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.

A heterogeneous LLM-augmented ensemble for robust drug-induced autoimmunity prediction

Tahsinul Haque Dhrubo

tahsinul.haque.dhrubo@gmail.com

BRAC University

Ayesha Siddika

BRAC University

Muhammad Iqbal Hossain

BRAC University

Research Article

Keywords: drug-induced autoimmunity, ensemble learning, chemical language models, large language models, QLoRA, out-of-distribution generalisation, conformal prediction, interpretability

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10493298/v1

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

Additional Declarations: No competing interests reported.

A heterogeneous LLM-augmented ensemble for robust drug-induced autoimmunity prediction

Tahsinul Haque Dhrubo[1*], Ayesha Siddika[1] and Muhammad Iqbal Hossain[2]

[1*]Bachelor in Computer Science, BRAC University, Dhaka, Dhaka, Bangladesh.
[2]Department of Computer Science and Engineering, BRAC University, Dhaka, Dhaka, Bangladesh.

*Corresponding author(s). E-mail(s): tahsinul.haque.dhrubo@gmail.com;
Contributing authors: ayeshasiddika17.sabr@gmail.com; iqbal.hossain@bracu.ac.bd;

Abstract

Drug-induced autoimmunity (DIA) is a rare, mechanistically heterogeneous adverse drug reaction whose prediction is constrained by small labelled corpora, severe class imbalance and chemically diverse scaffolds. The recently released InterDIA benchmark established the first dedicated machine-learning baseline, in which an EasyEnsembleClassifier trained on a 65-feature genetic-algorithm-selected RDKit descriptor subset reaches an external-test area under the receiver-operating-characteristic curve (AUC) of 0.8930 from a single molecular view, without out-of-distribution (OOD) evaluation or calibrated uncertainty. We propose a heterogeneous six-stream ensemble that fuses classical RDKit descriptors, Morgan ECFP6 fingerprints, two pretrained chemical language models (ChemBERTa-MTR and MoLFormer-XL), a descriptor-transformer hybrid view, and a 7-billion-parameter Qwen-2.5 large language model fine-tuned with QLoRA. A simple mean blend reaches a five-seed external-test AUC of $\mathbf{0.9324 \pm 0.0107}$, exceeding the baseline at every seed. The central finding is robustness rather than the marginal in-distribution gain: on the 68% of held-out drugs whose Bemis-Murcko scaffolds are absent from training, the ensemble degrades by only 0.054 AUC, whereas individual streams degrade by 0.10 to 0.20. We further report the first cross-corpus external evaluation of a DIA model on 427 drug-induced liver injury compounds, split-conformal prediction sets with 91.7% empirical coverage at the 90% target, and transparent SHAP and chain-of-thought

interpretability with explicit hallucination disclosure. Representation-level heterogeneity, not algorithmic novelty, yields a predictor that generalises to novel chemistry and reports calibrated confidence. All artefacts, per-seed predictions and code are released for reproduction.

Keywords: drug-induced autoimmunity, ensemble learning, chemical language models, large language models, QLoRA, out-of-distribution generalisation, conformal prediction, interpretability

1 Introduction

Drug-induced autoimmunity (DIA) is a non-IgE-mediated adverse drug reaction in which exposure to a small-molecule therapeutic provokes an autoimmune phenotype. The principal mechanisms are haptenisation, in which a reactive drug metabolite binds covalently to host proteins to create neo-antigens; direct activation of T cells; molecular mimicry of self-antigens; and oxidative-stress-driven release of damage-associated molecular patterns. Roughly one in ten cases of systemic lupus erythematosus is drug-induced, and recognised offenders span hydralazine, procainamide, isoniazid, minocycline and several anti-tumour-necrosis-factor biologicals [1]. Because the endpoint is rare, is usually detected only after marketing, and arises from chemically unrelated scaffolds, predictive models must cope simultaneously with small labelled datasets, strong class imbalance and broad chemical diversity.

Huang et al. [1] released InterDIA in 2025: a curated 597-drug corpus (477 training and 120 external test) with binary DIA labels and four families of precomputed descriptors, together with a class-imbalance-aware ensemble pipeline. Their strongest configuration, an EasyEnsembleClassifier (EEC) [2] trained on a 65-feature genetic-algorithm-selected RDKit subset (denoted RDKit_GA_65), attains external-test AUC 0.8930, accuracy 0.85, sensitivity 0.7667, specificity 0.8889 and MCC 0.6413. InterDIA is an important first benchmark, yet it leaves four gaps that matter for deployment. First, it relies on a single descriptor view, whereas complementary representations, including circular fingerprints, pretrained chemical language models and general-purpose large language models (LLMs), encode different and potentially synergistic molecular information. Second, it reports single-run point estimates, so it is unclear whether differences between models exceed run-to-run noise. Third, it performs no out-of-distribution (OOD) evaluation, even though a 597-drug corpus guarantees that held-out drugs will contain unseen scaffolds. Fourth, it provides no calibrated uncertainty, which is indispensable for clinical triage.

This study therefore asks whether *representation-level heterogeneity*, that is, fusing classical descriptors, circular fingerprints, pretrained chemical language models and a fine-tuned general-purpose LLM within a single ensemble, can produce a DIA predictor that is simultaneously more accurate, more robust to novel chemistry and better calibrated than the single-view InterDIA baseline on a small, imbalanced and chemically diverse corpus. Three subsidiary questions make that claim testable. The first concerns

discrimination: does a heterogeneous multi-representation ensemble improve external-test discrimination over the single-view baseline, and is any improvement reproducible across random seeds rather than an artefact of a single run? The second concerns robustness: does representation heterogeneity improve out-of-distribution generalisation, both to held-out Bemis-Murcko scaffolds and to an independent external corpus, relative to any single-representation stream? The third concerns uncertainty and interpretability: can the ensemble supply distribution-free calibrated uncertainty and faithful interpretability suitable for clinical triage, and what are the failure modes of LLM-generated reasoning in this setting? Section 4.3 reports the evidence on the first question, Sections 4.5 and 4.6 on the second, and Sections 4.9 and 5.2 on the third.

Answering these questions reframes the prediction problem around robustness and calibrated confidence rather than around a single headline number, and the study makes four contributions. The first is a six-stream heterogeneous ensemble (Figures 1 and 2) that fuses classical descriptors, ECFP6 fingerprints, two pretrained chemical language models with distinct pretraining objectives, a hybrid descriptor and transformer view, and a QLoRA-fine-tuned Qwen-2.5-7B LLM, whose simple mean blend reaches a five-seed external-test AUC of 0.9324 ± 0.0107. The second is a Bemis-Murcko scaffold OOD analysis showing that the ensemble preserves accuracy on novel chemistry 2.5 to 4.5 times better than any single stream, which we argue is the methodologically important result. The third is the first cross-corpus external evaluation of a DIA model, on 427 drug-induced liver injury (DILI) compounds, which reveals that broad-corpus pretraining (MoLFormer-XL) transfers better than descriptor pipelines tuned to the training chemistry. The fourth is a complete uncertainty and interpretability layer comprising bootstrap and DeLong significance testing, TreeExplainer SHAP attributions, chain-of-thought LLM rationales with explicit hallucination disclosure, a Graph Isomorphism Network baseline, split-conformal prediction sets, and an Uncertainty-Triggered LLM Consultation post-processor reported with an honest negative result.

The remainder of the paper is organised as follows. Section 2 reviews prior work on DIA prediction, imbalance-aware ensemble learning, molecular representations, large language models in chemistry and out-of-distribution evaluation, and closes with a comparative analysis of that work against the present study. Section 3 describes the corpus, the seven-step workflow, the ensemble architecture and the evaluation protocol. Section 4 reports the experimental results and Section 5 analyses and discusses them, including interpretability, comparison with the literature and limitations. Section 6 concludes.

2 Literature review

Prior work bearing on the four gaps identified above falls into five strands: computational prediction of DIA itself, imbalance-aware ensemble learning for low-prevalence toxicity, molecular representation learning from hand-crafted descriptors to chemical language models, general-purpose large language models in chemistry, and out-of-distribution evaluation together with uncertainty quantification and interpretability.

We review each strand in turn and then set the representative studies side by side with the present work in Table 1.

