on the three primary endpoints. Learning curves were generated at training fractions from 10% to 100% using 5-fold scaffold CV under the fixed configuration. McNemar's test[43] with continuity correction was used to compare the top two models per endpoint.

LDO threshold decomposition. To separate the effect of prior shift from other sources of performance loss, LDO MCC was evaluated at three decision thresholds: *default* ($t = 0.5$), *prior-corrected* (t set to the training positive rate, which is optimal assuming pure prior shift), and *oracle* (t chosen to maximize MCC on the test set, providing an upper bound on threshold-based recalibration).

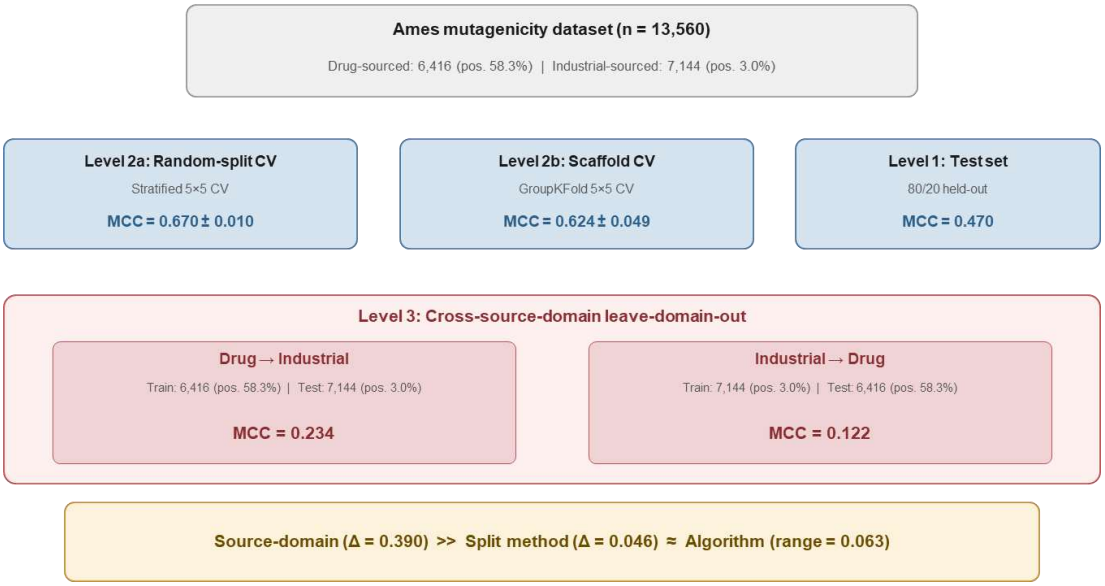


Figure 1: Evaluation cascade framework. All levels use the same fixed configuration (XGB, compact features, raw SMILES), enabling fair MCC comparison across evaluation methods.

Results

Best Observed Performance Across the Full Grid

Table 2 reports the best observed performance per endpoint across all 225 configurations. These values are an upper bound from extensive search and should be read as a descriptive benchmark, not unbiased estimates; the confirmatory analysis uses the fixed configuration (Table 4).

Table 2: Best observed performance per endpoint (scaffold-split test set, full grid). Best among 225 configurations; descriptive benchmark, not unbiased estimates.

Endpoint	Algo.	Preproc.	Features (n)	MCC [95% CI]	AUC	PR-AUC	Sens.	Spec.
Ames	XGB	raw	FP-2048 (522)	0.555 [.522,.588]	0.850	0.765	0.756	0.822
In vitro	LGBM	salt_str.	FP-1024 (394)	0.297 [.131,.450]	0.665	0.278	0.349	0.929
In vivo	XGB	salt_str.	FP-512 (266)	0.311 [−.014,.571]	0.860	0.237	0.319	0.983

PR-AUC baselines (prevalence): Ames 0.302, *in vitro* 0.124, *in vivo* 0.028.

Under fixed features and preprocessing (compact, raw SMILES), all seven algorithms on Ames spanned an MCC range of only 0.063 (Table 3, Figure 2); the five tabular models excluding LR spanned 0.023. Pairwise McNemar tests[43] between the top two models found no significant differences (Ames $p = 0.39$; *in vivo* $p = 0.72$). This in-domain equivalence did not extend to cross-source settings (see *Data Source Heterogeneity and Algorithm-Specific Robustness*).

The GCN underperformed fingerprint-based approaches across endpoints (Ames mean MCC 0.402 vs. 0.507 for XGB), consistent with prior reports that graph networks require considerably larger datasets to outperform engineered features.

Evaluation Cascade

For Ames, random-split CV reached MCC = 0.670 ± 0.010, in line with published benchmarks,[13,15] while scaffold-split CV fell to 0.624 ± 0.049 ($\Delta = 0.046$) and the held-out scaffold-based test set to 0.470 (Table 4, Figure 3). Cross-source LDO revealed a markedly larger

Table 3: All-algorithm comparison (compact features, raw SMILES, Ames scaffold-split test). Range across all 7 algorithms: 0.063; range excluding LR and GCN: 0.023.

