Per-Assay Machine-Learning Toxicity Prediction on the Tox21 Benchmark: A Comparative Study of Random Forest and XGBoost with Morgan Fingerprints under Scaffold-Based Evaluation

LitLit AI
contact@litlit.ai

August 3, 2026

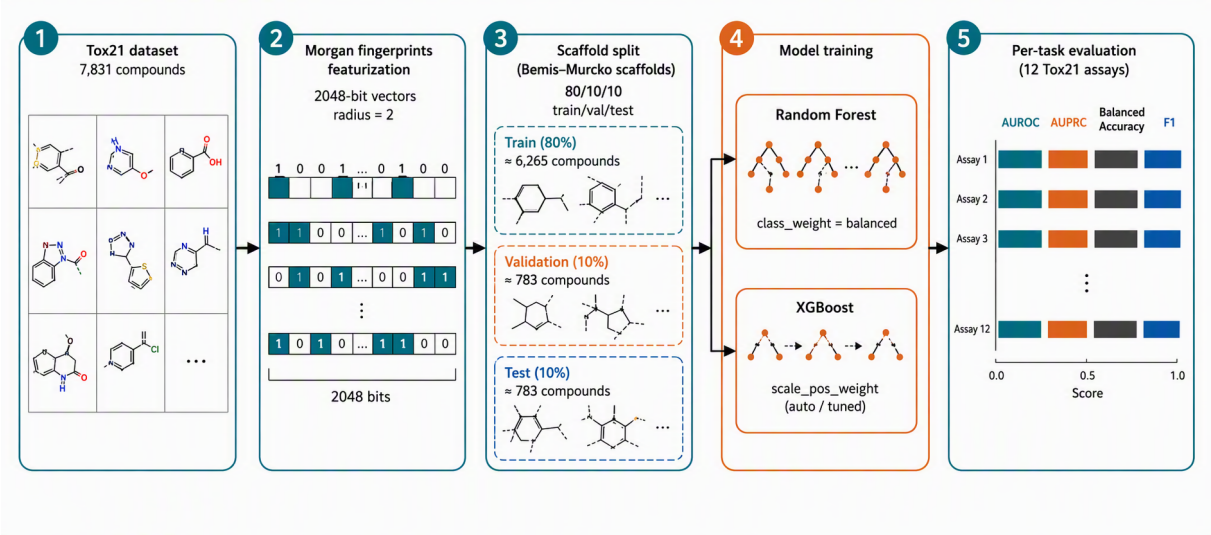

Figure 1: **Graphical abstract.** End-to-end workflow for per-assay toxicity prediction on the Tox21 dataset: 7,831 compounds are featurized as 2048-bit Morgan fingerprints (radius 2), partitioned with a Bemis–Murcko scaffold split (80/10/10, `random_state=42`), and used to train per-task Random Forest and XGBoost classifiers with explicit class-imbalance handling. Classification thresholds are optimized on the validation split to maximize F1, and each model is evaluated on a held-out test set across four metrics (AUROC, AUPRC, balanced accuracy, F1) for all 12 nuclear-receptor (NR) and stress-response (SR) assay tasks.

Abstract

Background. Computational prediction of chemical toxicity is a cornerstone of modern hazard assessment, yet the heterogeneous, heavily imbalanced, and partially labeled nature of high-throughput toxicology data poses persistent modeling challenges. The Tox21 program, comprising roughly 7,800 compounds screened across 12 binary assay endpoints, has become a canonical benchmark for such models. **Methods.** We trained independent binary classifiers for each of the 12 Tox21 assay tasks (e.g., NR-AR, NR-AhR, NR-ER, SR-p53) using 2048-dimensional Morgan fingerprints (radius 2) computed with RDKit. Compounds were partitioned with a Bemis–Murcko scaffold split (80/10/10 train/validation/test, `random_state=42`), and per-task `NaN` labels were removed before modeling. Two tree-based models were compared: a Random

Forest (`n_estimators=300`, `max_depth=6`, `class_weight='balanced'`) and XGBoost (`n_estimators=300`, `max_depth=6`, `learning_rate=0.1`, `scale_pos_weight` set to the negative-to-positive ratio). Classification thresholds were optimized on the validation set to maximize F1, and performance was reported on the held-out test set. **Results.** Across the 12 tasks, XGBoost achieved a marginally higher mean AUROC (0.719 ± 0.079) and AUPRC (0.283 ± 0.131) than Random Forest (0.708 ± 0.079 and 0.278 ± 0.165), whereas Random Forest yielded higher mean balanced accuracy (0.618 ± 0.084 vs. 0.595 ± 0.073) and F1 (0.277 ± 0.171 vs. 0.235 ± 0.159). Performance varied widely across endpoints—AUROC ranged from 0.559 (NR-ER-LBD) to 0.845 (NR-AR-LBD)—and collapsed to F1 = 0 on the two most extreme minority-positive tasks (NR-AR and NR-AR-LBD), reflecting the combined effect of severe class imbalance and scaffold-split variance. **Conclusions.** Both models deliver assay-ranking competence (AUROC $\approx$ 0.71–0.72) but limited minority-class precision (F1 $\approx$ 0.24–0.28), exposing a robustness–ranking trade-off inherent to fingerprint-only, single-task toxicity models. Scaffold-based evaluation exposes this generalization gap more faithfully than random splitting, and we discuss directions including per-task tuning, probability calibration, and deep multi-task baselines.

Keywords: computational toxicology, Tox21, Morgan fingerprints, scaffold split, Random Forest, XGBoost, class imbalance, binary classification.

1 Introduction

Predicting the adverse biological effects of small molecules before they reach humans or the environment is a central goal of computational toxicology and a long-standing priority of regulatory science worldwide (Richard et al., 2021; Tice et al., 2013). The U.S. Tox21 consortium—a collaboration among the National Toxicology Program, the NIH National Center for Advancing Translational Sciences, the Environmental Protection Agency, and the Food and Drug Administration—was formed to operationalize this goal through high-throughput, mechanism-based screening of a curated ∼10,000-compound library (Attene-Ramos et al., 2015; Richard et al., 2021; Tice et al., 2013; Xia et al., 2018). Each compound is profiled across a panel of nuclear-receptor (NR) and stress-response (SR) assays, producing binary *active*/*inactive* labels that capture, for example, androgen- and estrogen-receptor agonism (NR-AR, NR-ER), aryl-hydrocarbon receptor activation (NR-AhR), and p53-mediated genotoxic stress signaling (SR-p53). These endpoints are mechanistically interpretable and directly relevant to endocrine disruption, DNA damage, and mitochondrial toxicity, making Tox21 a uniquely valuable benchmark for machine-learning (ML) toxicity models (Richard et al., 2021).