DIA belongs to the family of idiosyncratic adverse drug reactions, which are rare, delayed and mechanistically heterogeneous, and are therefore poorly served by the classical dose-response paradigm of toxicology. Risk has historically been flagged through expert structural alerts for reactive-metabolite formation rather than through quantitative models. InterDIA [1] is, to our knowledge, the first dedicated machine-learning benchmark for the endpoint: it curates 597 drugs with binary DIA labels, computes four descriptor families, selects a compact RDKit descriptor subset by genetic algorithm, trains class-imbalance-aware ensembles, and interprets the model with SHAP. It establishes a strong single-view baseline (external-test AUC 0.8930) but evaluates on a single random held-out split, reports single-run point estimates, and provides neither out-of-distribution testing nor calibrated uncertainty. Our work takes InterDIA as both the baseline and the point of departure, retaining its data split and base learner for comparability while addressing each limitation in turn.

Low-prevalence toxicity endpoints are dominated by the negative class, so naive classifiers can achieve high accuracy while missing the rare positives that matter. Ensemble methods that couple resampling with bagging or boosting are the standard remedy. Balanced random forests [3], the EasyEnsembleClassifier [2], and balanced bagging with gradient-boosted base learners such as XGBoost [4] and LightGBM [5] repeatedly undersample the majority class and aggregate diverse learners, and are conveniently available through the imbalanced-learn toolbox [6]. These methods remain highly competitive on tabular molecular descriptors, which is why InterDIA adopts them and why we retain the EasyEnsembleClassifier as a shared base learner across five of our six streams: fixing the learner isolates representational diversity as the variable of interest, so that any gain is attributable to the fusion of molecular views rather than to a change of classifier.

A molecule can be encoded in many complementary ways, and the choice of representation often matters more than the choice of classifier. Physicochemical descriptors computed with RDKit [7] summarise interpretable properties such as surface area, lipophilicity and functional-group counts, and are strong on small datasets. Circular fingerprints, introduced as Morgan connectivity codes [8] and formalised as extended-connectivity fingerprints [9], encode the presence of local substructures and remain a robust baseline for structure-activity modelling. More recently, self-supervised chemical language models learn continuous representations directly from unlabelled SMILES: ChemBERTa and ChemBERTa-2 [10, 11] adapt the transformer encoder [12] to chemistry using masked-token and multi-task-regression pretraining on tens of millions of PubChem molecules, while MoLFormer-XL [13] scales pretraining to large ZINC and PubChem corpora and reports state-of-the-art transfer across property-prediction benchmarks. These learned representations capture information that is largely orthogonal to hand-crafted descriptors, which motivates fusing them rather than choosing between them; InterDIA, by contrast, uses a single descriptor view.

General-purpose LLMs have rapidly become a third representational paradigm for molecules, operating directly on textual SMILES and natural-language prompts. Guo

et al. [14] benchmark LLMs across eight chemistry tasks and find that in-context prompting is competitive on some property-prediction tasks yet still trails specialised models on others, indicating that prompting alone is insufficient. Crucially for data-scarce endpoints, Jablonka et al. [15] show that *fine-tuning* a general-purpose LLM on question-answer formatted molecular property data can match or exceed dedicated machine-learning models, particularly in the low-data regime that characterises DIA. Fine-tuning multi-billion-parameter models is made tractable on consumer hardware by parameter-efficient methods: low-rank adaptation (LoRA) [16] and its quantised variant QLoRA [17], which back-propagate through 4-bit base weights into small low-rank adapters, together with efficient instruction-tuned open models such as Qwen-2.5 [18] and optimised training kernels [19]. Despite this momentum, and to our knowledge, no prior work fine-tunes a multi-billion-parameter LLM specifically for DIA, nor integrates an LLM's chain-of-thought as an explicit ensemble stream. We do both, and, importantly, we report the resulting failure modes, including verdict bias and functional-group hallucination, rather than only the successes.

Random train-test splits systematically overstate performance because structurally similar analogues leak across the split. Scaffold-based evaluation, in which molecules are partitioned by their Bemis-Murcko frameworks [20], is the accepted way to estimate generalisation to novel chemistry and is standard in benchmarks such as MoleculeNet [21] and Chemprop [22]. Those same studies show that expressive end-to-end graph neural networks, including the Graph Isomorphism Network [23], often fail to beat strong descriptor baselines in low-data regimes, a caveat directly relevant to a 477-drug training set. Scaffold splitting nevertheless remains an internal test of a single corpus, and cross-corpus transfer to an independently curated set of drugs, which is the harder and more deployment-relevant condition, is rarely reported for rare toxicity endpoints at all.

Beyond point predictions, trustworthy deployment requires calibrated confidence: conformal prediction [24, 25] provides distribution-free coverage guarantees and has been applied to imbalanced bioactivity classification [26], while feature attribution via SHAP [27] and rigorous AUC comparison via the DeLong test [28] support interpretation and statistical rigour. InterDIA offers SHAP interpretation but neither scaffold-level nor cross-corpus OOD evaluation and no calibrated uncertainty, gaps that our evaluation protocol is designed to close.

Table 1 places the representative studies of each strand beside the present work along the dimensions that matter for deployment. The comparison makes the position of this study explicit. Each prior strand contributes one ingredient in isolation: InterDIA contributes an imbalance-aware learner and SHAP interpretation on a single descriptor view, the chemical language models contribute transferable pretrained representations without imbalance handling or uncertainty, the graph and benchmark literature contributes scaffold-aware evaluation without a rare-endpoint focus, the LLM literature contributes textual reasoning without robustness or calibration analysis, and the conformal literature contributes distribution-free guarantees without representation diversity. No prior study combines them. The design evaluated here unifies four ingredients rarely brought together for a single clinical endpoint: an imbalance-aware base learner, deliberately heterogeneous molecular representations

spanning descriptors, fingerprints, two chemical language models and a fine-tuned LLM, out-of-distribution evaluation at both the scaffold and the corpus level, and distribution-free calibrated uncertainty. That combination is the basis of the answers reported in Sections 4 and 5.

Table 1 Comparative analysis of representative prior work and the present study. Each row is the most representative study or studies of one literature strand; columns are the capabilities required for deployment on a rare, imbalanced and chemically diverse toxicity endpoint. "n/a" marks a dimension that is not applicable to the study design rather than an omission.

Study and endpoint	Molecular view	Imbal. aware	Seeds and stats	OOD evaluation	Uncertainty; interpretability
InterDIA [1]; DIA, 597 drugs	RDKit GA-65, single view	Yes	Single run	None	None; SHAP
ChemBERTa, ChemBERTa-2 [10, 11]; property benchmarks	SMILES transformer	No	Partial	Scaffold split	None; attention
MoLFormer-XL [13]; property benchmarks	SMILES transformer, large scale	No	Yes	Scaffold split	None; attention
MoleculeNet, Chemprop, GIN [21–23]; general benchmarks	Molecular graph, end to end	No	Yes	Scaffold split	Ensemble variance only
LLMs in chemistry [14, 15]; low-data property	Text prompt, fine-tuned LLM	No	Partial	None	None; unaudited rationales
Conformal QSAR [26]; imbalanced bioactivity	Descriptors, fingerprints	Yes	n/a	None	Conformal sets; no attribution
This work; DIA, 597 drugs plus 427 external	**Six views fused**	**Yes**	**Five seeds, CI, DeLong**	**Scaffold and cross-corpus**	**Conformal; SHAP and CoT, disclosed**

3 Methods

3.1 Overview of the approach

The study follows the seven-step workflow shown in Figure 1. The chain on the left is the sequence of operations; the annotations on the right record the artefact each step produces and hands to its successor. Every step is described in full in the sections that follow, and the remainder of this subsection walks through the workflow once end to end so that the design is legible before the details.

Step 1 curates and preprocesses the data. The public InterDIA corpus of 597 drugs is loaded with the authors' own partition into 477 training and 120 external-test molecules, and that partition is held fixed for the entire study so that no selection,

tuning or blending decision ever sees the test drugs. Preprocessing reproduces the original pipeline exactly: standardisation fitted on the training fold only, removal of zero-variance columns, removal of one member of every feature pair correlated above $|\rho| = 0.9$, and restriction to the 65 columns selected by the authors' genetic algorithm. Step 1 therefore outputs a fixed split and a verified descriptor matrix. The details are given in Section 3.2.

Step 2 is a fidelity gate rather than a modelling step, and it exists because every later claim is stated relative to a published baseline. All ten configurations of InterDIA Table 5 are retrained from the released artefacts and compared against the published numbers. Only when the reproduction matches does the study proceed to extend the pipeline; the outcome is reported in Section 4.1.

