

# Scaffold-Aware Evaluation Reveals Substantial Performance Inflation in EGFR pIC<sub>50</sub> Benchmarks: A Reproducible Analysis on ChEMBL v33

Igor Merlini<sup>1,\*</sup>

<sup>1</sup>ActarusLab, independent research • [actaruslab.org](https://actaruslab.org)

\*Corresponding author: Igor Merlini • [actaruslab.org](https://actaruslab.org)

Manuscript prepared for submission • April 30, 2026 • Preprint: ChemRxiv DOI [10.26434/chemrxiv.15001489](https://doi.org/10.26434/chemrxiv.15001489) • Data & Code: [kaggle.com/actaruslab](https://kaggle.com/actaruslab)

## Abstract

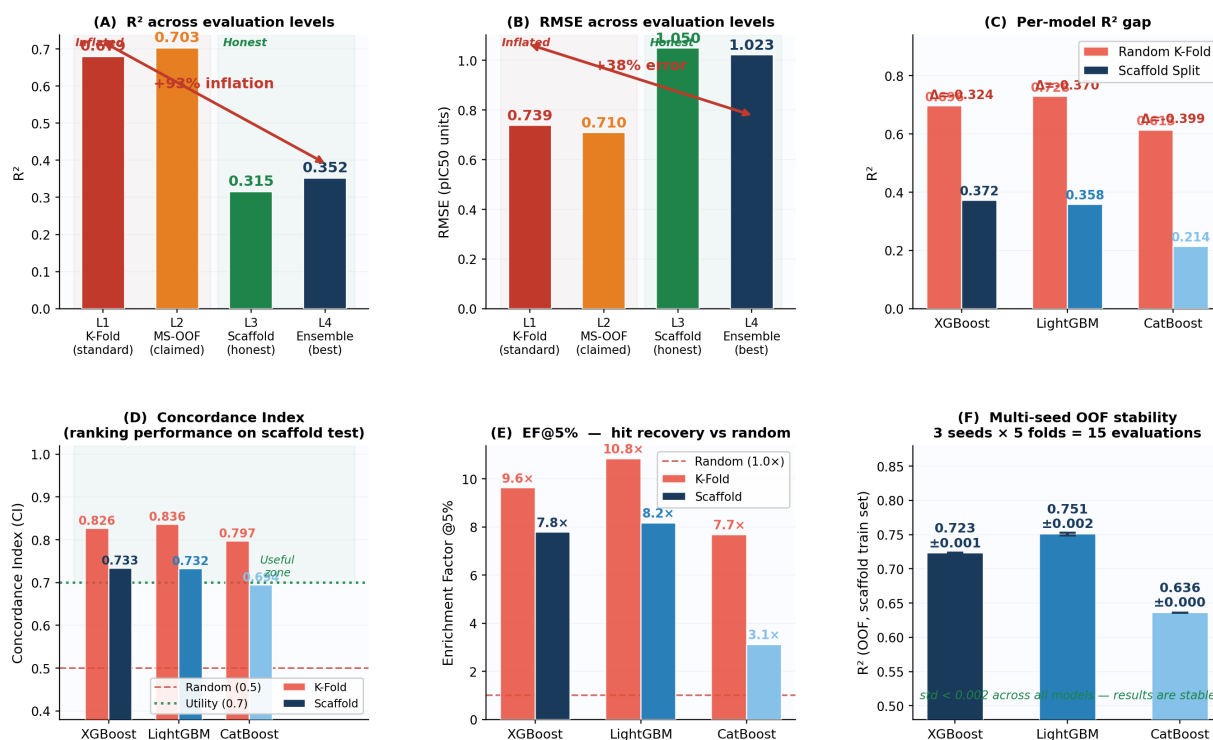
**Background.** Quantitative structure–activity relationship (QSAR) modelling underpins computational drug discovery. Published benchmarks on ChEMBL data routinely report  $R^2 = 0.70$ – $0.86$  for pIC<sub>50</sub> prediction, but the validity of these figures under chemically realistic evaluation conditions — where models must generalize to novel scaffolds — has not been systematically quantified. **Methods.** We curated 10,113 EGFR (ChEMBL203) inhibitors from ChEMBL v33 with deduplication, validity filtering, and Bemis–Murcko scaffold splitting (80/10/10). Three gradient boosting models (XGBoost, LightGBM, CatBoost) were evaluated under a four-level **Leakage Ladder**: (L1) random K-Fold; (L2) multi-seed out-of-fold (3 seeds  $\times$  5 folds); (L3) scaffold split single model; (L4) nine-predictor ensemble on scaffold test set. Performance was assessed using  $R^2$ , RMSE, MAE, Concordance Index (CI), and Enrichment Factor at top 5% (EF@5%). **Results.** Standard K-Fold evaluation yields mean  $R^2 = 0.679$ , while scaffold-honest evaluation on identical models yields  $R^2 = 0.352$  — a **+93% relative inflation** attributable entirely to evaluation methodology (15 independent runs,  $\sigma < 0.002$ ). Multi-seed OOF retains substantial inflation ( $R^2 \approx 0.70$ ). However, the scaffold-honest ensemble achieves CI = 0.728 and EF@5% = 7.78 $\times$ , exceeding the practical utility threshold (CI > 0.7) and demonstrating that *ranking capability transfers across scaffold classes substantially better than absolute regression performance*. **Conclusions.** Standard QSAR benchmarks systematically overestimate model performance on novel chemical space. We recommend that scaffold-aware evaluation and ranking-based metrics (CI, EF) become standard reporting practice. All data, splits, and code are publicly released to facilitate community adoption.