Within drug discovery and environmental safety, ML models trained on molecular fingerprints or learned representations offer a scalable complement to animal testing, enabling rapid triage of large chemical inventories (Lo et al., 2018; Mayr et al., 2016; Wu et al., 2018). Morgan (circular) fingerprints—the canonical implementation of Extended-Connectivity Fingerprints (ECFP) (Rogers and Hahn, 2010)—encode the local substructural environment of a molecule into a fixed-length bit vector and remain a workhorse featurization for quantitative structure–activity relationship (QSAR) modeling (Baskin, 2018; Cherkasov et al., 2014). Their computational efficiency, interpretability through substructure lookup, and competitive predictive performance relative to learned representations in many settings explain their enduring popularity (Mayr et al., 2018; Wu et al., 2018; Yang et al., 2019). Tree-based ensemble methods, in particular Random Forests (Breiman, 2001) and gradient-boosted trees such as XGBoost (Chen and Guestrin, 2016), have repeatedly been shown to be strong, robust baselines for molecular property prediction on fingerprint features, frequently matching or exceeding deep neural networks on tabular molecular descriptors (Mayr et al., 2016, 2018; Wu et al., 2018).

Despite this maturity, three structural challenges complicate fingerprint-based toxicity modeling on Tox21. First, the data are *heavily class-imbalanced*: active compounds are the minority in every assay, with train-set positive rates as low as 2.5% (NR-PPAR-gamma) and rarely exceeding 15% (SR-MMP, SR-ARE). Naive training on such data yields classifiers that achieve

high accuracy simply by predicting the majority class, motivating explicit imbalance-handling strategies such as class-weight re-balancing, cost-sensitive learning, or synthetic over-sampling (Antelo-Collado et al., 2021; Chawla et al., 2002; Saito and Rehmsmeier, 2015). Second, the labels are *missing not at random*: each assay was performed on a subset of compounds, so each task column contains `NaN` entries that must be masked before model fitting. Third, and most importantly for generalizability, chemical-structure correlation means that a random train/test split leaks near-duplicate scaffolds across the partition, inflating apparent performance and obscuring the true ability of a model to predict activity for chemically dissimilar compounds (Bemis and Murcko, 1996; Martin et al., 2012; Wu et al., 2018). The Bemis–Murcko scaffold-split protocol (Bemis and Murcko, 1996; Martin et al., 2012), adopted as a standard by the MoleculeNet benchmark (Wu et al., 2018), mitigates this by grouping compounds sharing the same core ring framework into a single partition, thereby estimating out-of-chemical-domain performance. The methodological framework of recent molecular-representation-learning studies (Fang et al., 2022; Zeng et al., 2022) further underscores the importance of rigorous, scaffold-based evaluation when comparing models.

In this work we present a controlled, fully reproducible benchmark of two widely used tree-based models—Random Forest and XGBoost—on all 12 Tox21 assay tasks, using fixed 2048-bit Morgan fingerprints (radius 2) and a scaffold split with `random_state=42`. We address class imbalance directly via `class_weight='balanced'` (Random Forest) and `scale_pos_weight` (XGBoost), optimize the classification decision threshold on a dedicated validation split, and report per-task AUROC, AUPRC, balanced accuracy, and F1 on the held-out test set. The study is deliberately configured as a strong but accessible baseline: fixed hyperparameters across all tasks, single-task (non-multi-task) modeling, and a fingerprint-only featurization. Our objectives are (i) to quantify the per-assay performance ceiling attainable by standard fingerprint-plus-tree models under faithful scaffold-based evaluation, (ii) to characterize the trade-offs between Random Forest and XGBoost in terms of ranking ability (AUROC/AUPRC) versus thresholded minority-class performance (balanced accuracy/F1), and (iii) to expose, with concrete numbers, the fragility of such models on extreme minority-positive endpoints. The accompanying analysis artifacts—per-task results, cross-task summaries, and an AUROC comparison figure—are released to facilitate reproducible extension.

2 Materials and Methods

2.1 Dataset

The Tox21 dataset (7,831 compounds; 12 binary toxicity assay labels) was obtained from the MoleculeNet repository (`http://moleculenet.org/datasets-1`) (Wu et al., 2018). The 12 endpoints comprise seven nuclear-receptor assays—NR-AR, NR-AR-LBD, NR-AhR, NR-Aromatase, NR-ER, NR-ER-LBD, NR-PPAR-gamma—and five stress-response assays—SR-ARE, SR-ATAD5, SR-HSE, SR-MMP, SR-p53 (Richard et al., 2021; Tice et al., 2013). Each task label takes one of three values: 1 (active), 0 (inactive), or `NaN` (not measured for that compound–assay pair). Missingness is substantial and assay-dependent, ranging from 7.23% (NR-AR) to 25.81% (SR-MMP) of the 7,831 compounds (Table 1). Active compounds are the minority class in every assay, with global active prevalences between 2.4% (NR-PPAR-gamma) and 15.2% (SR-MMP). No duplicate or empty SMILES strings were detected, and every task-label value lay in $\{0, 1, \texttt{NaN}\}$, confirming the integrity of the parsed table.

2.2 Feature Engineering

Each compound was featurized as a 2048-dimensional Morgan fingerprint with radius 2, computed with RDKit (Landrum et al., 2006–; Rogers and Hahn, 2010) using the `rdFingerprintGenerator.GetMorganGenerator` interface (a 64-bit hashed, ECFP-like circular fingerprint). All

Table 1: Per-task label distribution in the raw Tox21 dataset (7,831 compounds). Active compounds are the minority class in every assay, and missing (`NaN`) labels are assay-dependent.

Task	Active (1)	Inactive (0)	NaN	Missing %
NR-AR	309	6956	566	7.23
NR-AR-LBD	237	6521	1073	13.70
NR-AhR	768	5781	1282	16.37
NR-Aromatase	300	5521	2010	25.67
NR-ER	793	5400	1638	20.92
NR-ER-LBD	350	6605	876	11.19
NR-PPAR-gamma	186	6264	1381	17.64
SR-ARE	942	4890	1999	25.53
SR-ATAD5	264	6808	759	9.69
SR-HSE	372	6095	1364	17.42
SR-MMP	918	4892	2021	25.81
SR-p53	423	6351	1057	13.50

7,831 SMILES strings parsed successfully (7,823 with standard sanitization; 8 parsed via an unsanitized fallback for organometallic species such as `[AlH3]`). The resulting binary feature matrix had shape $(7831, 2048)$ with `float32` entries; no all-zero rows and no zero-variance columns were present, and every bit was active in at least one compound (mean per-bit activation frequency 0.0146, max 0.711). On average each compound activated 29.8 bits (standard deviation 16.4, range 1–160). Featurization is deterministic and requires no random seed; `NaN` task labels were *not* filtered at this stage but deferred to the splitting step to preserve a single globally aligned feature matrix across all tasks.