Algorithm	MCC	ROC-AUC
RF	0.492	0.838
ANN (MLP)	0.489	0.834
LGBM	0.479	0.833
SVM (RBF)	0.470	0.832
XGB	0.470	0.825
GCN	0.445	0.815
LR	0.430	0.800

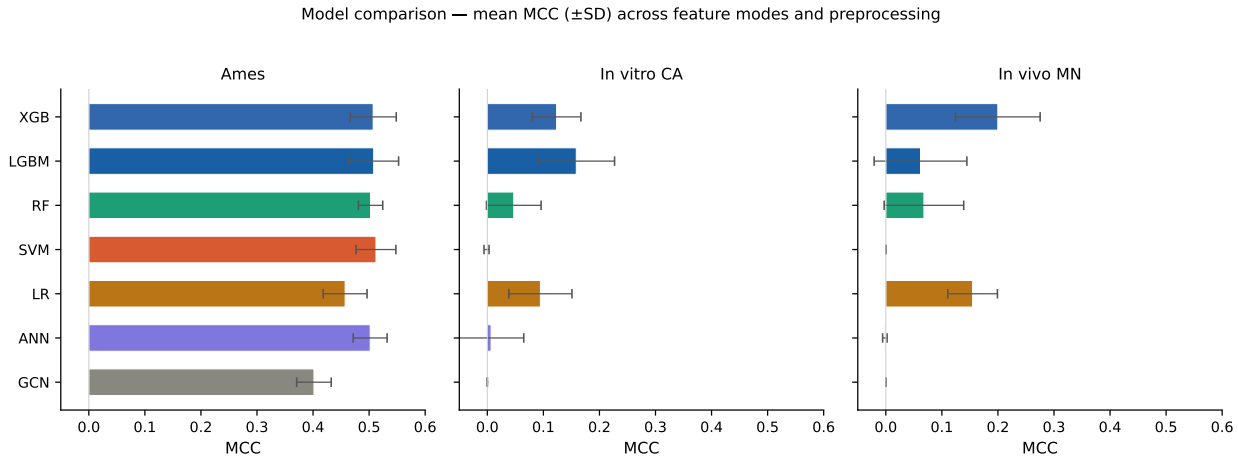


Figure 2: Mean MCC (±SD) by algorithm across all feature modes and preprocessing scenarios, grouped by endpoint.

decline: drug→industrial yielded MCC = 0.234 and industrial→drug yielded MCC = 0.122. The source heterogeneity gap ($\Delta = 0.390$) was considerably larger than the split-method gap ($\Delta = 0.046$). The LDO values also reflect differences in test set size and class distribution across directions, so $\Delta = 0.390$ captures a combined effect rather than source heterogeneity alone—but the relative magnitude of the two effects is unambiguous. For *in vitro* CA, random and scaffold splits yielded nearly identical results (0.180 vs. 0.189); for *in vivo* MN, both values were near chance level given their standard deviations.

Table 4: Evaluation cascade (fixed configuration: XGBoost, compact features, raw SMILES). Level 2: mean ± SD (25 folds). Level 1: MCC [95% bootstrap CI]. Level 3: Ames only.

Level	Evaluation method	Ames MCC	In vitro MCC	In vivo MCC
2a	Random-split 5×5 CV	0.670 ± 0.010	0.180 ± 0.053	0.181 ± 0.094
2b	Scaffold-split 5×5 CV	0.624 ± 0.049	0.189 ± 0.084	0.104 ± 0.111
1	Scaffold-split test	0.470 [.434,.502]	0.114 [−.023,.270]	0.071 [−.029,.285]
3a	LDO: drug → industrial	0.234	—	—
3b	LDO: industrial → drug	0.122	—	—

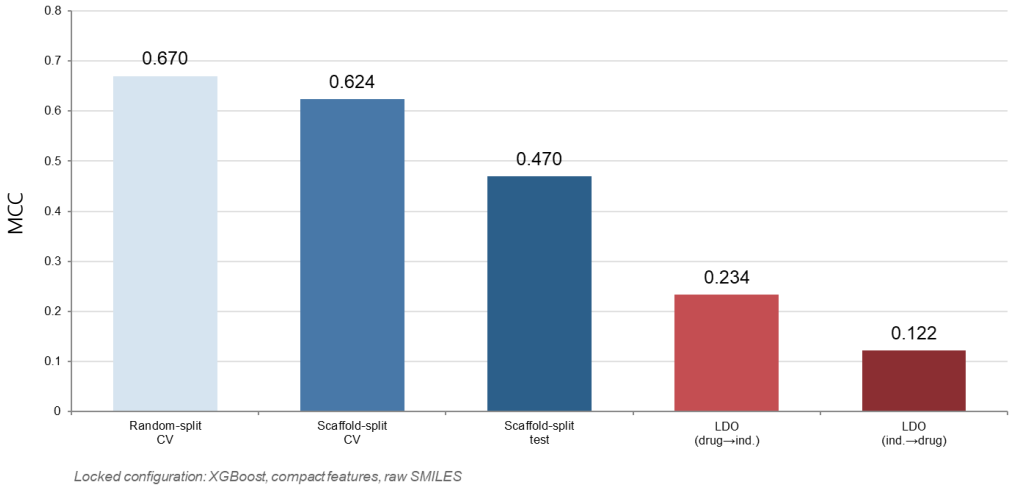


Figure 3: Evaluation cascade for Ames mutagenicity (fixed configuration). The data source heterogeneity effect ($\Delta = 0.390$) clearly exceeds the split-method effect ($\Delta = 0.046$).