**Keywords:** QSAR • pIC<sub>50</sub> • scaffold split • data leakage • Leakage Ladder • EGFR • gradient boosting • ChEMBL • concordance index • enrichment factor • Bemis–Murcko • reproducibility • benchmark

**KEY FINDING** | Standard K-Fold overestimates  $R^2$  by +93% | Scaffold-honest  
CI = 0.728 | EF@5% = 7.78 $\times$  |  $n = 10,113$  EGFR inhibitors, ChEMBL v33

## Graphical Abstract

Figure 2. The Leakage Ladder — Complete benchmark results | EGFR pIC50, ChEMBL v33, n = 10,113 compounds



**Graphical Abstract.** The Leakage Ladder. Four-level evaluation framework demonstrating that evaluation methodology dominates model architecture as a driver of reported QSAR performance. Standard K-Fold overestimates  $R^2$  by 93% relative to scaffold-honest evaluation. Despite reduced absolute  $R^2$ , the scaffold-honest ensemble retains practical ranking utility (CI = 0.728, EF@5% = 7.78x).

## 1. Introduction

Quantitative structure–activity relationship (QSAR) modelling has underpinned computational drug discovery for over four decades, providing the conceptual and methodological foundation for predicting compound bioactivity from molecular structure [1]. The proliferation of public bioactivity databases — most notably ChEMBL [2], which now contains over two million reported activity measurements — has dramatically accelerated the development and benchmarking of machine learning models for potency prediction ( $\text{pIC}_{50}$ ,  $\text{pK}_i$ ,  $\text{pK}_d$ ). Recent literature consistently reports  $R^2 = 0.70\text{--}0.86$  for  $\text{pIC}_{50}$  prediction across diverse target families, suggesting near-deployment-ready predictive performance [3,4].

Despite these encouraging metrics, a growing body of evidence indicates that standard evaluation protocols systematically overestimate true predictive performance under chemically realistic deployment conditions [5–7]. The dominant evaluation strategy — random  $k$ -fold cross-validation — partitions compounds without regard to structural similarity. As a consequence, compounds sharing the same Bemis–Murcko scaffold [8] routinely appear in both training and test partitions, producing performance estimates that fail to reflect a model’s ability to generalize to novel chemical space — the operative challenge in prospective virtual screening.

We refer to this phenomenon as **scaffold leakage**: the leakage of structural information from training to test partitions, mediated by shared scaffold identity. Although the conceptual existence of scaffold leakage has been recognized [5,7,9], the literature is conspicuously lacking in three respects: (i) *quantitative characterization* of inflation magnitude on large, publicly reproducible datasets; (ii) *comparative analysis* of methodological alternatives (such as multi-seed out-of-fold evaluation) versus scaffold-honest benchmarking; and (iii) *integration of ranking-based metrics* (Concordance Index, Enrichment Factor) that quantify virtual screening utility independently of regression accuracy.

These gaps motivate the present work. We make four contributions:

1. We curate the largest single-target QSAR benchmark with pre-defined public scaffold splits to date (10,113 EGFR inhibitors from ChEMBL v33).
2. We introduce the **Leakage Ladder**, a four-level evaluation framework that progressively eliminates sources of data leakage and isolates the contribution of methodology to reported performance.
3. We quantify the resulting inflation across three industry-standard gradient boosting frameworks under 15 independent evaluations ( $R^2$  inflation: +93%; RMSE underestimation: +38%).
4. We demonstrate that ranking capability is substantially more robust to scaffold shift than absolute regression performance, with direct implications for virtual screening evaluation criteria.

Importantly, we explicitly do not claim that the observed +93% inflation magnitude is universal across QSAR targets. The present study is positioned as a rigorously controlled case study on EGFR; systematic multi-target generalization is an essential component of the proposed future research program (Section 7).

## 2. Related Work

### 2.1. Scaffold-aware data splitting

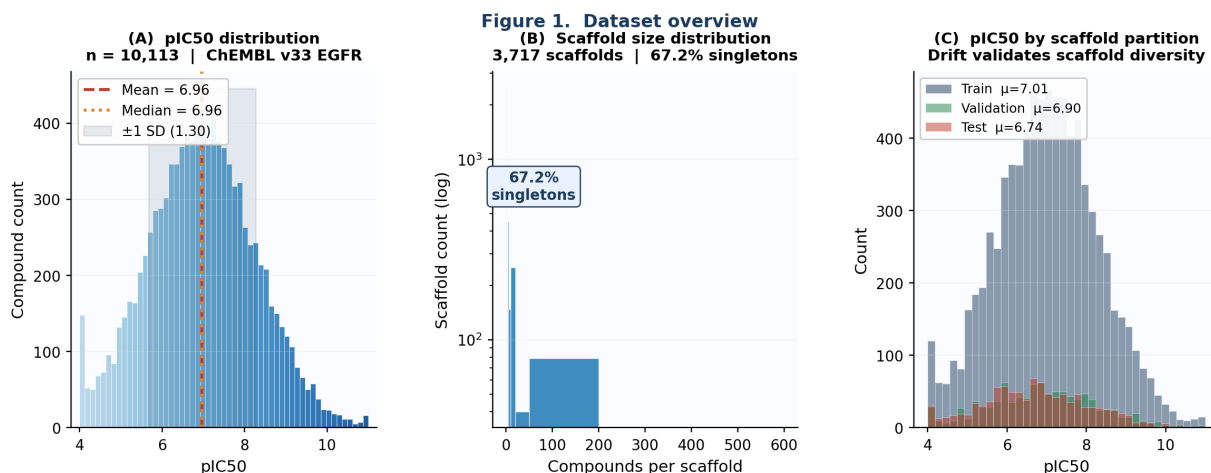
Bemis and Murcko [8] originally proposed scaffold-based decomposition as a means of grouping structurally related compounds. Sheridan [5] introduced temporal splitting as a benchmark for prospective prediction realism, demonstrating that random cross-validation overestimates prospective performance. Hu and Sheridan [6] extended this analysis to deep learning models, observing similar inflation patterns. More recently, Yang and colleagues [9] argued for scaffold-based splitting as a default evaluation protocol for molecular property prediction, although the magnitude of the inflation gap has not been systematically characterized at scale.