Step 3 performs multi-view molecular featurisation, and it is the step that carries the central hypothesis. The same SMILES string is encoded six different ways: as the 65 selected RDKit descriptors, as a 384-dimensional ChemBERTa-MTR embedding, as the 449-dimensional concatenation of those two, as a 2048-bit ECFP6 circular fingerprint, as a 768-dimensional MoLFormer-XL embedding, and as a natural-language prompt for a large language model. The views are chosen to be mechanistically complementary rather than merely numerous: hand-crafted descriptors carry interpretable physicochemistry, fingerprints carry local substructure, the two chemical language models carry statistics learned from tens to hundreds of millions of unlabelled molecules under different pretraining objectives, and the LLM carries general chemical knowledge expressed in text. Section 3.3 specifies each encoder.

Step 4 trains one model per view. Five of the six streams use the same EasyEnsembleClassifier configuration as the baseline paper, fitted under five-fold stratified cross-validation so that out-of-fold probabilities are leakage-free, while the sixth stream instead fine-tunes a 7-billion-parameter Qwen-2.5 model with QLoRA and reads its class probability from the first-token YES and NO logits. Fixing the learner across the five feature-based streams is deliberate: it makes representation the only variable that differs, so any complementarity observed later cannot be attributed to a change of classifier. The output of the step is six out-of-fold and six external-test probability vectors per seed.

Step 5 fuses those six probabilities into one score by an unweighted arithmetic mean (Equation 1). Weighted and learned alternatives were implemented and evaluated at this point, and rejected on out-of-fold evidence; the reasoning is given in Section 3.3. The simplicity of the fusion rule matters for the argument of the paper, because it removes any suspicion that the reported gains come from a tuned combiner rather than from the diversity of the views themselves.

Step 6 evaluates the fused predictor along two axes. Performance is measured by repeating steps 3 to 5 independently at five random seeds and reporting mean, standard deviation and confidence intervals, supported by bootstrap resampling and DeLong tests for pairwise AUC differences. Robustness is measured by partitioning the external test set according to whether a drug's Bemis-Murcko scaffold appears in training, and separately by refitting every stream and applying it to 427 independently curated drug-induced liver injury compounds. A Graph Isomorphism Network is trained as an end-to-end graph baseline for context. Section 3.5 gives the protocol.

Step 7 converts the score into a decision that carries its own confidence and explanation. Split-conformal calibration on the out-of-fold scores turns the ensemble probability into a prediction set with a distribution-free coverage guarantee, SHAP attributions on a gradient-boosted surrogate expose which descriptors drive predictions, and chain-of-thought rationales generated by the base LLM provide a textual account that is audited rather than trusted. An uncertainty-triggered consultation step, in which the LLM is re-queried only on molecules whose conformal set is not a singleton, closes the loop from uncertainty back to prediction.

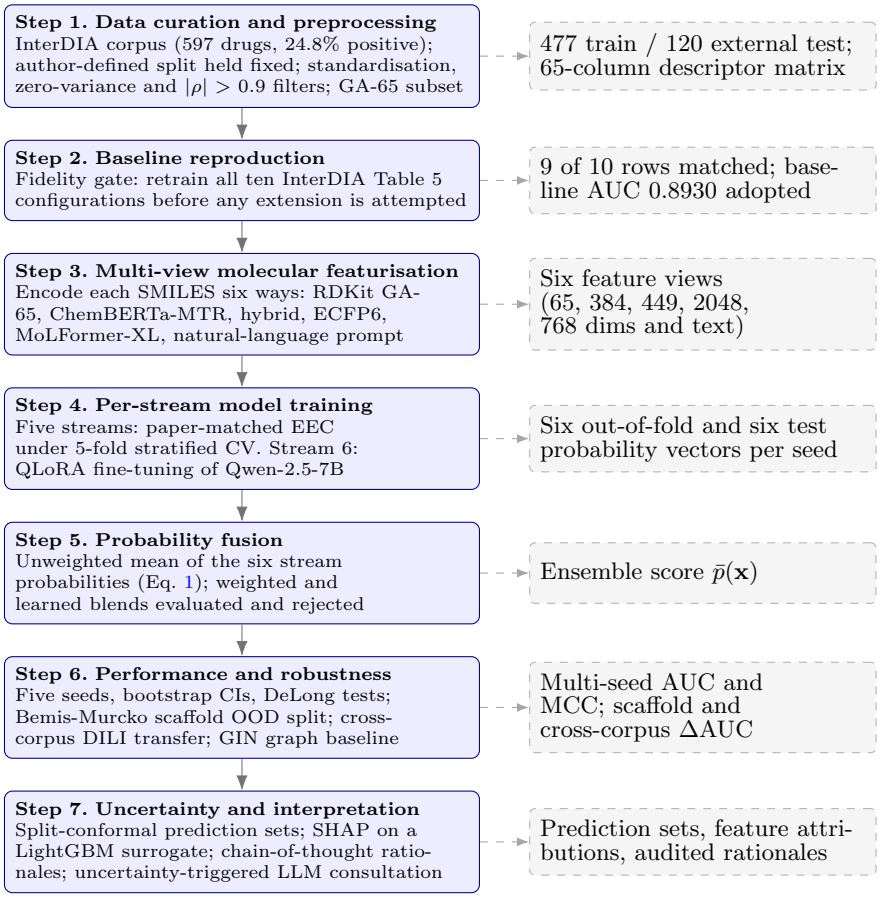


Fig. 1 Flow chart of the overall approach. The left-hand chain is the seven-step workflow executed in order; the dashed annotations on the right name the artefact each step produces and passes to the next. Steps 1 and 2 establish a verified starting point, steps 3 to 5 build the heterogeneous ensemble, and steps 6 and 7 evaluate it and equip it with calibrated uncertainty and interpretation.

3.2 Dataset and preprocessing

We use the public InterDIA corpus [1]: 597 small-molecule drugs with binary DIA labels, partitioned by the authors into 477 training drugs (118 positive, 359 negative; 75.3% negative prevalence) and 120 external-test drugs (30 positive, 90 negative; 75.0% negative prevalence). The 120-drug split is held out unchanged throughout all training and feature selection. For the descriptor stream we adopt the authors' RDKit feature matrix and load the GA-selected 65-column subset directly from their published artefact `Xtrain_std.pkl`, which guarantees exact feature-set identity. The authors' preprocessing pipeline is reproduced verbatim: standardisation fit on training data only, a zero-variance filter, a pairwise Pearson correlation filter at $|\rho| > 0.9$, and restriction to the GA-65 columns. To establish baseline fidelity we first reproduce all ten configurations of InterDIA Table 5 (Section 4.1).

3.3 Ensemble architecture

For an input drug $\mathbf{x}$ (a SMILES string) each stream k emits a probability $p_k(\mathbf{x}) \in [0, 1]$ of the DIA-positive class. The ensemble prediction is the arithmetic mean,

$$\bar{p}(\mathbf{x}) = \frac{1}{6} \sum_{k=1}^{6} p_k(\mathbf{x}). \tag{1}$$

The six streams (Figure 2, Table 2) span deliberately diverse molecular views and learners. Streams 1 to 5 share the paper-matched EEC base learner so that representational diversity, not learner diversity, drives complementarity; stream 6 is a generative LLM.

Table 2 Stream design. Streams 1 to 5 use the paper-matched EEC; stream H is an optional chain-of-thought integration evaluated in ablation only.

k	Representation	Pretrained source	Dim	Learner
1	RDKit_GA_65 descriptors	GA selection [1]	65	EEC
2	ChemBERTa-MTR embedding	ChemBERTa-2-77M-MTR	384	EEC
3	Hybrid (RDKit $\oplus$ MTR)	concatenation	449	EEC
4	ECFP6 fingerprint (radius 3)	RDKit [9]	2048	EEC
5	MoLFormer-XL embedding	MoLFormer-XL-both-10pct	768	EEC
6	Qwen-2.5-7B-Instruct + QLoRA	Qwen2.5-7B-Instruct (4-bit)	n/a	LLM softmax
H	sentence-transformer of CoT	all-MiniLM-L6-v2	384	EEC

Streams 1, 3 and 4 (descriptors, hybrid, fingerprints)

Stream 1 feeds the 65-column RDKit_GA_65 matrix to the EEC. Stream 4 computes Morgan circular fingerprints with radius 3 folded to 2048 bits [8, 9] using RDKit [7]. Stream 3 concatenates the 65-dimensional descriptor vector with the 384-dimensional ChemBERTa-MTR embedding (449 dimensions) to expose explicit cross-view interactions to the learner.