Data Source Heterogeneity and Algorithm-Specific Robustness

Drug-sourced compounds (n = 6,416, positive rate 58.3%) and industrial-sourced compounds (n = 7,144, positive rate 3.0%) differed by a factor of 19.5 in mutagenicity prevalence (Figure 4). A naive classifier predicting positive for drug-sourced and negative for industrial-sourced compounds achieved MCC = 0.608 and balanced accuracy = 0.834 without any structural information—approaching the random-split CV MCC of 0.670.

Comparing three algorithms under LDO (Table 5) revealed two distinct patterns. Drug→industrial generalization was consistent across all algorithms (MCC 0.232–0.235). The industrial→drug direction exposed sharp differences: XGB reached only MCC = 0.122 (sensitivity 0.062) and

LGBM 0.134, while RF maintained MCC = 0.325 (sensitivity 0.466). We hypothesize that RF's better cross-source robustness may reflect random feature subsampling at each split (random subspace method),[34] which reduces reliance on source-correlated feature subsets.

Table 5: Leave-domain-out results by algorithm (Ames). Drug: $n = 6{,}416$, pos. rate 58.3%. Industrial: $n = 7{,}144$, pos. rate 3.0%. Naive source classifier: MCC = 0.608.

Algorithm	Drug → Ind. MCC	Ind. → Drug MCC	Ind. → Drug Sens.
XGB	0.234	0.122	0.062
LGBM	0.232	0.134	0.081
RF	0.235	0.325	0.466

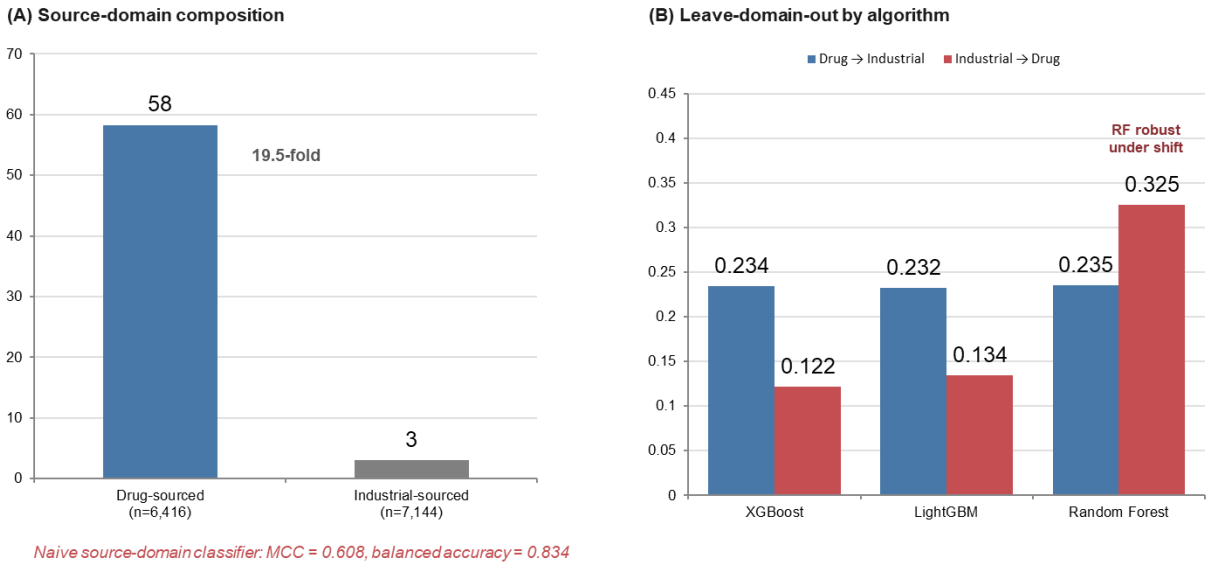


Figure 4: Data source heterogeneity in Ames. (A) Positive rate by source: drug 58.3% vs. industrial 3.0% (19.5-fold). (B) LDO MCC by algorithm; RF was appreciably more robust in the industrial→drug direction.

Preprocessing, *In Vivo* Behavior, and Feature Importance

Preprocessing. Salt stripping effects were strongly endpoint-specific (Figure 5). For *in vivo* MN, the best salt-stripped model reached MCC = 0.311 versus 0.086 for the best

raw model (+0.225). Salt fragments were present in 83% of *in vivo* SMILES. Among salt-containing compounds, the positive rate was 2.8% versus 7.1% for non-salt compounds, indicating that salts were disproportionately associated with non-mutagenic structures. For Ames, salt stripping had mixed effects ($\Delta = -0.03$ to $+0.02$). For *in vitro* CA it was generally detrimental (mean MCC 0.081 $\rightarrow$ 0.056). These results argue against uniform preprocessing across endpoints and support endpoint-specific optimization.