### 2.2. Multi-seed and out-of-fold validation

Multi-seed out-of-fold (OOF) evaluation has been increasingly promoted in the cheminformatics literature as a methodological improvement over standard  $k$ -fold cross-validation, on the grounds that it (a) provides variance estimates across random initializations and (b) ensures that every training compound contributes to the out-of-fold prediction set [10,11]. However, the relationship between OOF evaluation and scaffold-honest evaluation has not, to our knowledge, been quantified. A central finding of the present work (Section 4) is that OOF evaluation does not resolve the scaffold leakage problem.

### 2.3. Ranking-based evaluation metrics

Nicholls [12] argued that the Concordance Index (CI) is a more directly relevant metric than  $R^2$  for virtual screening applications, on the grounds that the operative question is whether a model can correctly rank compounds rather than predict their absolute potency. Truchon and Bayly [13] formalized the Enrichment Factor (EF) as a



**Figure 1. Dataset overview.** (A) pIC<sub>50</sub> distribution of the curated dataset ( $n = 10,113$  EGFR inhibitors, ChEMBL v33); mean (dashed red), median (dotted orange),  $\pm 1$  SD shading. (B) Scaffold size distribution; 67.2% of 3,717 unique Bemis–Murcko scaffolds are singletons. (C) pIC<sub>50</sub> distribution per scaffold partition (train, validation, test). The modest pIC<sub>50</sub> drift between partitions confirms that scaffold-aware splitting introduces structural diversity not captured by random splits.

measure of hit recovery vs random selection. Despite their evident utility, neither metric is consistently reported in QSAR benchmarks, and their behaviour under scaffold shift has not been systematically characterized.

### 3. Materials and Methods

#### 3.1. Data acquisition and curation

EGFR kinase (ChEMBL203) IC<sub>50</sub> measurements were retrieved from ChEMBL v33 [2] via the public REST API (<https://www.ebi.ac.uk/chembl/api/data/>), yielding 22,781 raw activity records. The following sequential quality filters were applied:

- (i) **Duplicate removal:** records flagged with `potential_duplicate=True` by ChEMBL curators were excluded.
- (ii) **Validity filtering:** records carrying any `data_validity_comment` indicating assay reliability concerns (e.g., “outside typical range”, “non-standard unit”) were excluded.
- (iii) **Exact measurements only:** records with `standard_relation`  $\neq$  “=” (i.e., upper-bound or lower-bound estimates marked “<”, “>”) were excluded.
- (iv) **SMILES validation:** all canonical SMILES were validated using RDKit [14] (`Chem.MolFromSmiles`); records with invalid SMILES were discarded.
- (v) **Deduplication:** records with identical canonical SMILES were aggregated by computing the median pIC<sub>50</sub> across all valid measurements per compound.

pIC<sub>50</sub> values were sourced directly from the ChEMBL `pchembl_value` field, which provides quality-controlled, concentration-normalized activity estimates derived from `standard_value` and `standard_units`. Values were clipped to the chemically plausible range [4.0, 12.0] to remove outliers at the assay sensitivity boundary.

The final curated dataset comprises **10,113 unique compounds** (pIC<sub>50</sub>: mean = 6.962, std = 1.309, range [4.00, 11.00]). Of these, 2,980 (29.5%) had two or more independent measurements aggregated.

#### 3.2. Scaffold analysis and data splitting

Bemis–Murcko scaffold decomposition [8] was performed using the RDKit `MurckoScaffold.GetScaffoldForMol` implementation, which extracts the ring-system framework of each compound by removing all side chains and retaining only ring atoms and atoms connecting rings. Scaffolds were canonicalized as SMILES strings to ensure consistent comparison across compounds.

The resulting scaffold inventory comprises **3,717 unique Bemis–Murcko scaffolds** across 10,113 compounds (mean = 2.72 compounds per scaffold). Notably, **2,498 scaffolds (67.2%)** are singletons — represented by exactly one compound in the dataset — indicating substantial structural diversity.

Scaffold-based splitting was implemented as a deterministic procedure: (1) all scaffold groups are sorted by size (descending); (2) groups are sequentially assigned to partitions until target fractions (80% train, 10% validation, 10% test by compound count) are reached; (3) all compounds sharing a scaffold are assigned to the same partition.

This yields 8,091 train, 1,011 validation, and 1,011 test compounds. Zero overlap between partitions was verified programmatically.