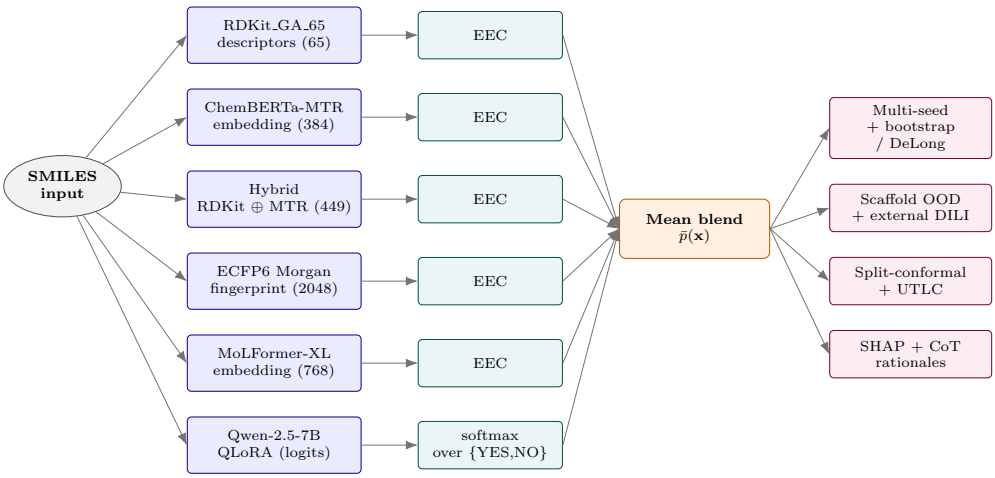


Fig. 2 Overview of the six-stream ensemble. Each stream encodes the input SMILES under a different molecular representation and learner; the five descriptor and embedding streams share the paper-matched EasyEnsembleClassifier (EEC), while the sixth is a QLoRA-fine-tuned Qwen-2.5-7B large language model whose class probability is a softmax over the first-token {YES, NO} logits. The six probabilities are combined by a simple mean and passed to multi-seed, out-of-distribution, conformal-uncertainty and interpretability analyses.

Streams 2 and 5 (chemical language models)

Stream 2 tokenises each SMILES with `DeepChem/ChemBERTa-2-77M-MTR` and mean-pools the final hidden state over the attention mask to a 384-dimensional embedding, standardised on the training fold. Stream 5 uses MoLFormer-XL [13], a chemical language model pretrained on roughly 100 million ZINC and PubChem SMILES, producing 768-dimensional mean-pooled embeddings. MoLFormer-XL is loaded with `trust_remote_code` in an isolated environment pinned to `transformers==4.40.0`, because its custom modelling file imports legacy ONNX interfaces removed in later releases; a defensive stub for the removed module is injected to survive incomplete installations.

Stream 6 (fine-tuned LLM)

We fine-tune `unsloth/Qwen2.5-7B-Instruct-bnb-4bit` [18, 19] with QLoRA [16, 17]. All hyperparameters are reported for reproducibility: 4-bit base weights, LoRA rank $r = 16$, $\alpha = 16$, dropout 0, target modules {q, k, v, o, gate, up, down}_proj; ten epochs at learning rate 5×10^{-5}, batch size 4, 8-bit AdamW, warmup ratio 0.05, bf16 precision, with minority-class oversampling per fold. The model is prompted in a chat template ("`SMILES: <SMILES>. Does this drug cause DIA? Answer YES or NO.`") and the positive-class probability is the softmax over the first-token YES and NO logits,

$$p_{\text{Qwen}}(\mathbf{x}) = \frac{\exp(\ell_{\text{YES}})}{\exp(\ell_{\text{YES}}) + \exp(\ell_{\text{NO}})}. \tag{2}$$

Fine-tuning consumes roughly 8.5 GB of GPU memory and 117 minutes per seed on a single consumer GPU.

Stream H (chain-of-thought integration)

To integrate LLM reasoning into the classifier rather than producing it post hoc, we generate base-Qwen rationales for all 597 molecules using the SHAP top-5 features as evidence and a four-step reasoning template, encode them with the sentence-transformer `all-MiniLM-L6-v2` [29] to 384-dimensional vectors, and train the paper-matched EEC on the embeddings. Stream H is evaluated as a seventh ensemble member in ablation (Section 4.8).

Blending and meta-learning

We compared the simple mean of Equation 1 with AUC-weighted and AUC-squared-weighted means and with a learned logistic-regression-plus-isotonic meta-learner on out-of-fold predictions. The simple mean was retained: weighting amplified single-stream noise and the meta-learner overfit the out-of-fold set.

3.4 Base learner

The EEC base learner reproduces the InterDIA configuration: an outer EasyEnsembleClassifier [2] with ten members, each an AdaBoost of 178 depth-7 decision trees at learning rate 0.92 with the SAMME.R algorithm. Per-seed five-fold stratified cross-validation provides leakage-free out-of-fold predictions; the full-train model produces external-test probabilities.

3.5 Evaluation protocol

Multi-seed reproducibility

All six streams are retrained from scratch at seeds $\{1, 7, 42, 99, 2025\}$, varying the cross-validation split, the EEC internal seed and the LLM training seed. Per-stream out-of-fold and test probabilities are persisted to disk; the blend is computed per seed and aggregated as mean $\pm$ standard deviation with 95% confidence intervals from a t_4 distribution.

Statistical significance

We report bootstrap 95% confidence intervals (1000 percentile resamples) and the DeLong test [28] for pairwise AUC differences on the external test set.

Scaffold out-of-distribution split

Bemis-Murcko scaffolds [20] are computed with RDKit for every drug. A test drug is labelled OOD when its scaffold is absent from training; 82 of 120 test drugs (68.3%) qualify. We report per-stream metrics on the in-distribution and OOD partitions and the paired degradation $\Delta_k = \mathrm{AUC}_k^{\mathrm{OOD}} - \mathrm{AUC}_k^{\text{in-dist}}$.

External-corpus out-of-distribution evaluation

We download the Therapeutics Data Commons DILI task [30, 31], canonicalise all SMILES, and drop 48 drugs overlapping InterDIA to obtain 427 external compounds (48.7% positive). Each stream is refit on full InterDIA training and applied to the DILI features; we report per-stream and ensemble AUC against the in-distribution values.

Interpretability

TreeExplainer SHAP [27] is computed on a LightGBM surrogate [5] trained on RDKit_GA_65 (surrogate test AUC 0.8767, comparable to the paper EEC), and we rank features by mean absolute SHAP. Base Qwen-2.5-7B generates four-step chain-of-thought rationales for all 597 molecules at temperature 0.3.

Baselines and uncertainty

A three-layer Graph Isomorphism Network [23] with 64-dimensional hidden states, global mean-pooling and a two-layer head is trained on five-dimensional atom features as an end-to-end graph baseline. Split-conformal prediction [24, 25] calibrates on the five-seed out-of-fold scores ($n_{\text{cal}} = 477$) with nonconformity $s_i = 1 - \bar{p}(y_i \mid \mathbf{x}_i)$ and threshold $\hat{q}_{1-\alpha}$ at the $\lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil / n_{\text{cal}}$ order statistic. Finally, an Uncertainty-Triggered LLM Consultation (UTLC) step re-queries base Qwen with a forced-binary prompt on molecules whose conformal set at the 90% level is not a singleton.

Implementation

The pipeline runs on a single NVIDIA RTX 5060 Ti (16 GB, Blackwell) under Python 3.12, across four isolated virtual environments to reconcile conflicting dependencies between scikit-learn, transformers 5.x, MoLFormer's legacy transformers 4.40 and the Therapeutics Data Commons loader. Total wall time is approximately six to seven hours, dominated by LLM fine-tuning.

4 Experimental results

This section reports the experimental evidence in the order in which it was produced: reproduction of the published baseline, per-stream and ensemble performance, multi-seed reproducibility and significance testing, the two out-of-distribution evaluations, the graph baseline and the chain-of-thought ablation, and finally conformal uncertainty. Interpretation of these findings, their relation to prior work and the limitations that bound them are deferred to Section 5.

4.1 Baseline reproduction

As a fidelity check we first reproduced all ten configurations of InterDIA Table 5. Nine of ten rows match the published values to four decimal places (Table 3). The only deviation is the EEC row ($\Delta\text{AUC} = -0.0145$), which is explained by documented non-determinism in the imbalanced-learn EasyEnsembleClassifier under parallel execution: the sampler ordering is not fully fixed by the random seed when `n_jobs` is set to use all cores. The deviation is bounded and does not alter the ranking of configurations,

so we adopt the published EEC values (AUC 0.8930, MCC 0.6413) as the comparison baseline.

Table 3 Reproduction of InterDIA Table 5 (top-half, RDKit_GA_65). Nine of ten rows match to four decimals; EEC drifts within documented sampler non-determinism.