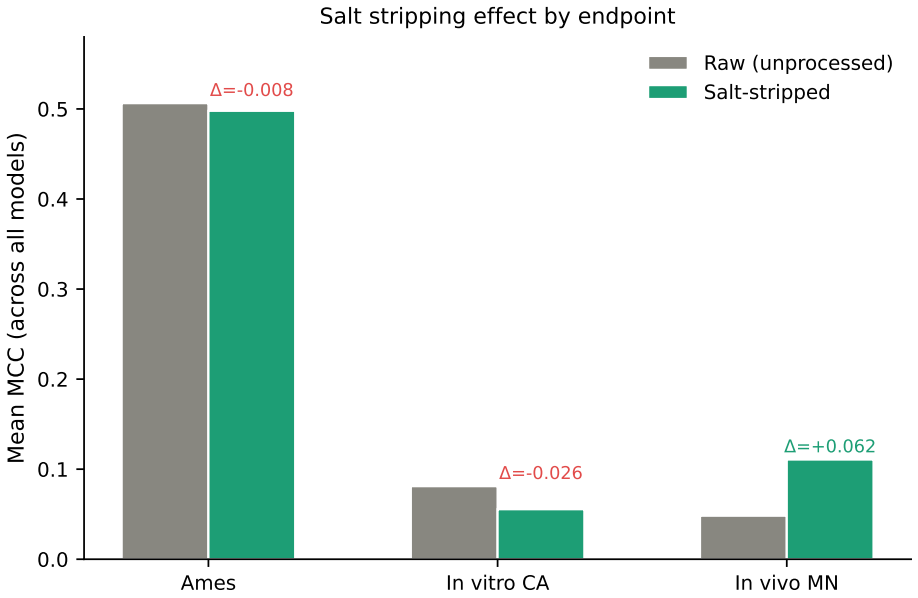


Figure 5: Effect of salt stripping on mean MCC (averaged across all models and feature modes). *In vivo* improved by +0.063 on average and +0.225 for the best model; Ames mixed; *in vitro* decreased.

***In vivo* micronucleus.** The best *in vivo* model achieved ROC-AUC = 0.860 and PR-AUC = 0.237 (an 8.5-fold enrichment over the 0.028 baseline), but MCC = 0.311 with a 95% CI that crossed zero (-0.014 to 0.571), meaning classification at any fixed threshold is unreliable. The test set contains only 12 positives among 431 compounds. The fixed-configuration learning curve was flat at MCC $\approx$ 0.055 from 10% to 100% of training data (Figure 7), though this curve was not computed for the salt-stripped configuration.

Feature importance. SHAP analysis[42] revealed consistent patterns (Figure 6). Physicochemical descriptors dominated the top-5 at every endpoint (alert count and MW for Ames;

LogP for *in vitro*; MW for *in vivo*). No individual fingerprint bit appeared in any top-5 list, despite fingerprints comprising 50–75% of features. Across all Ames configurations, compact features yielded mean MCC = 0.448 versus 0.523 for FP-2048 ($\Delta = +0.075$), indicating that fingerprint bits contribute weak signal individually but accumulate meaningful predictive information collectively.

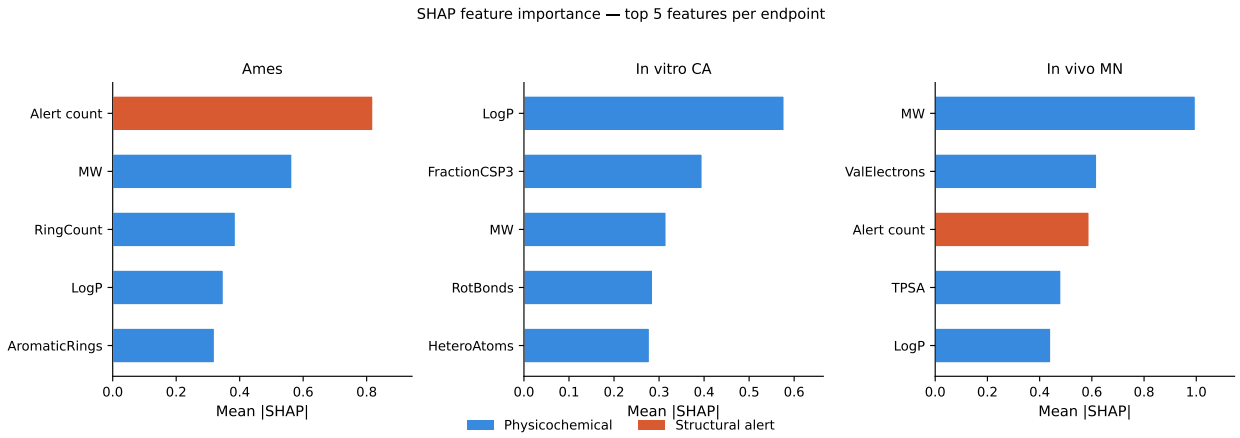


Figure 6: SHAP top-5 features per primary endpoint. Physicochemical descriptors dominate; structural alert count ranks first for Ames. No fingerprint bits appear in any top-5.

Learning curves. Fixed-configuration learning curves (Figure 7) showed three patterns: Ames MCC saturated early (0.543 at 10%, 0.640 at 100%); *in vitro* showed a noisy upward trend (0.070 → 0.170); *in vivo* remained flat (≈ 0.055).