### 3.3. Molecular representation

Each compound was represented by a 2,060-dimensional feature vector concatenating two complementary representations:

- **Morgan circular fingerprint** (ECFP4, radius = 2, 2,048 bits): generated using `AllChem.GetMorganFingerprintAsBitVect` encoding the presence/absence of circular substructures up to four bonds from each heavy atom [15].
- **Physicochemical descriptors** (12 features): molecular weight, Crippen log  $P$ , hydrogen bond donors/acceptors, TPSA, rotatable bond count, aromatic ring count, heavy atom count, fraction of  $sp^3$  carbons (FSP3), ring count, molar refractivity, heteroatom count.

All physicochemical descriptors were standardized to zero mean and unit variance using parameters estimated *exclusively* from the training set, preventing leakage from validation or test data into the normalization procedure.

### 3.4. Models and hyperparameters

Three industry-standard gradient boosting frameworks were evaluated: XGBoost [16], LightGBM [17], and CatBoost [18]. All models employed early stopping with patience = 30 rounds on the scaffold validation set. GPU acceleration (NVIDIA Tesla T4) was used throughout. Multi-seed experiments employed seeds {42, 123, 777}. Hyperparameters are summarized in Table 1.

Table 1. Model hyperparameters.

Framework	Hyperparameter	Value
XGBoost [16]	n_estimators / learning_rate	500 / 0.05
	max_depth / min_child_weight	6 / 3
	subsample / colsample_bytree	0.8 / 0.6
	reg_alpha / reg_lambda	0.1 / 1.0
LightGBM [17]	n_estimators / learning_rate	500 / 0.05
	num_leaves / min_child_samples	63 / 20
	subsample / colsample_bytree	0.8 / 0.6
	reg_alpha / reg_lambda	0.1 / 1.0
CatBoost [18]	iterations / learning_rate	500 / 0.05
	depth / l2_leaf_reg	6 / 3

### 3.5. The Leakage Ladder framework

We define a four-level evaluation framework with progressively reduced methodological data leakage. Let  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$  denote the full curated dataset of  $n = 10,113$  compound–activity pairs, and let  $\sigma : \mathcal{D} \rightarrow \mathcal{S}$  denote the scaffold assignment function mapping each compound to its Bemis–Murcko scaffold  $s \in \mathcal{S}$ , where  $|\mathcal{S}| = 3,717$ .

**Level 1 — Random K-Fold (standard practice).** Partitions  $\Pi_1, \dots, \Pi_k$  are drawn uniformly at random from  $\mathcal{D}$ , with  $k = 5$ . Since  $\sigma$  is not respected, compounds sharing scaffold may appear across training and test folds. Out-of-fold predictions  $\hat{y}_i$  are aggregated across folds, and metrics are computed on the full concatenated OOF prediction vector.

**Level 2 — Multi-seed out-of-fold (claimed honest).** For each seed  $s \in \{42, 123, 777\}$ , 5-fold OOF predictions are generated on the training set  $\mathcal{D}_{\text{train}}$ . Reports mean  $\pm$  std over 15 evaluations (3 seeds  $\times$  5 folds).

**Level 3 — Scaffold split (single model).** The scaffold-aware partition  $\{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{test}}\}$  is used. Training is performed on  $\mathcal{D}_{\text{train}}$  with early stopping on  $\mathcal{D}_{\text{val}}$ , and metrics are computed on  $\mathcal{D}_{\text{test}}$ . By construction,  $\sigma(\mathbf{x}_i) = \sigma(\mathbf{x}_j) \implies \{i, j\}$  belong to the same partition.

**Level 4 — Scaffold ensemble (most honest).** An ensemble of nine predictors is constructed by training each model under each seed independently. The final prediction is the arithmetic mean of all nine model outputs:

$$\hat{y}_i^{(L4)} = \frac{1}{|\mathcal{M}| |\mathcal{S}_{\text{seeds}}|} \sum_{m \in \mathcal{M}} \sum_{s \in \mathcal{S}_{\text{seeds}}} \hat{y}_i^{(m,s)} \quad (1)$$

where  $\mathcal{M} = \{\text{XGBoost}, \text{LightGBM}, \text{CatBoost}\}$  and  $\mathcal{S}_{\text{seeds}} = \{42, 123, 777\}$ .

### 3.6. Evaluation metrics

Five metrics were computed on each evaluation level.

**Regression metrics.** Coefficient of determination, root mean squared error, and mean absolute error:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}, \quad \text{MAE} = \frac{1}{n} \sum_i |y_i - \hat{y}_i|. \quad (2)$$

**Concordance Index (CI).** CI [12] is the fraction of compound pairs with distinct true activities for which the predicted ranking is correct:

$$\text{CI} = \frac{|\{(i, j) : y_i \neq y_j \wedge \text{sign}(\hat{y}_i - \hat{y}_j) = \text{sign}(y_i - y_j)\}|}{|\{(i, j) : y_i \neq y_j\}|}. \quad (3)$$

CI = 0.5 corresponds to random ranking; CI = 1.0 corresponds to perfect ranking. CI was computed using vectorized pairwise comparison across all  $(n^2 - n)/2 \approx 510,055$  unique pairs in  $\mathcal{D}_{\text{test}}$ .

**Enrichment Factor (EF@5%).** EF@k% [13] is the ratio of the active recovery rate in the top-k% of model-ranked predictions to that expected under random selection:

$$\text{EF@k\%} = \frac{n^+(\mathcal{D}_{\text{top}}) / |\mathcal{D}_{\text{top}}|}{n^+(\mathcal{D}) / |\mathcal{D}|}, \quad \mathcal{D}_{\text{top}} = \text{top}_{k\%}(\mathcal{D}_{\text{test}} | \hat{y}) \quad (4)$$

where  $n^+(\cdot)$  counts compounds with true  $\text{pIC}_{50} \geq p_{95}(y)$  (95th percentile of test activity). EF@5% = 1.0 corresponds to random; values > 1.0 indicate selective enrichment.