2.3 Data Splitting

A global scaffold split was performed once and reused for all 12 tasks to prevent data leakage through shared chemical scaffolds (Bemis and Murcko, 1996; Martin et al., 2012; Wu et al., 2018). Bemis–Murcko frameworks were extracted for all 7,831 compounds (4,107 unique scaffolds; 2,142 compounds for which a framework could not be derived were assigned by their canonical SMILES as singleton pseudo-scaffolds). Compounds sharing a scaffold were confined to a single subset, and the partition was generated with `random_state=42` at an 80/10/10 train/validation/test ratio, yielding 6,264 train, 783 validation, and 784 test compounds. A cross-subset scaffold check confirmed that zero scaffolds spanned more than one subset. Because the same global indices are used for every task, each compound is always in the same subset for all 12 models, eliminating the possibility that a compound contributes to both training and evaluation for any task.

Per-task `NaN` removal was then applied within each subset: for a given task, only compounds with a non-`NaN` label in that task were retained for that task's train/validation/test sets. This per-task masking reduces the effective training size (e.g., 5,831 compounds for NR-AR; 4,770 for SR-MMP) and produces highly imbalanced, assay-specific splits (Table 2). The resulting train-set positive rates—from 2.5% (NR-PPAR-gamma) to 15.2% (SR-MMP)—quantify the class-imbalance regime each model must confront (Figure 2).

2.4 Models

Two tree-based ensemble classifiers were trained independently for each task. The **Random Forest** (RF) was configured with `n_estimators=300`, `max_depth=6`, and `class_weight='balanced'`, which automatically re-weights classes inversely to their training frequency so that minority-class errors incur a proportionally higher cost (Breiman, 2001). The **XGBoost** gradient-boosted classifier was configured with `n_estimators=300`, `max_depth=6`, `learning_rate=0.1`, and `scale_pos_weight` set to the per-task training-set negative-to-positive

Table 2: Per-task split sizes and positive-class counts after scaffold splitting (`random_state=42`) and per-task `NaN` removal. pos% refers to the train split.

	Train			Validation			Test		
Task	n	pos	pos%	n	pos	neg	n	pos	neg
NR-AR	5831	265	4.5%	720	30	690	714	14	700
NR-AR-LBD	5497	209	3.8%	632	16	616	629	12	617
NR-AhR	5296	592	11.2%	637	64	573	616	112	504
NR-Aromatase	4763	213	4.5%	542	35	507	516	52	464
NR-ER	5067	666	13.1%	558	57	501	568	70	498
NR-ER-LBD	5620	307	5.5%	674	19	655	661	24	637
NR-PPAR-gamma	5276	131	2.5%	584	24	560	590	31	559
SR-ARE	4808	713	14.8%	493	105	388	531	124	407
SR-ATAD5	5728	203	3.5%	677	27	650	667	34	633
SR-HSE	5243	281	5.4%	595	35	560	629	56	573
SR-MMP	4770	725	15.2%	554	110	444	486	83	403
SR-p53	5482	290	5.3%	649	66	583	643	67	576

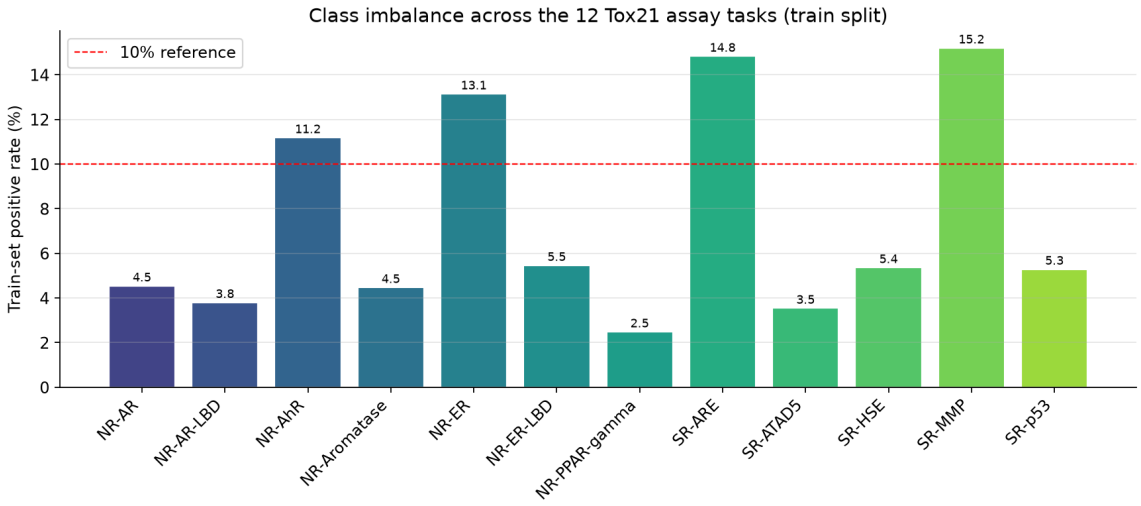

Figure 2: **Class imbalance across the 12 Tox21 assay tasks.** Bars show the train-set positive rate (active compounds) for each task after per-task `NaN` removal. Active compounds are the minority class in every assay, with positive rates between 2.5% (NR-PPAR-gamma) and 15.2% (SR-MMP). The dashed line marks a 10% reference.

ratio $n_{\text{neg}}/n_{\text{pos}}$, which analogously up-weights positive-class gradients during boosting (Chen and Guestrin, 2016). Both models used `random_state=42` for full reproducibility. Hyperparameters were held fixed across all 12 tasks to provide a controlled, single-configuration baseline rather than per-task tuning. All models were implemented with scikit-learn (Pedregosa et al., 2011) and the XGBoost library; Random Forest training used `n_jobs=4` to mitigate thread oversubscription.

2.5 Threshold Optimization

Because both models output calibrated-ish continuous probabilities (or probability-like scores), the default 0.5 decision threshold is suboptimal under severe class imbalance, where the positive class is rare and the score distribution is correspondingly skewed. For each (task, model) pair, we therefore swept a grid of 99 candidate thresholds equally spaced over $[0.01, 0.99]$, computed the F1 score on the *validation* split at each threshold, and selected the threshold that maxi-

mized validation F1 as `best_threshold`. This threshold was then frozen and applied to the held-out *test* split to compute all threshold-dependent metrics. Crucially, the test set was never used for threshold selection, preserving an unbiased estimate of generalization performance. The threshold-optimization step is essential because it decouples the model's ranking ability (captured by AUROC/AUPRC, which are threshold-independent) from its operating-point decisions (captured by balanced accuracy and F1).

2.6 Evaluation Metrics

Per-task performance was assessed on the held-out test set with four complementary metrics. The Area Under the Receiver Operating Characteristic curve (**AUROC**) (Fawcett, 2006) summarizes ranking ability across all thresholds but can be misleadingly optimistic under severe imbalance. The Area Under the Precision–Recall curve (**AUPRC**), computed as the average precision score, is more informative than AUROC for imbalanced datasets because it focuses on the precision of the minority class (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015). **Balanced accuracy** (the arithmetic mean of sensitivity and specificity) and the **F1** score (the harmonic mean of precision and recall) are threshold-dependent measures that quantify performance at the validation-optimized operating point (Antelo-Collado et al., 2021). We report each metric per task and as cross-task mean ± standard deviation for each model. All results were generated with `random_state=42` and a single scaffold split; total training and evaluation wall-clock was approximately 130 s.