Source	Model	AUC	ACC	SEN	SPE	MCC
Paper	BRF	0.8878	0.8083	0.7667	0.8222	0.5444
	EEC	0.8930	0.8500	0.7667	0.8889	0.6413
	BBC+XGBoost	0.8763	0.8083	0.7333	0.8333	0.5313
	BBC+GBDT	0.8974	0.8500	0.7333	0.8889	0.6093
	BBC+LightGBM	0.8915	0.8417	0.7333	0.8778	0.5926
Ours	BRF	0.8878	0.8083	0.7667	0.8222	0.5444
	EEC	0.8785	0.8167	0.8000	0.8222	0.5715
	BBC+XGBoost	0.8763	0.8083	0.7333	0.8333	0.5313
	BBC+GBDT	0.8974	0.8500	0.7333	0.8889	0.6093
	BBC+LightGBM	0.8915	0.8417	0.7333	0.8778	0.5926

4.2 Per-stream and ensemble performance

Table 4 reports the six streams and the ensemble on the external test set at seed 1. Individual streams range from AUC 0.78 (MoLFormer-XL) to 0.88 (EEC-RDKit). No single stream dominates across metrics: the LLM has the highest specificity (0.9667) but the lowest sensitivity (0.4333), reflecting a negative-class bias, whereas the hybrid view has balanced sensitivity. The mean blend reaches AUC 0.9285 at seed 1, above every stream, and improves MCC to 0.6825.

Table 4 Per-stream and ensemble metrics on the external test set (seed 1, $n = 120$). Bold marks the ensemble.

Stream	OOF AUC	Test AUC	ACC	SEN	SPE	MCC
EEC-RDKit_GA_65	0.7956	0.8785	0.8167	0.8000	0.8222	0.5715
ChemBERTa-MTR	0.7865	0.8233	0.7750	0.6667	0.8111	0.4480
Hybrid (RDKit + MTR)	0.7978	0.8615	0.7667	0.8000	0.7556	0.4948
ECFP6	0.7345	0.8167	0.8417	0.7000	0.8889	0.5826
MoLFormer-XL	0.7502	0.7800	0.7333	0.7333	0.7333	0.4156
Qwen-2.5-7B (QLoRA)	0.7122	0.8428	0.8333	0.4333	0.9667	0.5095
Ensemble (mean)	**0.8224**	**0.9285**	**0.8833**	0.7333	**0.9333**	**0.6825**

4.3 Multi-seed reproducibility

Across the five seeds the ensemble attains AUC 0.9324 ± 0.0107 (95% CI [0.919, 0.946]) and MCC 0.6490 ± 0.0513 (Table 5, Figure 3). The per-seed AUCs

$\{0.9285, 0.9219, 0.9333, 0.9500, 0.9281\}$ all exceed the paper baseline of 0.8930, and the confidence interval does not overlap it. The gain over the strongest single stream (EEC-RDKit) is +0.039 AUC, which corresponds to roughly 3.7 standard deviations of the seed-to-seed spread. We are deliberately cautious about this margin: on the 120-drug test set a DeLong test of the ensemble against the EEC baseline gives $p = 0.125$ (Section 4.4). The robust and reproducible claim is therefore consistency across seeds rather than a single significant jump, and the more important evidence is the OOD behaviour reported next.

Table 5 Five-seed external-test metrics (mean ± standard deviation [95% CI from t_4]). The n/a for MoLFormer MCC marks an undefined value at the default threshold on one seed.

Model	AUC	MCC
EEC-RDKit_GA_65	0.8936 ± 0.0104 [0.881, 0.907]	0.6194 ± 0.0414 [0.568, 0.671]
ChemBERTa-MTR	0.8301 ± 0.0137 [0.813, 0.847]	0.4711 ± 0.0276 [0.437, 0.505]
Hybrid	0.8542 ± 0.0181 [0.832, 0.877]	0.5104 ± 0.0214 [0.484, 0.537]
ECFP6	0.8928 ± 0.0169 [0.872, 0.914]	0.5615 ± 0.0268 [0.528, 0.595]
MoLFormer-XL	0.7833 ± 0.0301 [0.745, 0.821]	n/a
Qwen-2.5-7B	0.8361 ± 0.0182 [0.813, 0.859]	0.4727 ± 0.0370 [0.427, 0.519]
Ensemble	**0.9324 ± 0.0107 [0.919, 0.946]**	**0.6490 ± 0.0513 [0.585, 0.713]**
Paper EEC (single run)	0.8930	0.6413

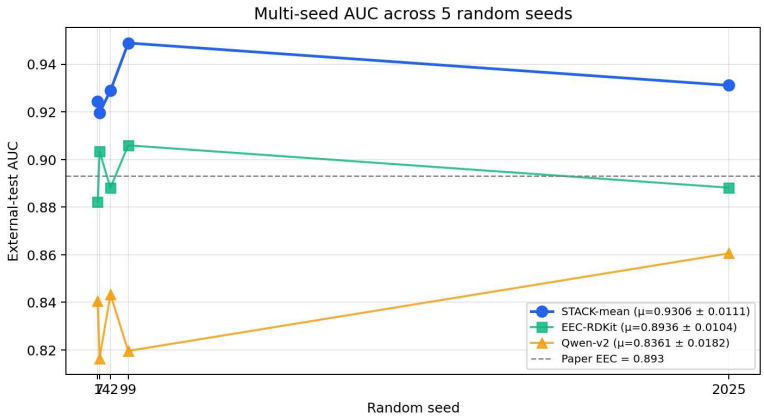


Fig. 3 External-test AUC across the five seeds for each stream and the ensemble. The dashed line marks the InterDIA EEC baseline (0.8930). The ensemble (top trajectory) exceeds the baseline at every seed.

4.4 Statistical significance

Figure 4 shows per-stream bootstrap AUC intervals at seed 1. The ensemble interval [0.865, 0.968] lies above every single stream's point estimate except that of

EEC-RDKit. DeLong tests place the ensemble significantly above five of six streams (ChemBERTa-MTR $p = 0.007$, ECFP6 $p = 0.004$, Qwen $p = 0.028$, hybrid $p = 0.024$, and the dropped ChemBERTa-MLM $p < 0.001$) but only marginally above the EEC baseline (ΔAUC $= +0.044$, $p = 0.125$) on $n = 120$. We report this openly: the in-distribution improvement is consistent but not individually significant on a small test set, which is precisely why a robustness-centred evaluation is warranted.

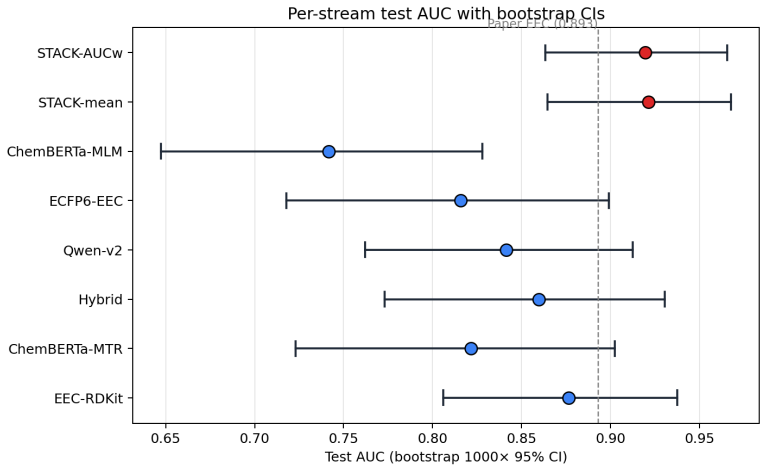


Fig. 4 Per-stream external-test AUC with bootstrap 95% confidence intervals (1000 resamples, seed 1). The ensemble (top) sits above every individual stream's point estimate except EEC-RDKit.

4.5 Scaffold-based out-of-distribution generalisation

This is the central result. On the 82 test drugs (68.3%) with Bemis-Murcko scaffolds absent from training, the ensemble degrades by only 0.054 AUC relative to the in-distribution partition, whereas individual streams degrade by 0.10 to 0.20 (Table 6, Figure 5). The degradation of the ensemble is 2.5 to 4.6 times smaller than that of any single stream, including the LLM, which shows the largest single-stream drop (-0.196). The multi-seed ensemble OOD AUC is 0.9073 ± 0.0074. Within the OOD partition the ensemble is significantly above ChemBERTa-MTR ($p = 0.037$), Qwen ($p = 0.018$) and the dropped MLM stream ($p = 0.002$); against the strongest individual streams the gap is positive but not significant given the reduced OOD sample size, so we treat the paired degradation magnitudes, not the within-partition tests, as the primary evidence.