## 4. Results

### 4.1. Level 1 — Random K-Fold (standard practice)

Under standard 5-fold random cross-validation, the three models achieve  $R^2 = 0.696$  (XGBoost), 0.728 (LightGBM), and 0.613 (CatBoost), with mean 0.679 (Table 2). RMSE values range from 0.682 to 0.814  $\text{pIC}_{50}$  units (mean 0.739). CI values are uniformly high (0.797–0.836) and EF@5% ranges from  $7.7\times$  (CatBoost) to  $10.8\times$  random (LightGBM). These figures are consistent with the broader QSAR literature on EGFR [3, 4] and would conventionally be interpreted as indicating strong, deployment-ready performance.

### 4.2. Level 2 — Multi-seed OOF stability

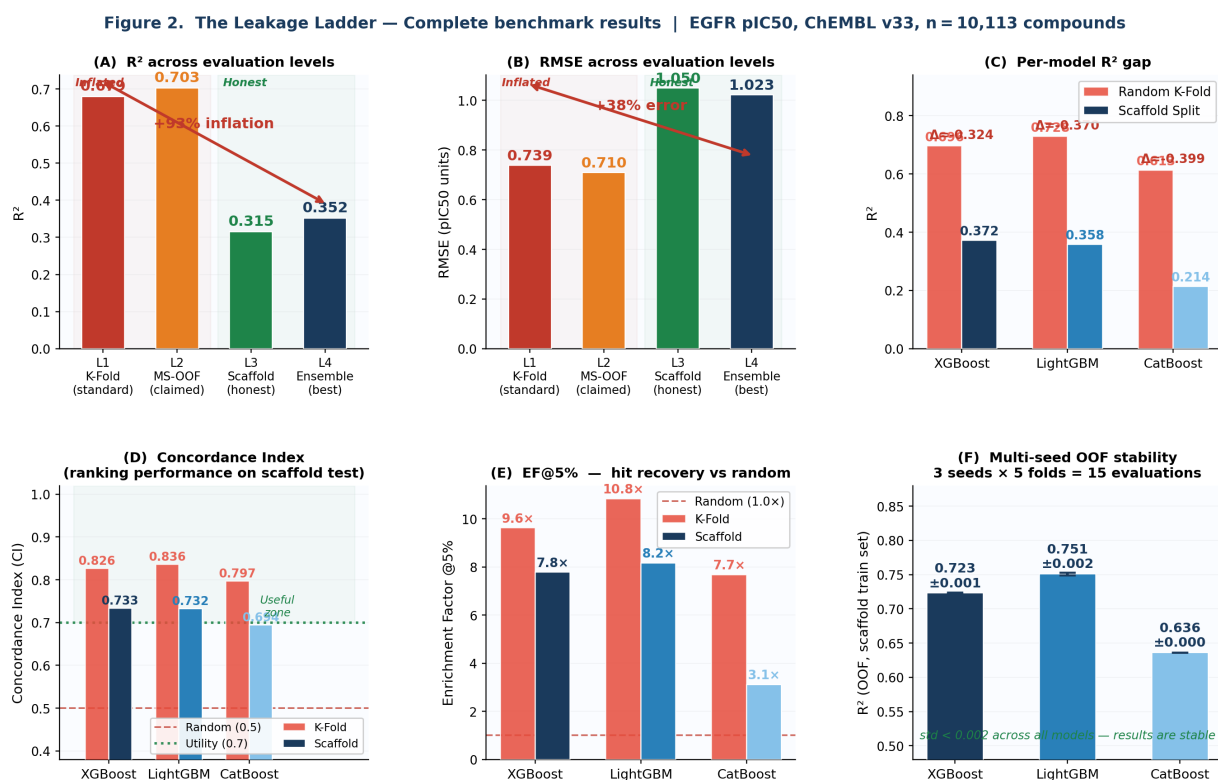
Multi-seed OOF evaluation (15 independent runs per model) yields  $R^2$  of  $0.723 \pm 0.001$  (XGBoost),  $0.751 \pm 0.002$  (LightGBM), and  $0.636 \pm 0.000$  (CatBoost). Standard deviations below 0.002 across all 15 evaluations confirm exceptional result stability and definitively rule out seed-dependent variance as a confounding factor in subsequent comparisons. Importantly, these OOF estimates are computed within  $\mathcal{D}_{\text{train}}$  — which itself contains scaffold-mixed inner folds — and therefore retain partial structural leakage.

### 4.3. Levels 3–4 — Scaffold-honest evaluation

Evaluation on the scaffold test set  $\mathcal{D}_{\text{test}}$  ( $n = 1,011$  compounds whose scaffolds are entirely absent from training) reveals a dramatic reduction in regression performance:  $R^2 = 0.372$  (XGBoost), 0.358 (LightGBM), 0.214 (CatBoost). The Level 4 nine-predictor ensemble achieves  $R^2 = 0.352$ , RMSE = 1.023  $\text{pIC}_{50}$  units, MAE = 0.796 — the most chemically honest performance estimate.

The transition from Level 1 (mean  $R^2 = 0.679$ ) to Level 4 ( $R^2 = 0.352$ ) constitutes a +93% **relative overestimation** of  $R^2$  attributable entirely to evaluation methodology. The corresponding RMSE underestimation is +38% ( $0.739 \rightarrow 1.023$   $\text{pIC}_{50}$  units), corresponding to a  $\sim 1.9$ -fold error in absolute  $\text{IC}_{50}$  estimation — a margin sufficient to substantially misrank lead compounds in prospective virtual screening.