Discussion

The Performance Variation Hierarchy

The fixed-configuration analysis yielded a clear hierarchy:

$$\text{Data source heterogeneity} \gg \text{Splitting method} \approx \text{In-domain algorithm choice}$$

The source heterogeneity gap ($\Delta = 0.390$) substantially exceeded the split-method gap ($\Delta = 0.046$). The split-method gap, in turn, was comparable to the in-domain algorithm

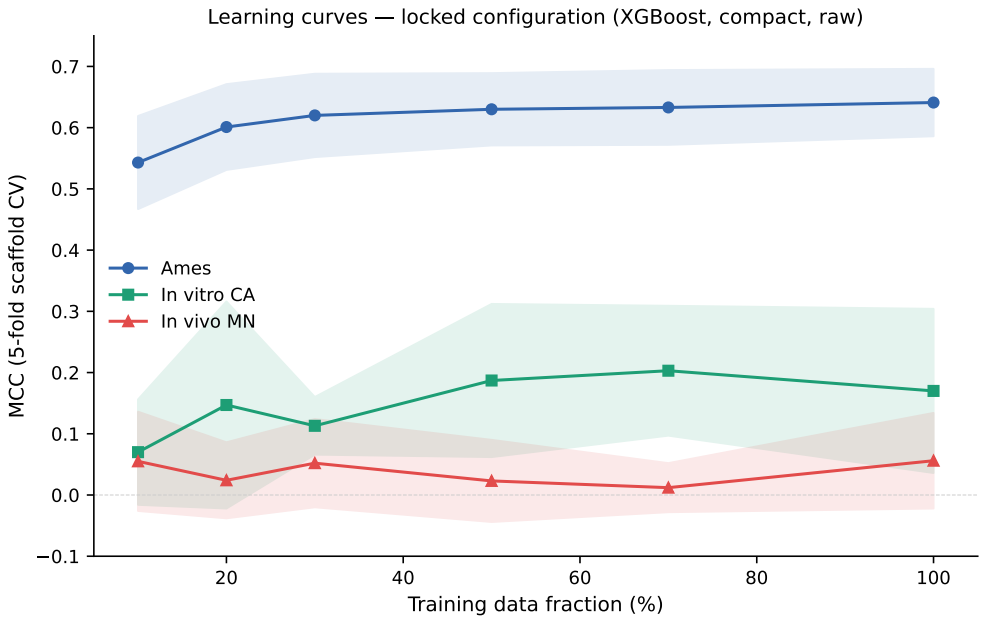


Figure 7: Learning curves (fixed configuration). Shaded regions: ±1 SD across 5-fold scaffold CV.

spread (MCC range = 0.063). Three caveats apply. First, the hierarchy was established on a single Ames benchmark. Second, the LDO Δ reflects differences in test set size and class distribution across directions and should therefore be treated as an upper bound. Third, algorithm robustness varied under source shift; algorithm choice may thus matter more in cross-source deployment than the in-domain figures suggest.

The LDO experiment should be interpreted as a diagnostic stress test, not a realistic deployment scenario. In practice, models are trained on all available data regardless of provenance. Mutagenic potential is a property of molecular structure, not database origin. A large LDO drop suggests that apparent in-domain performance depends strongly on the joint composition of the training and test sets. This pattern suggests incomplete learning of broadly generalizable structure-activity relationships. This is a concern for regulatory applications: new chemicals registered under EU REACH or K-REACH may occupy regions of chemical space that are underrepresented in pharmaceutical benchmarks, and a benchmark MCC of 0.67 on a mixed-source test set provides limited assurance in that context. Cross-source LDO reveals this dependence and offers a more conservative estimate of

out-of-distribution performance.

What Drives the Data Source Heterogeneity Gap?

Four non-exclusive factors may contribute, listed from most to least parsimonious:

1. **Prior shift.** The 19.5-fold difference in positive rates between sources miscalibrates the default decision threshold. This is consistent with XGB achieving a sensitivity of only 0.062 in the industrial→drug direction.

2. **Data quality and labeling differences.** Drug-sourced compounds come from curated benchmarks; industrial compounds come from regulatory submissions (ECHA, K-REACH) with varying protocols. Prevalence differences may therefore arise independently of underlying chemical differences.

3. **Selection bias.** Drug candidates are optimized for biological activity, while industrial chemicals are selected for functional or commercial utility. These represent fundamentally different sampling criteria.

4. **Chemical space divergence.**[44,45] Even a perfectly calibrated model may generalize poorly if the two source populations occupy distinct regions of chemical space.

To estimate the contribution of prior shift, each LDO direction was evaluated at three thresholds: *default* ($t = 0.5$), *prior-corrected* (t set to the training positive rate, optimal under pure prior shift), and *oracle* (t maximizing MCC on the test set, an upper bound on recalibration). This analysis used Morgan FP-1024 plus 13 physicochemical descriptors rather than the full 138-feature compact set, so absolute MCC values differ from Table 5 (notably RF industrial→drug: 0.102 here vs. 0.325 in the main analysis). The three-threshold comparison within each row is nonetheless internally valid.