4.6 Cross-corpus external evaluation on DILI

To probe generalisation beyond scaffold shift we evaluate on 427 DILI compounds (Table 7, Figure 6). Drug-induced autoimmune hepatitis is a recognised DIA subtype, so DILI is a defensible if imperfect proxy. Three observations follow. First, the ensemble AUC falls by 0.25 to 0.6804. Second, MoLFormer-XL alone reaches 0.6919 and thereby

Table 6 Bemis-Murcko scaffold OOD evaluation (seed 1).
$\Delta\text{AUC} = \text{AUC}^{\text{OOD}} - \text{AUC}^{\text{in-dist}}$; ratio is $|\Delta_k|$ relative to the ensemble.

Stream	AUC in-dist ($n = 38$)	AUC OOD ($n = 82$)	ΔAUC	ratio
Qwen-2.5-7B	0.9589	0.7629	−0.196	4.6
Hybrid	0.9714	0.8234	−0.148	3.4
EEC-RDKit	0.9607	0.8274	−0.133	3.1
ECFP6	0.9750	0.8492	−0.126	2.9
ChemBERTa-MTR	0.9000	0.7863	−0.114	2.6
Ensemble	**0.9536**	**0.8992**	−0.054	**1.0**

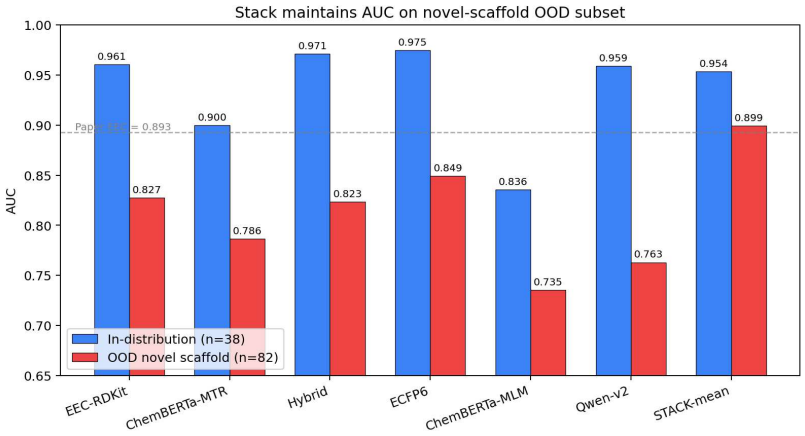


Fig. 5 Per-stream AUC on the in-distribution partition (blue, $n = 38$) and the novel-scaffold OOD partition (red, $n = 82$, 68% of the test set). The ensemble (right) is the only model that retains AUC above 0.89 on novel scaffolds.

outperforms the full ensemble, reversing the in-distribution ranking in which it is the weakest stream; its degradation of only 0.091 is roughly one third of that of the descriptor and fingerprint streams. Third, the LLM's sensitivity collapses from 0.43 in distribution to 0.17 on DILI. To our knowledge this is the first cross-corpus external evaluation of a DIA model; Section 5.1 takes up what the reversal implies for stream selection.

4.7 Graph neural network baseline

A three-layer Graph Isomorphism Network [23] trained end to end on molecular graphs reaches out-of-fold AUC 0.6473 and external-test AUC 0.7515, roughly 18 percentage points below the ensemble and below every one of the six individual streams. The result is reported here as context for the representation choices of Section 3.3 and is discussed in Section 5.1.

Table 7 In-distribution InterDIA versus external-corpus DILI ($n = 427$). In-dist is the five-seed mean; DILI is the seed-1 refit.

Stream	AUC in-dist	AUC DILI	ΔAUC
EEC-RDKit	0.8936	0.6675	−0.226
ChemBERTa-MTR	0.8301	0.6071	−0.223
Hybrid	0.8542	0.6366	−0.218
ECFP6	0.8928	0.6205	−0.272
MoLFormer-XL	0.7833	**0.6919**	−0.091
Qwen-2.5-7B	0.8361	0.6334	−0.203
Ensemble	0.9324	0.6804	−0.252

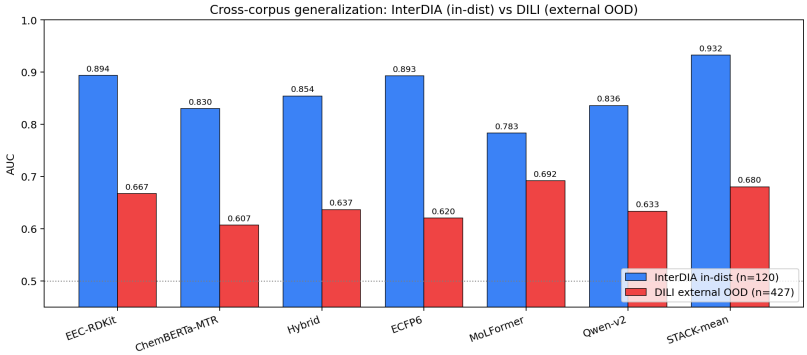


Fig. 6 Cross-corpus generalisation: per-stream AUC on InterDIA in-distribution (blue, $n = 120$, five-seed mean) versus TDC DILI (red, $n = 427$, seed 1). MoLFormer-XL shows the smallest drop and the highest single-stream DILI AUC.

4.8 Chain-of-thought integration

Integrating the LLM's verbal reasoning as an explicit stream (stream H) leaves the headline unchanged (Table 8). Stream H alone reaches AUC 0.6772 ± 0.0203, the weakest of all streams; added as a seventh member it shifts the ensemble AUC by -0.002, within seed noise. The integration is therefore functionally complete but not beneficial on this corpus. The cause is diagnostic: base Qwen commits to a MEDIUM verdict in 77.5% of rationales, so the sentence-transformer embeddings cluster by rhetorical structure rather than by class. A rationale-tuned LLM, trained on explicit step-by-step DIA targets, is the natural remedy.

4.9 Conformal uncertainty and consultation

Split-conformal prediction yields calibrated coverage (Table 9, Figure 7). At the 90% target the empirical coverage is 91.7%, and the confident-singleton subset (106 of 120 drugs, 88.3%) reaches 90.6% accuracy, providing a usable triage layer in which 88% of drugs receive a high-confidence single-label decision. The reliability diagram (Figure 8) shows that the ensemble is moderately calibrated (expected calibration

Table 8 Chain-of-thought integration ablation (five seeds).

Variant	Streams	AUC	MCC
Ensemble (headline)	6	**0.9324 ± 0.0107**	**0.6490 ± 0.0513**
+ CoT stream H	7	0.9302 ± 0.0061	0.6323 ± 0.0402
+ CoT and retained MLM	8	0.9250 ± 0.0079	0.6506 ± 0.0491

error 0.130 versus 0.105 for the EEC reference), with mild over-confidence in the middle probability range where stream disagreement is highest. Tuning the decision threshold to the Youden-optimal value (0.4576) on the out-of-fold predictions recovers the paper-level sensitivity of 0.7667 while raising AUC to 0.9393.

Table 9 Split-conformal prediction calibrated on five-seed out-of-fold scores ($n_{\text{cal}} = 477$).

α	target	empirical coverage	singletons	ACC at singleton
0.05	95%	99.17%	82 (68.3%)	98.78%
0.10	**90%**	**91.67%**	**106 (88.3%)**	**90.57%**
0.20	80%	83.33%	113 (94.2%)	88.50%
0.30	70%	78.33%	102 (85.0%)	92.16%

Finally, the Uncertainty-Triggered LLM Consultation step was applied to the 14 drugs whose conformal set at the 90% level is not a singleton. Re-querying base Qwen with a forced-binary prompt raises accuracy on that subset from 0.500 to 0.643, and over all 120 drugs it moves accuracy from 0.8583 to 0.8750 and MCC from 0.6265 to 0.6516. The LLM returned the negative label for all 14 uncertain drugs, with a mean positive probability near 10^{-7}.

5 Analysis and discussion

The preceding section reported what was measured. This section interprets those measurements: why representation heterogeneity behaves as it does on novel chemistry (Section 5.1), what the model's explanations do and do not support (Section 5.2), how the uncertainty layer should be used in practice (Section 5.3), how the study stands relative to the literature reviewed in Section 2 (Section 5.4), and what bounds the claims (Section 5.5).