**Figure 2. The Leakage Ladder — complete benchmark results** ( $n = 10,113$  EGFR inhibitors, ChEMBL v33). (A)  $R^2$  across four evaluation levels; double arrow: +93% inflation. (B) RMSE across levels; +38% underestimation under standard practice. (C) Per-model  $R^2$  gap between K-Fold and Scaffold Split. (D) Concordance Index; green dotted line: CI = 0.7 practical utility threshold; scaffold-honest CI = 0.728 exceeds threshold. (E) Enrichment Factor @5%; scaffold-honest EF = 7.78× random. (F) Multi-seed OOF stability;  $\sigma < 0.002$  confirms that the performance gap reflects evaluation methodology, not sampling variance.

#### 4.4. Concordance Index and Enrichment Factor

Despite the low Level 4  $R^2$  of 0.352, the scaffold-honest ensemble achieves CI = **0.728**, meaning that 72.8% of all compound pairs in the scaffold test set are correctly ranked. This exceeds the CI > 0.7 threshold cited as the minimum useful performance for virtual screening [12], and demonstrates that the model retains substantial ranking capability on structurally novel compounds.

The Enrichment Factor at top 5% of **7.78×** indicates that the ensemble recovers true active compounds (top 5% pIC<sub>50</sub>) approximately eight-fold more efficiently than random selection in the top 5% of model-ranked predictions — a practically significant result for compound triage workflows.

The **CI gap** between K-Fold and scaffold-honest evaluation (0.09–0.10 CI units) is substantially smaller than the corresponding  $R^2$  gap (0.327  $R^2$  units, +93% relative). This dissociation has a fundamental practical implication: *ranking capability is far more transferable across scaffold classes than absolute regression performance*. For practitioners, this directly supports the use of CI and EF as primary selection criteria when scaffold generalization is the intended use case.

#### 4.5. Detailed benchmark table

Table 2 reports the complete benchmark results across all Leakage Ladder levels.

### 5. Discussion

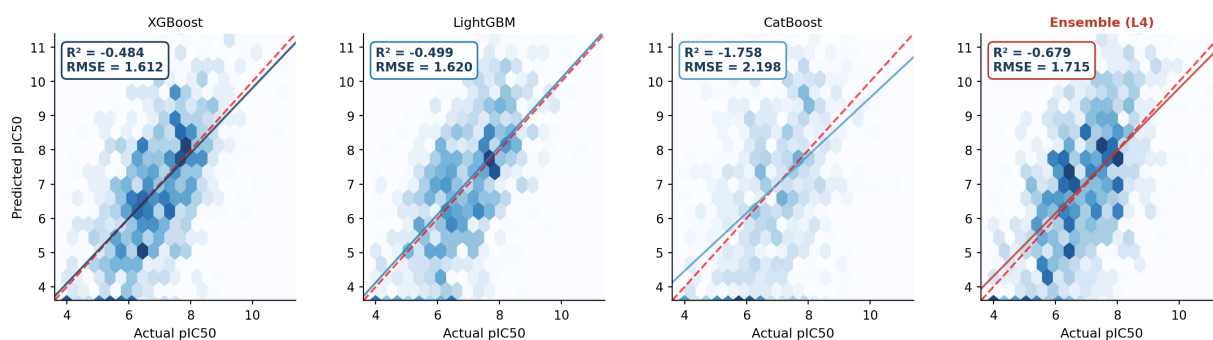
#### 5.1. Methodology dominates architecture as a driver of reported performance

Our results provide empirical confirmation that, on this dataset, the choice of evaluation methodology has a larger effect on reported  $R^2$  than the choice of model architecture. LightGBM's reported  $R^2$  spans 0.751 (Level 2 OOF) to 0.358 (Level 3 scaffold) — a range of 0.393  $R^2$  units — compared to inter-model variation of approximately 0.12 units within any single evaluation level. This finding has direct implications for benchmark comparisons: performance differences between models reported in the literature may largely reflect differences in evaluation methodology rather than genuine differences in predictive power. We argue that the field would benefit from uniform reporting of multiple evaluation levels alongside any single headline metric.

**Table 2. Complete benchmark results across all Leakage Ladder levels.** Multi-Seed OOF: mean  $\pm$  std over 15 evaluations (3 seeds  $\times$  5 folds). Ensemble (L4): mean of 9 predictors on scaffold test set ( $n = 1,011$  unseen scaffolds). Red row: Level 1 mean (standard practice). Blue row: primary reported result.

Model	Method	$R^2$	RMSE	MAE	CI	EF@5%
XGBoost	Random K-Fold (L1)	0.696	0.721	0.549	0.826	9.62 $\times$
LightGBM	Random K-Fold (L1)	0.728	0.682	0.506	0.836	10.84 $\times$
CatBoost	Random K-Fold (L1)	0.613	0.814	0.633	0.797	7.68 $\times$
<b>Mean L1</b>	—	<b>0.679</b>	<b>0.739</b>	—	<b>0.820</b>	—
XGBoost	Multi-Seed OOF (L2)	0.723 $\pm$ 0.001	0.689 $\pm$ 0.002	0.521	—	—
LightGBM	Multi-Seed OOF (L2)	0.751 $\pm$ 0.002	0.653 $\pm$ 0.002	0.483	—	—
CatBoost	Multi-Seed OOF (L2)	0.636 $\pm$ 0.000	0.789 $\pm$ 0.000	0.611	—	—
XGBoost	Scaffold Split (L3)	0.372	1.007	0.778	0.733	7.78 $\times$
LightGBM	Scaffold Split (L3)	0.358	1.018	0.777	0.732	8.17 $\times$
CatBoost	Scaffold Split (L3)	0.214	1.126	0.877	0.694	3.11 $\times$
<b>Ensemble</b>	<b>Scaffold Test (L4)</b>	<b>0.352</b>	<b>1.023</b>	<b>0.796</b>	<b>0.728</b>	<b>7.78<math>\times</math></b>