The results (Table S7) show that prior-shift correction gave modest gains (mean +0.075; range −0.024 to +0.239). Oracle MCC topped out at 0.27–0.36, well below the in-domain

scaffold-CV MCC of 0.624. This leaves a residual gap of 0.26–0.35 that threshold adjustment alone cannot close. Sensitivity to threshold choice also varied by algorithm: RF gained +0.243 MCC from default to oracle in the industrial→drug direction, while gradient boosting models gained only +0.07 to +0.10. Prior shift thus accounts for part of the LDO performance loss, but a substantial residual gap remains. The practical implication holds regardless of the underlying cause: published Ames QSAR performance is highly sensitive to source composition, and recalibration alone is insufficient under cross-source evaluation.

In Vivo Micronucleus and Tiered Testing

In vivo MN models showed strong ranking performance (ROC-AUC = 0.860, PR-AUC = 0.237), but classification at any fixed threshold was unreliable (MCC 95% CI crossed zero). This indicates that current 2D-descriptor models are not suitable for regulatory binary hazard decisions under ICH M7[7] or REACH.[4] The 8.5-fold enrichment over baseline prevalence, however, suggests potential utility as a prioritization tool within a tiered screening workflow.[46]

Performance was also highly sensitive to preprocessing: salt stripping improved MCC by +0.22. These findings indicate that reported *in vivo* performance can differ appreciably depending on undocumented preprocessing choices.[21] Mechanistically, micronucleus induction is governed by ADME, metabolic activation, tissue distribution, and DNA repair[47,48]—processes that 2D descriptors do not represent.

Cross-Endpoint Overlap and Preprocessing

More than 90% of the *in vitro* and *in vivo* compounds also appear in the Ames dataset. Performance differences across endpoints (Ames MCC = 0.555 vs. *in vivo* 0.311) therefore most likely reflect intrinsic endpoint predictability from 2D structure, not differences in chemical coverage.[18] Multi-task learning may be a productive avenue for future work.[49]

Preprocessing effects were endpoint-specific: salt stripping improved *in vivo* performance,

produced mixed results for Ames, and reduced *in vitro* performance. The reduction for *in vitro* may reflect clastogenic activity of some counterions. Although standardized preprocessing pipelines are widely recommended,[20,21] the present results indicate that endpoint-specific optimization is preferable.

Practical Recommendations

The following reporting standards are proposed for genotoxicity QSAR:

1. Use scaffold-based splitting.
2. When source labels are available, report cross-source LDO performance with disclosed source composition.
3. Optimize preprocessing separately for each endpoint.
4. Report bootstrap CIs and applicability domain coverage.
5. For class-imbalanced endpoints, report both threshold-free (ROC-AUC, PR-AUC) and threshold-dependent (MCC, balanced accuracy) metrics.
6. Test algorithm superiority claims under source shift, not only in-domain.

Several limitations apply. Only a single scaffold split was used. The GCN lacked edge features and pretraining.[50] Three-dimensional descriptors, metabolite predictions, within-AD vs. out-of-AD stratification for LDO, and multiple-comparison correction for McNemar tests[24] were not evaluated.

Conclusion

Applying an evaluation cascade to 225 model configurations across three genotoxicity datasets shows that, for the Ames mutagenicity benchmark studied here, mixing drug-like and industrial compounds—which differ sharply in positive rate and data provenance—produces a

source heterogeneity gap ($\Delta = 0.390$, scaffold CV to LDO) that is considerably larger than what comes from choosing between random and scaffold splitting ($\Delta = 0.046$). Threshold decomposition attributed only part of this gap to prior shift; a meaningful residual remains that recalibration cannot fix. Algorithm differences are small within a single source domain (MCC range = 0.063) but become pronounced under source shift, with RF performing considerably better than gradient boosting. *In vivo* micronucleus models can rank compounds but do not yet provide reliable binary classification with 2D descriptors. Taken together, these results call for a source-aware evaluation cascade as a standard reporting requirement for mutagenicity QSAR.

Data and Software Availability

The curated datasets used in this study—Ames mutagenicity (n = 14,771), *in vitro* chromosomal aberration (n = 1,656), and *in vivo* micronucleus (n = 2,199)—including molecular structures (SMILES), experimental labels, and data source annotations (drug-domain vs. industrial-domain), are provided as Supporting Information in machine-readable CSV format. The complete pipeline code (v12.1), configuration files, scaffold split assignments (global seed = 42), and all locked results tables are available at `https://github.com/smyoon96/Genotox_model_v2` (DOI: [to be assigned upon acceptance]). The repository includes step-by-step instructions for exact reproduction of all reported numbers. All analyses were performed using open-source software: RDKit (2023.09), scikit-learn (1.3), XGBoost, LightGBM, and PyTorch.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

acknowledgement

The author has no acknowledgements to make.

Funding

This research did not receive funding.

Authors' contributions

1. Seungmin, Yoon : Designed the study, curated the datasets, developed the pipeline, conducted all analyses, and wrote the manuscript

2. Sanghun Kim : Contributed to data curation and manuscript review

3. Hyunpyo Jeon : Supervised the study and reviewed the manuscript. All authors reviewed and approved the final manuscript.

References

(1) Honma, M.; Kitazawa, A.; Cayley, A.; Williams, R. V.; Barber, C.; Hanser, T.; Saiakhov, R.; Chakravarti, S.; Myatt, G. J.; Cross, K. P.; others Improvement of Quantitative Structure–Activity Relationship (QSAR) Tools for Predicting Ames Mutagenicity: Outcomes of the Ames/QSAR International Challenge Project. *Mutagenesis* **2019**, *34*, 3–16.