5.1 Discrimination, robustness and the role of representation heterogeneity

The in-distribution and out-of-distribution results should be read together, because they carry very different evidential weight. In distribution, the ensemble improves on the baseline at every seed and by +0.039 AUC over the strongest single stream, roughly 3.7 standard deviations of the seed-to-seed spread, yet a DeLong test on

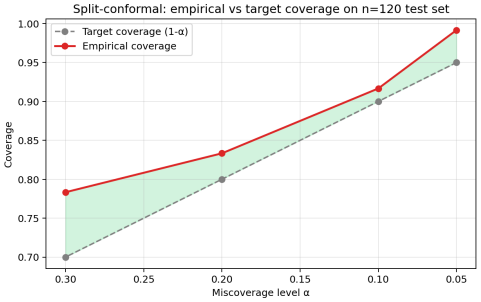


Fig. 7 Empirical versus target coverage of the split-conformal sets across four miscoverage levels; empirical coverage tracks or exceeds the target.

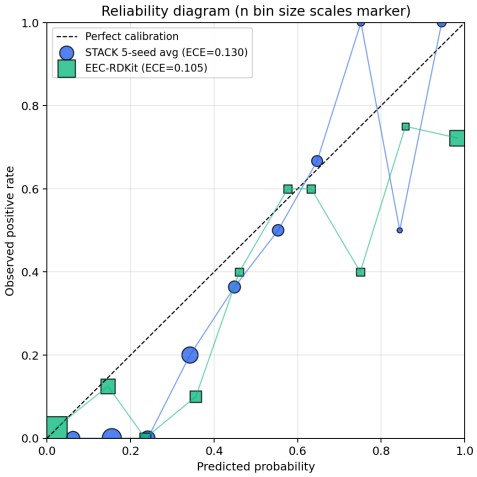


Fig. 8 Ten-bin reliability diagram for the ensemble and the EEC reference. Both are moderately calibrated; the ensemble is mildly over-confident at mid-range probabilities.

$n = 120$ returns $p = 0.125$. Both facts are true, and the honest reading is that the in-distribution gain is consistent but underpowered rather than established. Had the study stopped there it would have added a marginal number to an existing benchmark.

The scaffold analysis is what changes the picture. On the 68.3% of test drugs whose Bemis-Murcko framework never appears in training, the ensemble loses 0.054 AUC while individual streams lose between 0.114 and 0.196, a gap of 2.5 to 4.6 times that is far larger than the in-distribution margin. This is the classical ensemble-diversity effect [21] operating on representations rather than on model seeds. Each view offers a different shortcut when a scaffold is familiar: the descriptor stream can key on a physicochemical range typical of the training actives, the fingerprint stream can key on a specific substructural bit, and the fine-tuned LLM can effectively memorise a drug family it has seen. Those shortcuts are idiosyncratic to each representation, so they fail on different molecules, and averaging cancels the errors that do not agree. The prediction that follows is testable and consistent with the data: the stream with the sharpest in-distribution fit, the fine-tuned LLM at AUC 0.9589 on familiar scaffolds, also degrades the most on unfamiliar ones (-0.196), whereas the ensemble is never the best model on familiar chemistry and is the only model above 0.89 AUC on novel chemistry. Heterogeneity therefore does not buy peak accuracy; it buys the property that matters when a screening model meets a drug unlike anything in its training set.

The cross-corpus result sharpens the same point and complicates it. When the test distribution shifts not only in scaffold but in the labelling concept itself, the ensemble is no longer the best predictor: MoLFormer-XL alone, the weakest stream in distribution at AUC 0.7833, becomes the strongest on DILI at 0.6919 and degrades by only 0.091 against 0.203 to 0.272 for the other streams. The natural explanation is pretraining breadth. MoLFormer-XL was pretrained on the order of 10^8 ZINC and

PubChem molecules, so its representation is not tuned to the 477 training drugs and has correspondingly less to lose; the descriptor, hybrid and fingerprint streams are fitted directly to that narrow chemistry, and the fine-tuned LLM most of all, whose sensitivity falls from 0.43 to 0.17. Uniform averaging spends part of its budget on those brittle streams, which is why an equal-weight blend that is optimal under mild scaffold shift becomes suboptimal under corpus shift. The practical implication is that stream weighting should itself be a function of how far a query molecule sits from the training distribution, which we identify as the most promising extension of this work. It should be said plainly that all DILI AUCs lie between 0.61 and 0.69, so this comparison ranks degrees of degradation and does not claim a deployable cross-endpoint predictor.

The Graph Isomorphism Network baseline is best understood in the same frame. At external-test AUC 0.7515 it is not a competitive stream, which is unsurprising when an end-to-end message-passing network must learn both a representation and a classifier from 477 labelled molecules; this is exactly the low-data failure mode documented for graph models against strong descriptor baselines [21, 22]. Its role in this study is to pre-empt the reasonable objection that a graph network would have been the better modelling choice, and to support the underlying design principle that on corpora of this size the productive move is to import representations pretrained elsewhere rather than to learn one from scratch.

5.2 Interpretability

SHAP attributions on the LightGBM surrogate (Figure 9) are dominated by the van der Waals surface-area family (SMR_VSA, PEOE_VSA, SlogP_VSA and EState_VSA descriptors), which encode polar surface area and electronic character relevant to protein binding and reactive-metabolite formation. The primary-amine count `fr_NH2` ranks among the top features, consistent with the recognised primary-amine DIA risk class (hydralazine, procainamide, isoniazid); amide-bond and bicyclic-system counts also rank highly, matching haptenisation and aromatic-system mechanisms. These attributions agree with the chemistry highlighted by InterDIA, which supports the surrogate as a faithful explanatory proxy. The agreement is worth stating carefully: it establishes that the descriptor stream keys on mechanistically plausible features, not that the ensemble as a whole does, since the embedding and LLM streams are not covered by this surrogate.

Base-Qwen chain-of-thought rationales were generated for all 120 test molecules and give a complementary, and more fragile, form of explanation. The model identifies real functional groups in most cases (amines, aromatics, amides, ketones and fluorinated rings) and invokes appropriate mechanisms. We disclose two pathologies transparently: a commitment bias toward MEDIUM verdicts, in 77.5% of rationales, and occasional functional-group hallucination, with at least one rationale referencing a nitroimidazole ring absent from the input. The two failures have different consequences. The verdict bias is what makes the chain-of-thought stream uninformative as a classifier, since embeddings of near-identical hedged text cannot separate the classes, and it explains the null ablation result of Section 4.8 more convincingly than any tuning artefact would. The hallucination is the more serious for deployment, because a fluent rationale referencing a substructure the molecule does not contain is precisely

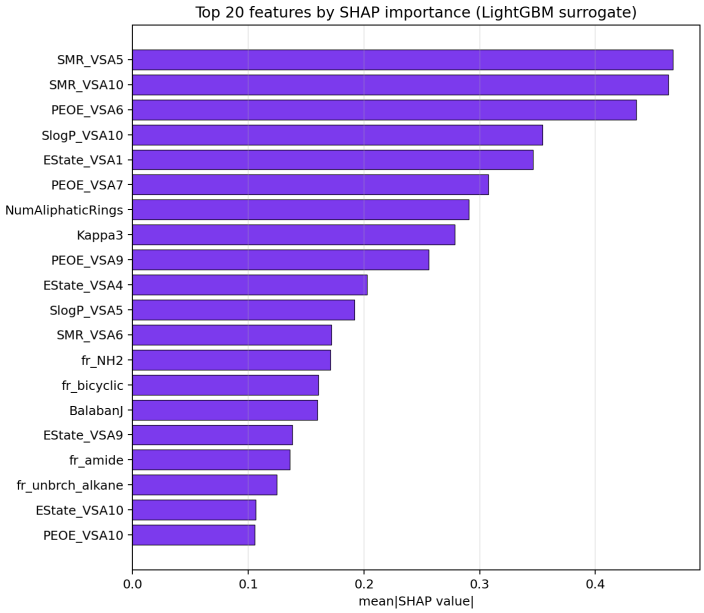


Fig. 9 Top 20 features by mean absolute SHAP value on the LightGBM surrogate (test AUC 0.8767). Van der Waals surface-area descriptors and the primary-amine count dominate.

the kind of error a non-specialist reader cannot catch. We therefore treat the rationales as auditable text rather than as evidence, and we regard a chemist-blinded rating study and rationale-tuned fine-tuning as prerequisites for using them in any decision path.

5.3 Uncertainty and its use in practice

The conformal layer is the component closest to being deployable. Empirical coverage of 91.7% against a 90% target confirms the distribution-free guarantee holds on this corpus, and the operationally useful statistic is that 88.3% of drugs receive a singleton prediction set at 90.6% accuracy. That supports a concrete triage policy: automate the singletons and route the remaining 11.7% to expert review, which concentrates human effort on the molecules where the streams disagree. The reliability diagram tempers this, since the ensemble is mildly over-confident at mid-range probabilities (expected calibration error 0.130 against 0.105 for the EEC reference), exactly the region where stream disagreement is highest; averaging sharpens the decision without fully correcting the confidence attached to it.