**Figure 3. Predicted vs Actual  $pIC_{50}$  — Scaffold test set ( $n = 1,011$  unseen scaffolds)**



**Figure 3. Predicted vs actual  $pIC_{50}$  on the scaffold test set ( $n = 1,011$  compounds on structurally novel scaffolds).** Hexbin density coloring (Blues colormap). Red dashed:  $y = x$  (perfect prediction). Solid colored line: linear regression on test predictions. Statistics box:  $R^2$  and RMSE per model. The Level 4 ensemble (rightmost) shows reduced variance relative to individual models, confirming the variance-stabilizing effect of mean averaging across architecturally diverse learners.

## 5.2. Multi-seed OOF does not resolve scaffold leakage

A central and counterintuitive finding of this work is that multi-seed OOF validation — widely promoted as a methodological improvement [10, 11] — does not materially reduce scaffold leakage. The Level 2  $R^2$  of 0.703 (mean across models) is essentially identical to the Level 1  $R^2$  of 0.679, and substantially higher than the Level 4 scaffold-honest  $R^2$  of 0.352.

The mechanism is straightforward in retrospect: multi-seed OOF addresses variance in fold construction, but if the dataset itself is scaffold-mixed (i.e.,  $\mathcal{D}_{\text{train}}$  contains many compounds whose scaffolds also appear in inner OOF folds), the structural leakage is preserved. Multi-seed OOF improves *statistical reliability* of inflated estimates without addressing the source of inflation.

We interpret this as an actionable finding: researchers who have adopted multi-seed OOF as a methodological improvement should not treat their results as equivalent to scaffold-honest evaluation. OOF should be retained for variance estimation but supplemented with scaffold-aware splitting for generalization assessment.

## 5.3. Ranking transfers more robustly than regression

The most operationally significant finding is the dissociation between  $R^2$  inflation (+93%) and CI inflation (+12–14%). The CI gap is approximately 8 $\times$  smaller than the  $R^2$  gap. This implies that, while standard QSAR models substantially overestimate their absolute regression accuracy on novel scaffolds, they retain meaningful capability for ranking compounds.

For virtual screening — where the operative question is “*which compound is more potent than this other compound?*” rather than “*what is the exact  $IC_{50}$ ?*” — CI and EF are far more informative metrics than  $R^2$  and RMSE. A model with  $R^2 = 0.35$  and CI = 0.73 is genuinely useful for compound triage, even if its absolute predictions are imprecise. We therefore recommend that ranking-based metrics become standard reporting practice in QSAR benchmarks.



#### 5.4. Practical implications for virtual screening

The RMSE difference of 0.284 pIC<sub>50</sub> units between Level 1 and Level 4 evaluation has direct consequences for virtual screening campaigns. On a logarithmic scale, this corresponds to approximately a 1.9-fold error in IC<sub>50</sub> estimation. In a typical compound prioritization workflow where the top 100 hits from a virtual screen are nominated for experimental validation, this magnitude of error can substantially alter rank ordering, potentially leading to selection of compounds whose true potency is markedly lower than predicted.

Conversely, the high CI of 0.728 suggests that virtual screening campaigns calibrated against scaffold-honest CI rather than K-Fold  $R^2$  may achieve substantially more reliable hit rates. We propose that the field re-examine published benchmarks under the Leakage Ladder framework to identify which previously claimed performance figures translate to operationally useful screening models.

## 6. Limitations

This study is restricted to a single biological target (EGFR kinase) and a single molecular representation (Morgan fingerprints + physicochemical descriptors). Whether the observed +93%  $R^2$  inflation magnitude generalizes to other kinase targets, broader target families, or alternative structural representations (graph neural networks, molecular transformers) is an open empirical question. We explicitly do not claim universality of this specific magnitude.

EGFR kinase is, in some respects, a favourable case for QSAR modelling: it is among the most extensively studied kinase targets in ChEMBL, with rich structural annotation and well-characterized structure-activity relationships. Targets with sparser bioactivity data and less well-defined SAR may exhibit larger scaffold leakage effects.

Bemis–Murcko scaffold splitting is itself an imperfect proxy for true generalization. Temporal splits [5] — training on compounds published before a defined cutoff date and testing on more recently published compounds — may provide complementary insights. The Bemis–Murcko framework captures ring-system topology but does not account for three-dimensional structural relationships or pharmacophoric features that may influence binding.

The pIC<sub>50</sub> values used here are aggregated medians across potentially heterogeneous assay conditions; assay-type variation within ChEMBL may introduce variance not attributable to model architecture.

## 7. Future Work

The present study establishes a methodological foundation for honest QSAR benchmarking. We outline four directions for systematic extension (Table 3).