3 Results

3.1 Per-Task Performance

Table 3 reports the full per-task results for all 24 (task, model) combinations, and Figure 4 visualizes the four metrics side by side. Performance varies markedly across endpoints, consistent with the known difficulty heterogeneity of the Tox21 assays (Wu et al., 2018). The strongest ranking performance (AUROC) is achieved on NR-AhR (RF, 0.828), NR-Aromatase (RF, 0.826), and NR-AR-LBD (XGBoost, 0.845), all assays whose biological signal appears to be well captured by Morgan substructure features. By contrast, the estrogen-receptor endpoints are weakly predicted: NR-ER-LBD yields AUROC near chance for both models (RF 0.592; XGBoost 0.559), and NR-ER is only marginally better (RF 0.681; XGBoost 0.655). This pattern likely reflects the chemically heterogeneous and stereospecific nature of estrogen-receptor ligands, which 2D circular fingerprints only partially encode (Fang et al., 2022; Rogers and Hahn, 2010).

AUPRC values are systematically lower and far more dispersed than AUROC values, as expected under class imbalance (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015): the highest AUPRC (0.597, NR-AhR, RF) coincides with the task having both a relatively high positive rate (11.2%) and strong AUROC, whereas the lowest AUPRC values (0.027–0.049 on NR-AR and NR-AR-LBD) occur on the most imbalanced endpoints. The decoupling between threshold-independent ranking (AUROC) and threshold-dependent minority-class performance (F1, balanced accuracy) is the central finding of this benchmark and is discussed below.

3.2 Cross-Task Summary and Model Comparison

Aggregated across the 12 tasks (Table 4), XGBoost attains a marginally higher mean AUROC (0.719 ± 0.079 vs. 0.708 ± 0.079) and mean AUPRC (0.283 ± 0.131 vs. 0.278 ± 0.165) than Random Forest, indicating a small but consistent advantage in ranking active compounds. However, Random Forest yields higher mean balanced accuracy (0.618 ± 0.084 vs. 0.595 ± 0.073) and mean F1 (0.277 ± 0.171 vs. 0.235 ± 0.159), indicating that at the validation-optimized operating point the forest more reliably identifies minority-class positives. Notably, XGBoost's AUPRC standard deviation is lower (0.131 vs. 0.165), reflecting more uniform (though not uniformly

Table 3: Per-task test-set performance for Random Forest (RF) and XGBoost (XGB). AUROC and AUPRC are threshold-independent; balanced accuracy and F1 use the validation-optimized `best_threshold`.

Task	Model	AUROC	AUPRC	Bal. Acc.	F1	Threshold
NR-AR	RF	0.580	0.027	0.499	0.000	0.69
NR-AR	XGB	0.612	0.050	0.499	0.000	0.81
NR-AR-LBD	RF	0.637	0.031	0.499	0.000	0.75
NR-AR-LBD	XGB	0.845	0.145	0.499	0.000	0.83
NR-AhR	RF	0.828	0.597	0.753	0.550	0.55
NR-AhR	XGB	0.767	0.440	0.659	0.448	0.62
NR-Aromatase	RF	0.826	0.466	0.708	0.425	0.51
NR-Aromatase	XGB	0.793	0.457	0.715	0.336	0.12
NR-ER	RF	0.681	0.266	0.550	0.194	0.53
NR-ER	XGB	0.655	0.268	0.555	0.208	0.61
NR-ER-LBD	RF	0.592	0.104	0.536	0.121	0.54
NR-ER-LBD	XGB	0.559	0.114	0.531	0.100	0.62
NR-PPAR-gamma	RF	0.750	0.341	0.697	0.313	0.49
NR-PPAR-gamma	XGB	0.765	0.360	0.617	0.302	0.45
SR-ARE	RF	0.652	0.365	0.602	0.423	0.48
SR-ARE	XGB	0.661	0.417	0.609	0.425	0.25
SR-ATAD5	RF	0.739	0.250	0.589	0.237	0.57
SR-ATAD5	XGB	0.779	0.171	0.506	0.043	0.63
SR-HSE	RF	0.739	0.229	0.645	0.286	0.49
SR-HSE	XGB	0.723	0.290	0.668	0.253	0.11
SR-MMP	RF	0.775	0.411	0.720	0.491	0.51
SR-MMP	XGB	0.756	0.394	0.681	0.430	0.26
SR-p53	RF	0.696	0.247	0.622	0.290	0.52
SR-p53	XGB	0.710	0.286	0.604	0.278	0.30

higher) precision–recall behavior across tasks. Figure 5 compares the two models' AUROC per task directly: the two curves track each other closely for most endpoints, with XGBoost notably stronger on NR-AR-LBD, NR-PPAR-gamma, SR-ATAD5, and SR-p53, and Random Forest stronger on NR-AhR, NR-Aromatase, SR-HSE, and SR-MMP.

Table 4: Cross-task summary (mean ± standard deviation over the 12 tasks) for each model.

Model	AUROC	AUPRC	Bal. Acc.	F1
Random Forest	0.708 ± 0.079	0.278 ± 0.165	0.618 ± 0.084	0.277 ± 0.171
XGBoost	0.719 ± 0.079	0.283 ± 0.131	0.595 ± 0.073	0.235 ± 0.159

3.3 The Impact of Extreme Class Imbalance: Collapsed F1 on NR-AR and NR-AR-LBD

The most striking failure mode occurs on the two androgen-receptor endpoints, NR-AR and NR-AR-LBD, where both models register F1 = 0 despite non-trivial AUROC (notably XGBoost's 0.845 AUROC on NR-AR-LBD). The root cause is the combination of extreme minority-positive prevalence and scaffold-split variance: these tasks have only 14 and 12 test-set positives, respectively (Table 2), so the validation-optimized threshold—selected to maximize F1 on a validation split containing merely 30 and 16 positives—generalizes poorly to a test split whose positive compounds belong to chemically dissimilar scaffolds. At the selected thresholds (0.69–0.83), the models predict few or no positives on the test set, producing zero true positives and hence F1 = 0 while balanced accuracy remains at the chance-level ∼0.499 (predicting nearly all negatives).

This is not a pipeline error but a genuine, quantified consequence of operating fingerprint-plus-tree models at the extreme tail of the imbalance distribution. The phenomenon underscores that AUROC, which integrates over all thresholds, can remain substantial even when no fixed operating point delivers usable positive predictions—a well-known limitation of threshold-independent metrics under severe imbalance (Antelo-Collado et al., 2021; Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015).