(2) Furuhama, A. et al. Evaluation of QSAR Models for Predicting Mutagenicity: Outcome of the Second Ames/QSAR International Challenge Project. *SAR and QSAR in Environmental Research* **2023**, *34*, 983–1001.

(3) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.

(4) European Parliament and Council Regulation (EC) No 1907/2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). Official Journal of the European Union, L 396, 2006.

(5) Ministry of Environment, Republic of Korea Act on Registration, Evaluation, Etc. of Chemicals (K-REACH), Act No. 15511. 2018.

(6) European Chemicals Agency *Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.6: QSARs and grouping of chemicals*; 2008.

(7) International Council for Harmonisation ICH M7(R2): Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk. 2017; `https://www.ich.org/page/multidisciplinary-guidelines`, Step 4 version.

(8) Ames, B. N.; Durston, W. E.; Yamasaki, E.; Lee, F. D. Carcinogens Are Mutagens: A Simple Test System Combining Liver Homogenates for Activation and Bacteria for Detection. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, *70*, 2281–2285.

(9) Mortelmans, K.; Zeiger, E. The Ames Salmonella/Microsome Mutagenicity Assay. *Mutat. Res.* **2000**, *455*, 29–60.

(10) OECD OECD Test Guideline 473: In Vitro Mammalian Chromosomal Aberration Test. 2016.

(11) OECD OECD Test Guideline 474: Mammalian Erythrocyte Micronucleus Test. 2016.

(12) Russell, W. M. S.; Burch, R. L. *The Principles of Humane Experimental Technique*; Methuen: London, 1959.

(13) Hansen, K.; Mika, S.; Schroeter, T.; Sber, A.; Benigni, R.; Bossa, C.; Netzeva, T. I.; Helma, C.; Kramer, S.; Müller, K.-R. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *J. Chem. Inf. Model.* **2009**, *49*, 2077–2081.

(14) Xu, C.; Cheng, F.; Chen, L.; Du, Z.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In Silico Prediction of Chemical Ames Mutagenicity. *J. Chem. Inf. Model.* **2012**, *52*, 2840–2847.

(15) Li, S.; Wan, F.; Shu, H.; Jiang, T.; Zhao, D.; Zeng, J. Machine Learning–Predicting Ames Mutagenicity of Small Molecules. *Toxicol. Rep.* **2021**, *8*, 1570–1577.

(16) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(17) Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783–790.

(18) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476–488.

(19) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability Domain: Towards a More Formal Definition. *SAR QSAR Environ. Res.* **2016**, *27*, 893–909.

(20) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.

(21) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J. Chem. Inf. Model.* **2016**, *56*, 1243–1252.

(22) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; others Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(23) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*, 80.

(24) Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.

(25) Klimisch, H.-J.; Andreae, M.; Tillmann, U. A Systematic Approach for Evaluating the Quality of Experimental Toxicological and Ecotoxicological Data. *Regulatory Toxicology and Pharmacology* **1997**, *25*, 1–5.

(26) Landrum, G.; others RDKit: Open-Source Cheminformatics Software. `https://www.rdkit.org`, Version 2023.03.

(27) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(28) Benigni, R.; Bossa, C. Structure Alerts for Carcinogenicity, and the Salmonella Assay System: A Novel Approach Using the Concept of Complementary Groups of Alerts. *J. Chem. Inf. Model.* **2008**, *48*, 71–79.

(29) Benigni, R.; Bossa, C.; Tcheremenskaia, O. Nongenotoxic Carcinogenicity of Chemicals: Mechanisms of Action and Early Recognition through a New Set of Structural Alerts. *Chem. Rev.* **2013**, *113*, 2940–2957.

(30) Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48*, 312–320.

(31) Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating Mutual Information. *Phys. Rev. E* **2004**, *69*, 066138.

(32) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; pp 785–794.

(33) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems. 2017; pp 3146–3154.

(34) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

(35) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.

(36) Cox, D. R. The Regression Analysis of Binary Sequences. *J. R. Stat. Soc., Ser. B* **1958**, *20*, 215–242.

(37) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533–536.

(38) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. International Conference on Learning Representations (ICLR). 2017.

(39) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; others Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(40) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.

(41) Brier, G. W. Verification of Forecasts Expressed in Terms of Probability. *Mon. Weather Rev.* **1950**, *78*, 1–3.

(42) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 2017; pp 4765–4774.

(43) McNemar, Q. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157.

(44) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26, Reprinted 2004.

(45) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623.

(46) Dobo, K. L.; Greene, N.; Fred, C.; Glowienke, S.; Harvey, J. S.; Hasselgren, C.; Jolly, R.; Kenyon, M. O.; Munzner, J. B.; Muster, W.; others In Silico Methods Combined with Expert Knowledge Rule Out Mutagenic Potential of Pharmaceutical Impurities: An ICH M7 Score. *Regul. Toxicol. Pharmacol.* **2012**, *62*, 482–489.