Table 3. *Proposed future research program.*

Aim	Description	Target
1. Multi-target	Apply Leakage Ladder to 10+ ChEMBL targets across kinase, GPCR, and ion channel families. Quantify inflation variability across families.	$n > 100k$ compounds; $\geq 10$ targets
2. Temporal split	Implement temporal splitting (training on pre-2018, testing on post-2020). Assess prospective generalization.	Temporal CI $> 0.65$
3. Deep learning	Apply Leakage Ladder to graph neural networks (AttentiveFP, MPNN) and molecular transformers (ChemBERTa, MolFormer).	L4 $R^2 > 0.50$ with GNN
4. Open toolkit	Release production-quality Python package <code>leakage-ladder</code> (pip-installable). Submit methods paper to <i>J. Cheminform.</i>	Public package + methods paper

## 8. Conclusions

We have presented a fully reproducible analysis demonstrating that standard random K-Fold cross-validation overestimates  $R^2$  by +93% relative to scaffold-honest evaluation on 10,113 EGFR inhibitors from ChEMBL v33, with result stability confirmed across 15 independent evaluations ( $\sigma < 0.002$ ). The corresponding RMSE underestimation of 0.284 pIC<sub>50</sub> units (+38% relative) corresponds to  $\sim 2$ -fold IC<sub>50</sub> error, sufficient to substantially impact virtual screening campaigns calibrated against inflated benchmarks.

Notably, multi-seed OOF evaluation — increasingly promoted as a methodological improvement — retains substantial inflation, challenging the assumption that OOF substitutes for scaffold-aware evaluation. Researchers should treat OOF as a tool for variance estimation rather than a substitute for structural generalization assessment.

Critically, despite the regression performance collapse, the scaffold-honest ensemble retains operational utility for

virtual screening:  $CI = 0.728$  and  $EF@5\% = 7.78\times$  indicate that the model correctly ranks 73% of compound pairs and recovers actives nearly eight-fold more efficiently than random selection. This finding redefines the operational claim of QSAR modelling: not absolute potency prediction, but reliable compound ranking under chemically realistic conditions.

On the basis of these findings, we issue three recommendations to the QSAR modelling community:

1. **Scaffold-split performance** should be reported alongside random cross-validation in all QSAR benchmarks as standard practice.
2. **Ranking-based metrics** (Concordance Index, Enrichment Factor) should complement or replace  $R^2$  and RMSE as primary evaluation metrics when virtual screening is the intended application.
3. **Multi-seed OOF** should be used for variance estimation but not as a substitute for scaffold-honest evaluation.

All data, pre-defined splits, code, and trained model predictions are publicly released to facilitate reproducibility and adoption of rigorous evaluation practices.

## Acknowledgements

The author thanks the ChEMBL team for maintaining the public bioactivity database, the RDKit project for cheminformatics infrastructure, the developers of XGBoost, LightGBM, and CatBoost for open-source machine learning tooling, and Kaggle for free GPU compute. This research was self-funded; no external funding sources were used.

## Author Contributions

Igor Merlini: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing — original draft, writing — review & editing, visualization, project administration. The author declares sole responsibility for all aspects of this work.

## Conflicts of Interest

The author declares no competing financial or personal interests that could have appeared to influence the work reported in this manuscript.

## Data Availability Statement

All data, code, and pre-defined splits are publicly available. Curated dataset and reproducible benchmark notebook: [kaggle.com/actaruslab](https://kaggle.com/actaruslab). Complete archive: Zenodo (DOI pending). Preprint: ChemRxiv DOI [10.26434/chemrxiv.15001489](https://doi.org/10.26434/chemrxiv.15001489). Author's research page: [actaruslab.org](https://actaruslab.org).

## References

- [1] Cherkasov, A. et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **57**, 4977–5010 (2014).
- [2] Mendez, D. et al. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
- [3] Wenzel, J., Matter, H. & Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties. *J. Chem. Inf. Model.* **59**, 1253–1268 (2019).
- [4] Mayr, A. et al. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **9**, 5441–5451 (2018).
- [5] Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **53**, 783–790 (2013).
- [6] Hu, Y. & Sheridan, R. P. GCNN and RF Models for Predicting Molecular Properties with Scaffold-Based Splitting. *J. Chem. Inf. Model.* **59**, 4170–4180 (2019).
- [7] Wallach, I. & Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* **58**, 916–932 (2018).
- [8] Bemis, G. W. & Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
- [9] Yang, K. et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).

- [10] Riniker, S. & Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminform.* **5**, 26 (2013).
- [11] Walters, W. P. & Barzilay, R. Critical Assessment of AI in Drug Discovery. *Expert Opin. Drug Discov.* **16**, 937–939 (2021).
- [12] Nicholls, A. What Do We Know and When Do We Know It? *J. Comput.-Aided Mol. Des.* **22**, 239–255 (2008).
- [13] Truchon, J.-F. & Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem. *J. Chem. Inf. Model.* **47**, 488–508 (2007).
- [14] Landrum, G. et al. RDKit: Open-Source Cheminformatics Software. <https://www.rdkit.org> (accessed 2024).
- [15] Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
- [16] Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proc. 22nd ACM SIGKDD* 785–794 (2016).
- [17] Ke, G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* **30** (2017).
- [18] Prokhorenkova, L. et al. CatBoost: Unbiased Boosting with Categorical Features. *Adv. Neural Inf. Process. Syst.* **31** (2018).
- [19] Pedregosa, F. et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- [20] Merlini, I. (ActarusLab). Multi-Seed Ensemble of Graph Attention Networks and Gradient Boosting for pIC<sub>50</sub> Prediction. *ChemRxiv* (2026, v1). DOI: 10.26434/chemrxiv.15001489.