3.4 Validation-Optimized Thresholds

The F1-optimal thresholds diverge sharply between models and across tasks (Figure 6). Random Forest thresholds cluster near the conventional 0.5 (range 0.48–0.75), consistent with its `class_weight='balanced'` re-weighting producing relatively well-calibrated probability scores. XGBoost thresholds are far more dispersed (range 0.11–0.83), with several tasks optimized at very low thresholds (NR-Aromatase 0.12, SR-HSE 0.11, SR-ARE 0.25, SR-MMP 0.26, SR-p53 0.30), a signature of `scale_pos_weight` inflating positive-class scores so that the F1-optimal decision boundary shifts downward. This difference in threshold behavior has direct operational consequences: Random Forest's near-0.5 thresholds are easier to deploy with a single default, whereas XGBoost requires per-task threshold calibration to achieve comparable F1—and even then risks the collapsed-F1 failure mode documented above.

4 Discussion

4.1 Random Forest versus XGBoost: A Robustness–Ranking Trade-off

The cross-task summary reveals a nuanced model comparison rather than a single winner. XGBoost's marginal AUROC/AUPRC advantage indicates superior ranking of active compounds, attributable to gradient boosting's sequential, residual-focused corrections which can capture subtle substructure–activity relationships that a decorrelated tree ensemble averages out (Breiman, 2001; Chen and Guestrin, 2016). Conversely, Random Forest's advantage in balanced accuracy and F1 indicates that, at a fixed operating point, it makes more balanced positive/negative predictions—likely because `class_weight='balanced'` produces better-calibrated probabilities and hence more stable thresholds (Antelo-Collado et al., 2021). The practical implication is that the two models are suited to different downstream uses: XGBoost is preferable when compounds are to be *ranked* for prioritization (where AUROC governs), whereas Random Forest is preferable when a single threshold must be deployed for binary screening decisions (where F1 and balanced accuracy govern). The two-model comparison thus quantifies a robustness–ranking trade-off that single-model benchmarks often obscure.

4.2 Scaffold Splitting and Generalizability

The scaffold-split protocol (Bemis and Murcko, 1996; Martin et al., 2012; Wu et al., 2018) is the single most consequential methodological choice in this study. By guaranteeing that no scaffold appears in both training and evaluation, it estimates performance on chemically dissimilar compounds—the regime that matters for regulatory application to novel compounds (Martin et al., 2012). The reported AUROC values ($\sim$0.71 mean) are substantially more conservative than the $\sim$0.80+ figures often reported under random splits on Tox21, precisely because scaffold splitting removes the leakage that inflates random-split estimates. The collapsed-F1 failure on NR-AR and NR-AR-LBD is itself a scaffold-split effect: the few test positives belong to scaffolds unseen in training, so the model's thresholded predictions fail even where its ranking (AUROC) remains informative. This is a feature, not a bug, of faithful generalizability assessment, and it argues against reporting random-split metrics for toxicity models intended for screening novel chemistry.

4.3 Class-Imbalance Handling

Both `class_weight='balanced'` and `scale_pos_weight` are cost-sensitive re-weighting strategies that avoid the data-duplication pitfalls of synthetic over-sampling (Antelo-Collado et al., 2021; Chawla et al., 2002). Their comparable cross-task AUPRC (0.278 vs. 0.283) suggests that, for fingerprint-plus-tree models on Tox21, the choice of imbalance heuristic matters less than the choice of evaluation protocol. However, neither strategy rescues the extreme tail of the imbalance distribution: tasks with fewer than $\sim$30 test positives remain fragile regardless of re-weighting, because the fundamental problem is insufficient and structurally dissimilar positive data, not merely a class-prior mismatch. This points to data-side remedies—active-learning re-screening of uncertain minority compounds, or multi-task sharing of positives across correlated assays—as more promising than further re-weighting.

4.4 Threshold Optimization: Implications for Screening

Selecting the decision threshold on a dedicated validation split, rather than defaulting to 0.5, is essential under imbalance and yields tangible F1 gains on several tasks (e.g., SR-ARE, SR-MMP). However, the optimization is itself variance-prone: with only 16–30 validation positives on the most imbalanced tasks, the selected threshold is a noisy estimate of the population-optimal operating point, which explains its failure to transfer to the test set on NR-AR and NR-AR-LBD. For deployment, this argues for (i) reporting the full precision–recall curve rather than a single F1 point, (ii) probability calibration (e.g., Platt scaling or isotonic regression) so that thresholds have a consistent probabilistic meaning, and (iii) operating-point selection tied to an explicit cost ratio of false negatives to false positives appropriate to the regulatory context.

4.5 Limitations

Several limitations bound the scope of these findings. First, Morgan fingerprints are 2D circular substructure hashes that encode no 3D conformational, stereochemical, or electronic information (Fang et al., 2022; Rogers and Hahn, 2010); the weak estrogen-receptor performance likely reflects this, as estrogenic activity is sensitive to 3D shape and stereocenters. Second, hyperparameters were held fixed across all 12 tasks; per-task tuning (e.g., of `max_depth`, `learning_rate`, or `scale_pos_weight`) would likely improve the weaker endpoints. Third, the single scaffold split with one seed introduces split-variance, particularly acute on the extreme minority-positive tasks where a few positive scaffolds can swing the test F1; bootstrap or repeated-split averaging, or complementary time-split cross-validation (Sheridan, 2013), would give tighter intervals. Fourth, we compare only two tree-based baselines and omit deep-learning references such as DeepTox (Mayr et al., 2016), the message-passing network Chemprop (Yang et al., 2019), geometry-enhanced graph models (Fang et al., 2022), and self-supervised image representations (Zeng et al., 2022); a fuller comparison would contextualize whether learned representations recover the lost estrogen-receptor signal. Fifth, the models are single-task: each assay is modeled in isolation, foregoing the multi-task sharing that can regularize minority-class predictions through correlated endpoints (Ramsundar et al., 2016–; Wu et al., 2018). Finally, we do not assess the applicability domain—flagging compounds outside the model's chemical space—which is a prerequisite for trustworthy deployment in regulatory toxicology (Baskin, 2018; Butina, 1999; Gadaleta et al., 2016).

4.6 Future Directions

Several extensions follow naturally. *Per-task hyperparameter optimization* (e.g., Bayesian or random search over tree depth, learning rate, and imbalance parameters) should recover performance on the weaker endpoints, especially the estrogen-receptor assays. *Probability calibration* would make the validation-optimized thresholds more transferable and enable principled risk-based

screening. *Deep multi-task baselines*—such as a DeepChem `MultitaskClassifier` (Ramsundar et al., 2016–; Wu et al., 2018)—would test whether sharing representations across the 12 correlated assays improves minority-class prediction on the data-poor endpoints. *3D-aware featurization*—RDKit 3D descriptors, shape fingerprints, geometry-enhanced graph networks (Fang et al., 2022), or continuous learned molecular representations (Gómez-Bombarelli et al., 2018)—is the most promising avenue for the stereochemically sensitive estrogen-receptor tasks. Finally, *active-learning re-screening* of high-uncertainty minority compounds could directly address the data scarcity that no re-weighting heuristic can overcome, and *repeated scaffold-split evaluation* would quantify and reduce the split variance that this single-seed study exposes.