The consultation experiment is an instructive negative result and we report it as such. Accuracy on the 14 uncertain drugs rises from 0.500 to 0.643, which looks like successful LLM assistance until the mechanism is examined: the model answered negative for all 14. Since the test set is 75% negative and the uncertain subset sits near the decision boundary, a constant negative response is expected to gain roughly this much. The apparent improvement is population-prior correction rather than chemistry

reasoning, and the correct conclusion is that consulting an instruction-tuned LLM at the point of uncertainty adds nothing on this endpoint. This is consistent with the verdict bias documented above and points to the same remedy, a model fine-tuned on reasoning traces rather than on labels alone. We keep the result in the paper because the positive-looking headline number would be easy to report without the diagnosis, and because the diagnosis is the useful part.

5.4 Comparison with prior work

Table 10 summarises the head-to-head comparison with InterDIA, and extends the qualitative positioning of Table 1 with the quantitative outcome. Beyond the headline metrics, the ensemble adds capabilities absent from the baseline: multi-seed confidence intervals, scaffold OOD robustness, cross-corpus evaluation and conformal uncertainty. Relative to the broader literature, the results confirm that class-imbalance-aware ensembles remain strong for low-prevalence toxicity [2], that broad-corpus chemical language models transfer better across chemistry than descriptor pipelines [13], and that end-to-end graph models underperform in low-data regimes [21, 22]. They also qualify the optimism of the LLM-for-chemistry literature [14, 15]: a QLoRA-fine-tuned 7-billion-parameter model is a genuinely useful ensemble member, at AUC 0.8361 across seeds and the highest specificity of any stream, but it is the least robust member under scaffold shift and its free-text reasoning does not survive audit. The contribution of this work is therefore not any single one of these ingredients but the demonstration, on a real clinical endpoint, that representation-level heterogeneity converts a marginal in-distribution gain into a substantial and reproducible robustness gain.

Table 10 Comparison with the InterDIA baseline [1].

Capability	InterDIA EEC	This work (ensemble)
External-test AUC	0.8930 (single run)	0.9324 ± 0.0107 (five seeds)
MCC	0.6413	0.6490 ± 0.0513
Multi-seed CIs	not reported	95% CI [0.919, 0.946]
Scaffold OOD	not reported	ΔAUC -0.054 on 68% novel scaffolds
External-corpus OOD	not reported	AUC 0.6804 on 427 DILI drugs
Calibrated uncertainty	none	91.7% conformal coverage at 90% target
Interpretability	SHAP	SHAP + LLM rationales (disclosed)

5.5 Limitations

Several limitations bound the claims. The external test set is small ($n = 120$), so in-distribution pairwise comparisons are underpowered and we rely on multi-seed consistency rather than single-test significance. No clean public DIA-labelled external corpus exists; DILI is a proxy whose label set includes non-immune hepatotoxicity, which inflates the measured domain-shift penalty. The LLM rationales hallucinate occasionally and are not yet validated by expert chemists. The fine-tuned LLM is a

binary classifier while rationales come from the base model, so a unified rationale-tuned model remains open. Finally, the UTLC consultation is only weakly validated and behaves as a prior corrector on this corpus.

6 Conclusion

We presented a heterogeneous, LLM-augmented ensemble for drug-induced autoimmunity prediction that fuses classical descriptors, circular fingerprints, two pretrained chemical language models, a hybrid view and a QLoRA-fine-tuned 7-billion-parameter LLM. The ensemble reaches a five-seed external-test AUC of 0.9324 ± 0.0107, exceeding the InterDIA baseline at every seed, but its more important property is robustness: on the 68% of test drugs with novel Bemis-Murcko scaffolds it loses only 0.054 AUC, 2.5 to 4.5 times less than any single stream. We additionally reported the first cross-corpus external evaluation of a DIA model, where broad-corpus pretraining transferred best; calibrated split-conformal prediction sets suitable for triage; transparent SHAP and chain-of-thought interpretability; and two honest negative results, in chain-of-thought integration and LLM consultation, that point to rationale-tuned models as the next step. The work shows that for small, imbalanced and chemically diverse toxicity endpoints, representation-level heterogeneity is a practical route to models that generalise and quantify their own confidence. All data, per-seed predictions and code are released to support reproduction and extension.

Acknowledgements. We thank Huang et al. for releasing the InterDIA dataset and code, and the Unsloth team for the QLoRA optimisations that made 7-billion-parameter fine-tuning feasible on a single consumer GPU.

Declarations

Funding: No specific funding was received for this work.

Competing interests: The authors declare no competing interests.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Data availability: The InterDIA dataset is publicly available from Huang et al. [1]. The DILI corpus is available through the Therapeutics Data Commons [30]. All per-stream and per-seed prediction arrays, conformal sets, SHAP values and LLM rationales accompany this work in the project repository.

Materials availability: Not applicable.

Code availability. All code accompanies this work in the project repository with command-by-command reproduction instructions. Project Repository Link: Exile404/Heterogeneous-llm-ensemble-dia

Author contributions: T.H.D. designed the study, implemented the pipeline, ran all experiments and drafted the manuscript. A.S. contributed to validation and writing. M.I.H. supervised the work and reviewed and edited the manuscript. All authors read and approved the final version.

References

[1] Huang, X., Wang, Q., Li, R., *et al.*: InterDIA: interpretable prediction of drug-induced autoimmunity through ensemble machine learning approaches. Toxicology **511**, 154064 (2025) https://doi.org/10.1016/j.tox.2025.154064

[2] Liu, X.-Y., Wu, J., Zhou, Z.-H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **39**(2), 539–550 (2009)

[3] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)

[4] Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[5] 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. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30 (2017)

[6] Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research **18**(17), 1–5 (2017)

[7] Landrum, G., et al.: RDKit: open-source cheminformatics. Version 2024.03 (2024). https://www.rdkit.org

[8] Morgan, H.L.: The generation of a unique machine description for chemical structures. Journal of Chemical Documentation **5**(2), 107–113 (1965)

[9] Rogers, D., Hahn, M.: Extended-connectivity fingerprints. Journal of Chemical Information and Modeling **50**(5), 742–754 (2010)

[10] Chithrananda, S., Grand, G., Ramsundar, B.: ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885 (2020)

[11] Ahmad, W., Simon, E., Chithrananda, S., Grand, G., Ramsundar, B.: ChemBERTa-2: towards chemical foundation models. arXiv preprint arXiv:2209.01712 (2022)

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30 (2017)

[13] Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., Das, P.: Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence **4**(12), 1256–1264 (2022)

[14] Guo, T., Guo, K., Nan, B., Liang, Z., Guo, Z., Chawla, N.V., Wiest, O., Zhang, X.: What can large language models do in chemistry? A comprehensive benchmark on eight tasks. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2023)

[15] Jablonka, K.M., Schwaller, P., Ortega-Guerrero, A., Smit, B.: Leveraging large language models for predictive chemistry. Nature Machine Intelligence **6**(2), 161–169 (2024) https://doi.org/10.1038/s42256-023-00788-1

[16] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)

[17] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: efficient finetuning of quantized LLMs. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 36 (2023)

[18] Yang, A., Yang, B., Zhang, B., et al.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)

[19] Han, D., Han, M., Unsloth team: Unsloth: 2x faster LLM fine-tuning (2024). https://github.com/unslothai/unsloth

[20] Bemis, G.W., Murcko, M.A.: The properties of known drugs. 1. molecular frameworks. Journal of Medicinal Chemistry **39**(15), 2887–2893 (1996)

[21] 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. Chemical Science **9**(2), 513–530 (2018)

[22] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., *et al.*: Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling **59**(8), 3370–3388 (2019)

[23] Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: International Conference on Learning Representations (ICLR) (2019)

[24] Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)

[25] Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)

[26] Norinder, U., Boyer, S.: Binary classification of imbalanced datasets using conformal prediction. Journal of Chemical Information and Modeling **57**(7), 1591–1597 (2017)

[27] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30 (2017)

[28] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics **44**(3), 837–845 (1988)

[29] Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using siamese BERT-networks. Proceedings of EMNLP-IJCNLP, 3982–3992 (2019)

[30] Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C.W., Xiao, C., Sun, J., Zitnik, M.: Therapeutics Data Commons: machine learning datasets and tasks for drug discovery and development. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2021)

[31] Chen, M., Suzuki, A., Thakkar, S., Yu, K., Hu, C., Tong, W.: DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. Drug Discovery Today **21**(4), 648–653 (2016)