(47) Tweats, D. J.; Blakey, D.; Heflich, R. H.; Jacobs, A.; Jacobsen, S. D.; Morita, T.; Nishi, T.; Pfuhler, S.; Scott, D.; Treumann, S.; others Report of the IWGT Working Group on Strategies and Interpretation of Regulatory In Vivo Tests I. Increases in Micronucleated Bone Marrow Cells in Rodents That Do Not Indicate Genotoxic Hazards. *Mutat. Res.* **2007**, *627*, 78–91.

(48) Kirkland, D.; Reeve, L.; Gatehouse, D.; Vanparys, P. A Core In Vitro Genotoxicity Battery Comprising the Ames Test Plus the In Vitro Micronucleus Test Is Sufficient to Detect Rodent Carcinogens and In Vivo Genotoxins. *Mutat. Res.* **2011**, *721*, 27–73.

(49) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively Multitask Networks for Drug Discovery. *arXiv preprint arXiv:1502.02072* **2015**,

(50) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-Training Graph Neural Networks. International Conference on Learning Representations (ICLR). 2020.

Supporting Information

The Supporting Information is available free of charge. Summary tables are shown below; complete per-configuration data are provided as supplementary CSV files in the code repository.

Table S0. Endpoint classification criteria

The full mapping of OECD, EPA, and EU Test Guidelines to the three endpoint categories (Ames, in vitro CA, in vivo MN) used for curating all datasets is provided as `table_s0_endpoint_classification.xlsx`. This table was used to assign each raw study record to one of the target endpoints prior to any modeling.

Table S1. Random versus scaffold split comparison

Table 6: Random vs. scaffold split (XGBoost, compact features, raw SMILES, 5×5 CV).

Endpoint	n	Random MCC	Scaffold MCC	Δ
Ames	13,560	0.670 ± 0.010	0.624 ± 0.049	0.046
In vitro CA	1,619	0.180 ± 0.053	0.189 ± 0.084	-0.009
In vivo MN	2,153	0.181 ± 0.094	0.104 ± 0.111	0.077

Table S2. Full 225-experiment results

Complete metrics (MCC, bootstrap CI, ROC-AUC, PR-AUC, balanced accuracy, sensitivity, specificity, Brier score, AD coverage) for all 225 configurations are provided in `table_s2_full_results.csv`

Table S3. Hyperparameter tuning comparison

Default versus tuned MCC for all configurations, including optimized parameter values, are provided in `table_s3_hp_tuning.csv`.

Table S4. Applicability domain coverage

Per-configuration AD coverage (Tanimoto ≥ 0.3, 1024-bit Morgan FP) and mean similarity values are provided in `table_s4_ad_coverage.csv`. Coverage ranged from 48.5% (in vitro) to 58.9% (in vivo) across configurations.

Table S5. Cross-endpoint compound overlap

Pairwise compound overlap and label concordance (κ) between the three endpoint datasets are provided in `table_s5_cross_endpoint.csv`. Overlap with Ames was 97.5% for in vitro and 92.1% for in vivo.

Table S6. SHAP feature importance

SHAP values (mean |SHAP|, TreeExplainer) for all features across all tree-based models are provided in `table_s6_shap.csv`.

Table S7. LDO threshold-tuning analysis (prior shift vs. source shift)

To distinguish prior shift (calibration mismatch from differing prevalence) from genuine source shift, we evaluated each LDO direction at three decision thresholds: default ($t = 0.5$), prior-corrected (t = training positive rate), and oracle (t maximizing MCC on the test set, providing an upper bound). Per-direction and per-algorithm results are shown below.

Note on features: For computational tractability this analysis used Morgan FP-1024 + 13 physicochemical descriptors rather than the full 138-feature compact set used in the main pipeline. Absolute MCC values therefore differ from Table 5 (notably RF industrial→drug: 0.102 here vs. 0.325 in the main analysis). The internal three-threshold comparison within each row is valid because identical features are used across thresholds.

Table 7: LDO threshold-tuning analysis (Morgan FP-1024 + 13 physchem features). Default = MCC at $t = 0.5$. Prior-corrected = MCC at t equal to the training positive rate. Oracle = upper bound MCC across all thresholds. The gap between oracle LDO MCC and the in-domain scaffold-CV MCC ($\approx$0.624) isolates the irreducible source shift that cannot be corrected by threshold tuning.

Direction	Algo.	MCC default	MCC prior-corr.	MCC oracle	Oracle threshold
Drug $\rightarrow$ Ind.	XGBoost	0.231	0.263	0.310	0.788
Drug $\rightarrow$ Ind.	LightGBM	0.241	0.264	0.304	0.781
Drug $\rightarrow$ Ind.	RF	0.250	0.299	0.359	0.690
Ind. $\rightarrow$ Drug	XGBoost	0.200	0.267	0.273	0.041
Ind. $\rightarrow$ Drug	LightGBM	0.239	0.214	0.267	0.180
Ind. $\rightarrow$ Drug	RF	0.102	0.341	0.345	0.037

Mean improvement (default $\rightarrow$ prior-corrected): +0.075 (range -0.024 to +0.239). Mean improvement (default $\rightarrow$ oracle): +0.108 (range +0.028 to +0.243). Oracle MCC range: 0.267–0.359, all substantially below the in-domain scaffold-CV MCC of 0.624.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.zip