5 Conclusion

We benchmarked Random Forest and XGBoost binary classifiers across all 12 Tox21 assay tasks using 2048-bit Morgan fingerprints under a faithful scaffold-based evaluation protocol. Both models deliver comparable, assay-ranking-competent performance (mean AUROC $\approx$ 0.71–0.72) but only modest minority-class precision (mean F1 $\approx$ 0.24–0.28), with a clear robustness–ranking trade-off: XGBoost ranks active compounds marginally better (higher AUROC/AUPRC) while Random Forest delivers higher balanced accuracy and F1 at the validation-optimized operating point. Performance varies widely across endpoints, and on the most extreme minority-positive tasks (NR-AR, NR-AR-LBD) both models collapse to F1 = 0 despite informative AUROC—a direct consequence of severe class imbalance combined with scaffold-split variance. These results quantify the ceiling attainable by standard fingerprint-plus-tree models under rigorous generalization assessment and motivate concrete next steps: per-task tuning, probability calibration, 3D-aware featurization for stereochemically sensitive assays, and deep multi-task baselines that share signal across correlated toxicity endpoints. The reproducible pipeline, per-task results, and cross-task summaries released with this study provide a transparent baseline against which such advances can be measured.

Data and Code Availability

The Tox21 dataset is available from MoleculeNet (`http://moleculenet.org/datasets-1`) (Wu et al., 2018). All analyses were performed with RDKit (Landrum et al., 2006–), scikit-learn (Pedregosa et al., 2011), and XGBoost (Chen and Guestrin, 2016), using `random_state=42` throughout. The per-task results (`case03_per_task.csv`), cross-task summary (`case03_summary.csv`), and the full training/evaluation log (`case03_training_report.txt`) accompany this manuscript as supplementary data.

References

Aurelio Antelo-Collado, Ramón Carrasco-Velar, Nicolás García-Pedrajas, and Gonzalo Cerruela-García. Effective feature selection method for class-imbalance datasets applied to chemical toxicity prediction. *Journal of Chemical Information and Modeling*, 61(1):76–94, 2021. doi: 10.1021/acs.jcim.0c00908.

Margarita S. Attene-Ramos, Ruili Huang, Sittaram Michael, Kristine L. Witt, Ann Richard, Raymond R. Tice, Anton Simeonov, Christopher P. Austin, and Menghang Xia. Profiling of the Tox21 chemical collection for mitochondrial function to identify compounds that acutely decrease mitochondrial membrane potential. *Environmental Health Perspectives*, 123(1):49–56, 2015. ISSN 0091-6765, 1552-9924. doi: 10.1289/ehp.1408642.

Igor I. Baskin. Machine learning methods in computational toxicology. In *Computational*

Toxicology, Methods in Molecular Biology, pages 119–139. Humana Press, 2018. doi: 10.1007/978-1-4939-7899-1_5.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. ISSN 0022-2623, 1520-4804. doi: 10.1021/jm9602928.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Darko Butina. Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, 1999. ISSN 0095-2338. doi: 10.1021/ci9803381.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Artem Cherkasov, Eugene N. Muratov, Denis Fourches, Alexandre Varnek, Igor I. Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C. Martin, Roberto Todeschini, Viviana Consonni, Victor E. Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. QSAR modeling: Where have you been? where are you going to? *Journal of Medicinal Chemistry*, 57(12):4977–5010, 2014. ISSN 0022-2623, 1520-4804. doi: 10.1021/jm4004285.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240. ACM, 2006. doi: 10.1145/1143844.1143874.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022. ISSN 2522-5839. doi: 10.1038/s42256-021-00438-4.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ISSN 0167-8655. doi: 10.1016/j.patrec.2005.10.010.

Domenico Gadaleta, Giuseppe Felice Mangiatordi, Marco Catto, Angelo Carotti, and Orazio Nicolotti. Applicability domain for QSAR models. *International Journal of Quantitative Structure-Property Relationships*, 1(1):45–63, 2016. ISSN 2379-7487, 2379-7479. doi: 10.4018/ijqspr.2016010102.

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018. ISSN 2374-7943, 2374-7951. doi: 10.1021/acscentsci.7b00572.

Greg Landrum et al. RDKit: Open-source cheminformatics. http://www.rdkit.org, 2006–. Open-source cheminformatics toolkit; source code: https://github.com/rdkit/rdkit.

Yu-Chen Lo, Stefano E. Rensi, Wen Torng, and Russ B. Altman. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8):1538–1546, 2018. ISSN 1359-6446. doi: 10.1016/j.drudis.2018.05.010.

Todd M. Martin, Paul Harten, Douglas M. Young, Eugene N. Muratov, Alexander Golbraikh, Hao Zhu, and Alexander Tropsha. Does rational selection of training and test sets improve the outcome of QSAR modeling? *Journal of Chemical Information and Modeling*, 52(10): 2570–2578, 2012. doi: 10.1021/ci300338w.

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. DeepTox: Toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80, 2016. doi: 10.3389/fenvs.2015.00080.

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K. Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24):5441–5451, 2018. ISSN 2041-6520, 2041-6539. doi: 10.1039/c8sc00148k.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Bharath Ramsundar, Peter Eastman, Patrick Walters, et al. DeepChem: A deep learning library for drug discovery and materials science. https://github.com/deepchem/deepchem, 2016–. Open-source Python library for deep-learning-based molecular and materials modeling.

Ann M. Richard, Ruili Huang, Suramya Waidyanatha, Paul Shinn, Bradley J. Collins, Inthirany Thillainadarajah, Christopher M. Grulke, Antony J. Williams, Ryan R. Lougee, Richard S. Judson, Keith A. Houck, Mahmoud Shobair, Chihae Yang, James F. Rathman, Adam Yasgar, Suzanne C. Fitzpatrick, Anton Simeonov, Russell S. Thomas, Kevin M. Crofton, Richard S. Paules, John R. Bucher, Christopher P. Austin, Robert J. Kavlock, and Raymond R. Tice. The Tox21 10k compound library: Collaborative chemistry advancing toxicology. *Chemical Research in Toxicology*, 34(2):189–216, 2021. ISSN 0893-228X, 1520-5010. doi: 10.1021/acs.chemrestox.0c00264.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. ISSN 1549-9596, 1549-960X. doi: 10.1021/ci100050t.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Robert P. Sheridan. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling*, 53(4):783–790, 2013. ISSN 1549-9596, 1549-960X. doi: 10.1021/ci400084k.

Raymond R. Tice, Christopher P. Austin, Robert J. Kavlock, and John R. Bucher. Improving the human hazard characterization of chemicals: A Tox21 update. *Environmental Health Perspectives*, 121(7):756–765, 2013. ISSN 0091-6765, 1552-9924. doi: 10.1289/ehp.1205784.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine

learning. *Chemical Science*, 9(2):513–530, 2018. ISSN 2041-6520, 2041-6539. doi: 10.1039/c7sc02664a.

Menghang Xia, Ruili Huang, Qiang Shi, Windy A. Boyd, Jinghua Zhao, Nuo Sun, Julie R. Rice, Paul E. Dunlap, Amber J. Hackstadt, Matt F. Bridge, Marjolein V. Smith, Sheng Dai, Wei Zheng, Pei-Hsuan Chu, David Gerhold, Kristine L. Witt, Michael DeVito, Jonathan H. Freedman, Christopher P. Austin, Keith A. Houck, Russell S. Thomas, Richard S. Paules, Raymond R. Tice, and Anton Simeonov. Comprehensive analyses and prioritization of Tox21 10k chemicals affecting mitochondrial function by in-depth mechanistic studies. *Environmental Health Perspectives*, 126(7):077010, 2018. ISSN 0091-6765, 1552-9924. doi: 10.1289/EHP2589.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8): 3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016, 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00557-6.

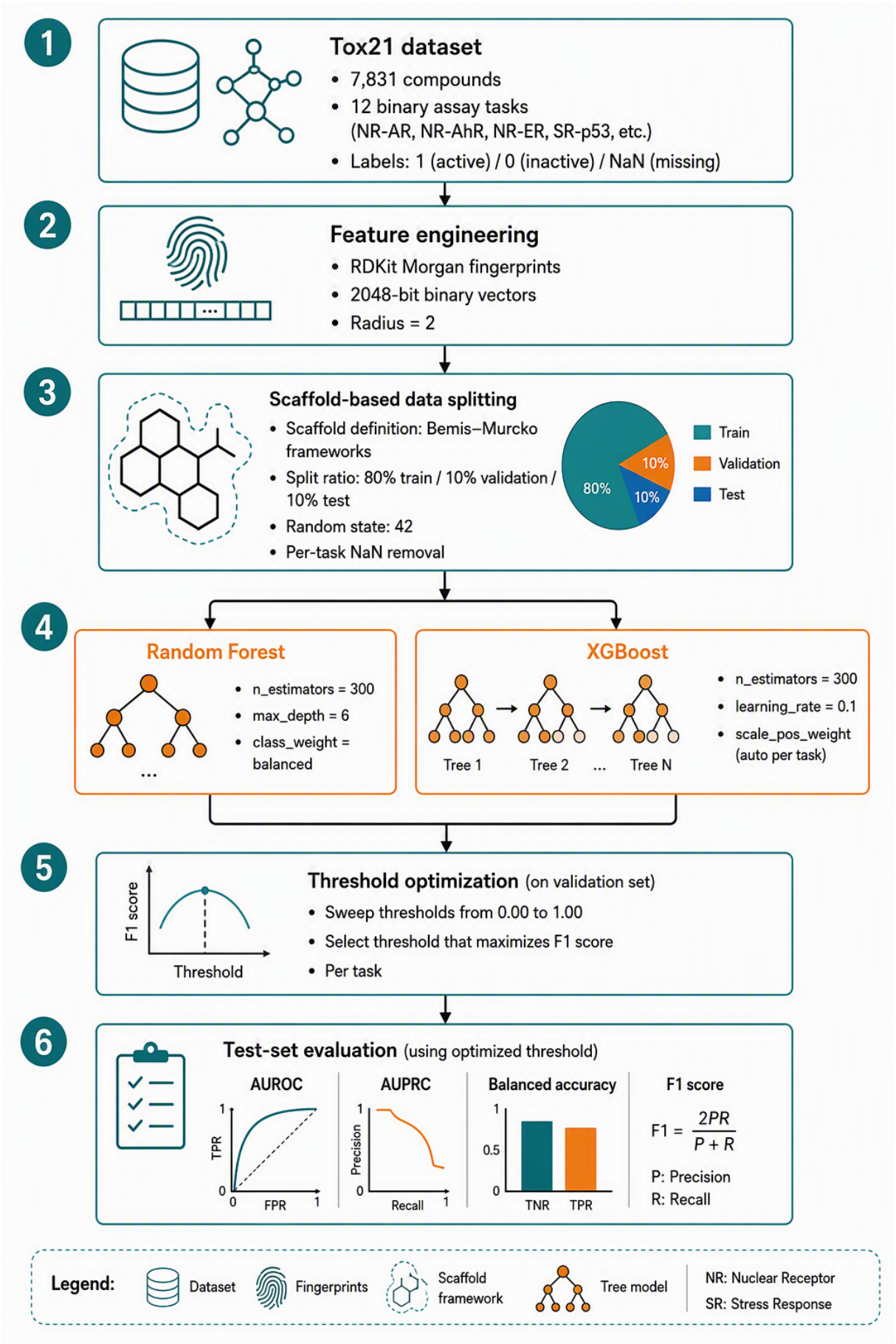

Figure 3: **Analysis pipeline.** The six stages of the benchmark: (1) Tox21 dataset with 12 binary assay labels and `NaN` missingness; (2) RDKit Morgan fingerprint featurization (2048 bits, radius 2); (3) global scaffold split (80/10/10, `random_state=42`) with per-task `NaN` removal; (4) parallel Random Forest and XGBoost training with explicit class-imbalance handling; (5) F1-optimal threshold selection on the validation split; (6) test-set evaluation across AUROC, AUPRC, balanced accuracy, and F1.

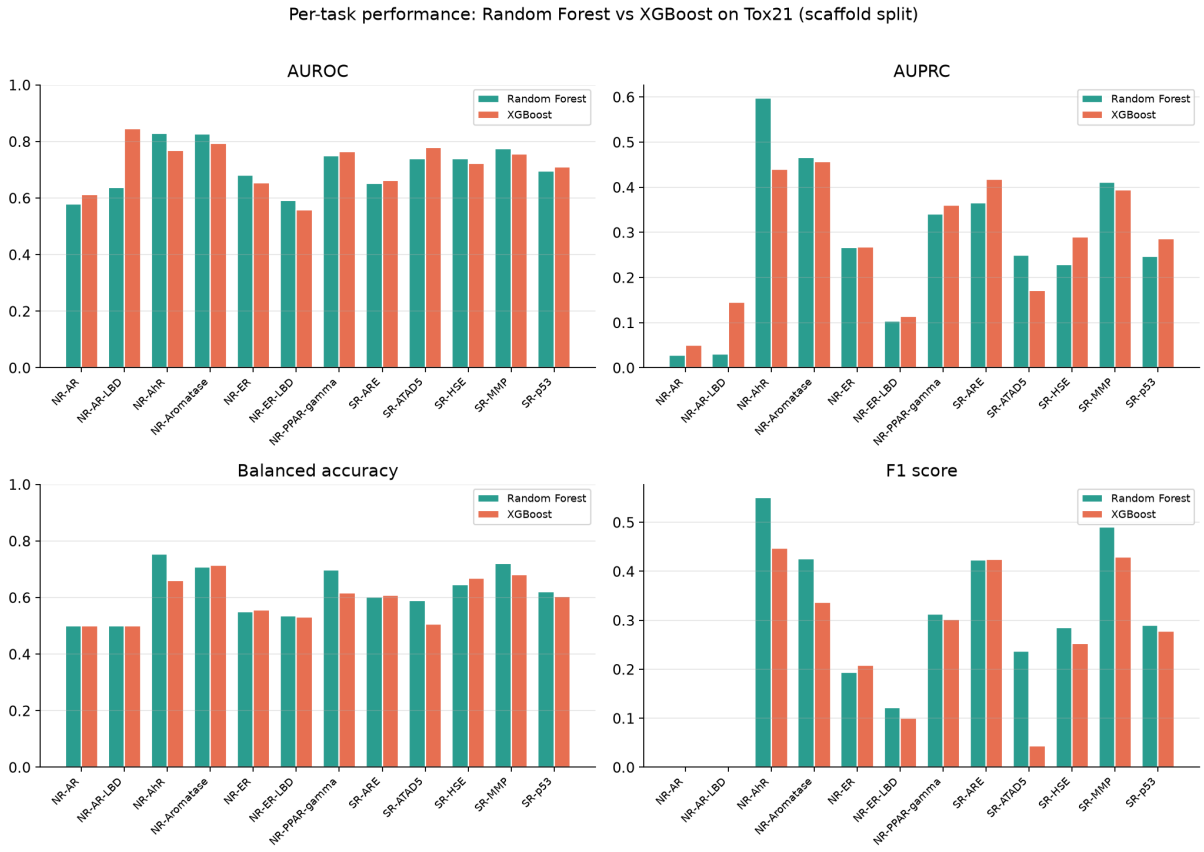

Figure 4: **Per-task performance across four metrics.** Grouped bars compare Random Forest (teal) and XGBoost (orange) for AUROC, AUPRC, balanced accuracy, and F1 across all 12 assay tasks. AUROC is comparatively uniform across tasks for a given model, whereas F1 collapses to zero on NR-AR and NR-AR-LBD, illustrating the divergence between ranking and thresholded minority-class performance under class imbalance.

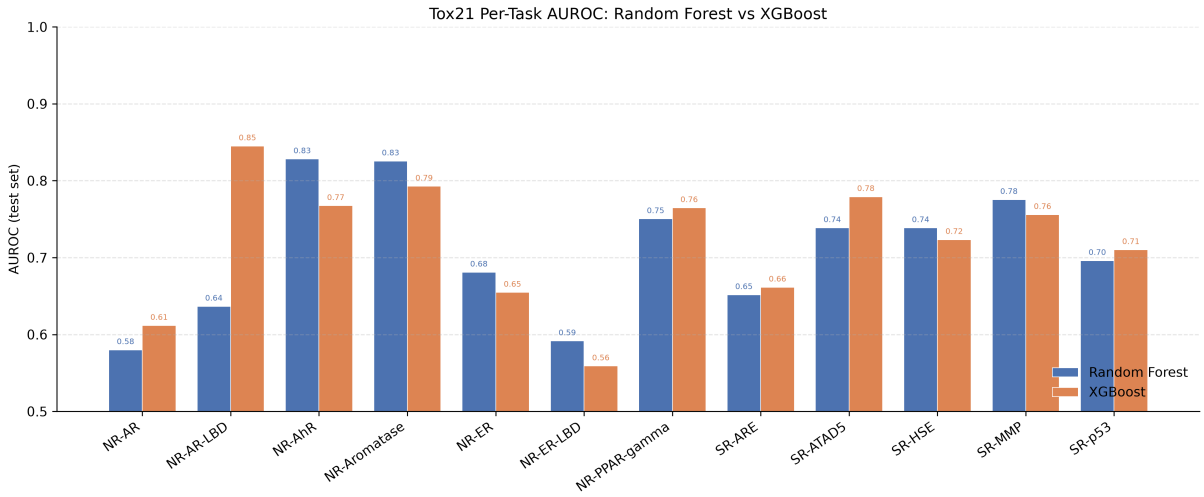

Figure 5: **Per-task AUROC: Random Forest vs. XGBoost.** Grouped bar chart comparing the two models' test-set AUROC across the 12 Tox21 assay tasks. Dashed bars of zero F1 (visible in Figure 4) on NR-AR and NR-AR-LBD coexist with non-trivial AUROC on the same tasks (e.g., XGBoost AUROC of 0.845 on NR-AR-LBD), illustrating that ranking ability and thresholded positive prediction can diverge sharply.

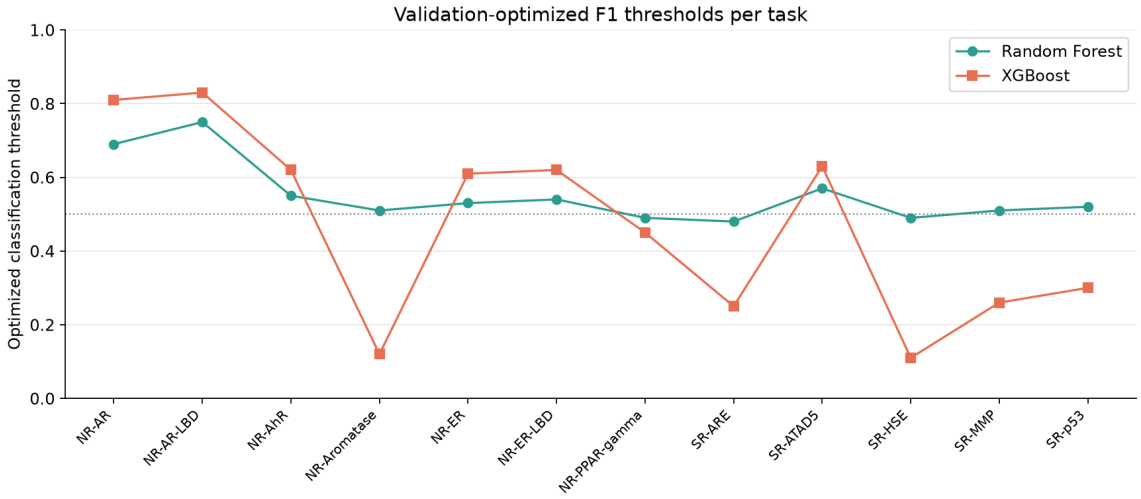

Figure 6: **Validation-optimized F1 thresholds per task.** Random Forest (teal) thresholds cluster near 0.5, while XGBoost (orange) thresholds are widely dispersed, with several tasks optimized far below 0.5—a consequence of `scale_pos_weight` inflating positive-class scores.