A Standard Machine-Learning Virtual Screening Pipeline Produces Candidates Indistinguishable From Random Drug-Like Compounds: A Quantitative Audit Using COX-2 As Case Study

Winston Zai Lin
Blair Academy, Blairstown, NJ 07825, USA
linzaiwinston413@gmail.com
ORCID: 0009-0003-6036-8149

July 2026

Abstract

Motivation: Machine-learning (ML) virtual screening pipelines report strong retrospective benchmark performance (Pearson $R > 0.8$) and convert it into prioritized candidate lists. Whether those candidates dock better than random drug-like compounds against the actual target is rarely tested. We test it.

Methods: Random Forest regressors were trained on COX-2 (CHEMBL230, $n = 4{,}445$; Pearson $R = 0.756$, RMSE $= 0.81$ pIC_{50}) and COX-1 (CHEMBL221, $n = 1{,}652$; $R = 0.682$, RMSE $= 0.72$ pIC_{50}) and applied to 249,455 ZINC250k compounds. Potency, Lipinski, solubility-window ($-1 \leq \log P \leq 3$), and PAINS filters returned 42 candidates, each annotated with predicted COX-1 selectivity, Murcko scaffold, max-Tanimoto applicability-domain coverage, synthetic accessibility (SA), and BRENK alerts. All 42 candidates and a 100-compound random ZINC250k control (sampled before any ML scoring) were docked into the chain-A binding pocket of human COX-2 (PDB 1CX2) with AutoDock Vina. Celecoxib was redocked as positive control; pose recovery was scored by heavy-atom RMSD to the crystal ligand.

Results: Celecoxib redocked at -11.23 kcal/mol with 0.63 Å RMSD to the crystal ligand—protocol valid. All 42 candidates docked successfully (Vina best -11.06, median -7.58, worst -3.63 kcal/mol; 11 reached ≤ -8.0); **none beat celecoxib**. 39 of 42 candidates (93%) lay outside the applicability domain ($T \geq 0.40$; mean max-Tanimoto $= 0.243$). Under leakage-free Bemis–Murcko scaffold cross-validation the benchmark $R = 0.756$ itself collapsed to 0.503 ± 0.128 (RMSE $0.81 \rightarrow 1.06$)—the model explains only $\sim$25% of pIC_{50} variance on novel scaffolds, which is the regime the screen operates in. The candidate Vina distribution was **statistically indistinguishable from the random baseline** (Mann-Whitney $p = 0.89$; Welch $p = 0.68$; Cohen's $d = -0.08$; Hodges–Lehmann shift -0.02 kcal/mol, 95% CI $[-0.42, +0.34]$). ML pIC_{50} and Vina rankings were uncorrelated (Spearman $\rho = -0.086$, 95% CI $[-0.35, +0.20]$). The top-20 ML-prioritized candidates engaged celecoxib's pocket residues at parity with random ZINC (Val523: 16/20 vs. 16/20; mean contact recovery 0.73 vs. 0.66). A Vina-on-training validation showed that Vina at our settings *did* separate genuine ChEMBL COX-2 actives from random ZINC (actives median -8.40 vs. random -7.53 kcal/mol, $p = 0.006$)—yet the ML candidates docked at the random level (-7.58), not the actives level. Candidates were synthesizable (SA ≤ 5 for all 42, mean 3.01); 19/42 carried BRENK alerts.

Conclusions: A standard Random Forest screen at $R = 0.756$ produces no candidate enrichment over random ZINC by any objective measure: no docking gain, no positive ML–Vina rank correlation, 93% out-of-domain prediction. Critically, the same docking oracle *does* separate genuine COX-2 actives from random ZINC—so the failure is a property of the ML candidate list, not of the evaluation. The benchmark $R = 0.756$ is itself inflated by scaffold leakage—it falls to ≈ 0.50 under scaffold-split cross-validation—and does not predict prospective hit-finding on cross-library screens. Scaffold-split validation, applicability-domain

coverage, and a same-library random-baseline comparison should be mandatory reporting elements in ML virtual screening papers.

Keywords: virtual screening; applicability domain; molecular docking; COX-2; cyclooxygenase; Random Forest; AutoDock Vina; selectivity; Murcko scaffold; ML reliability

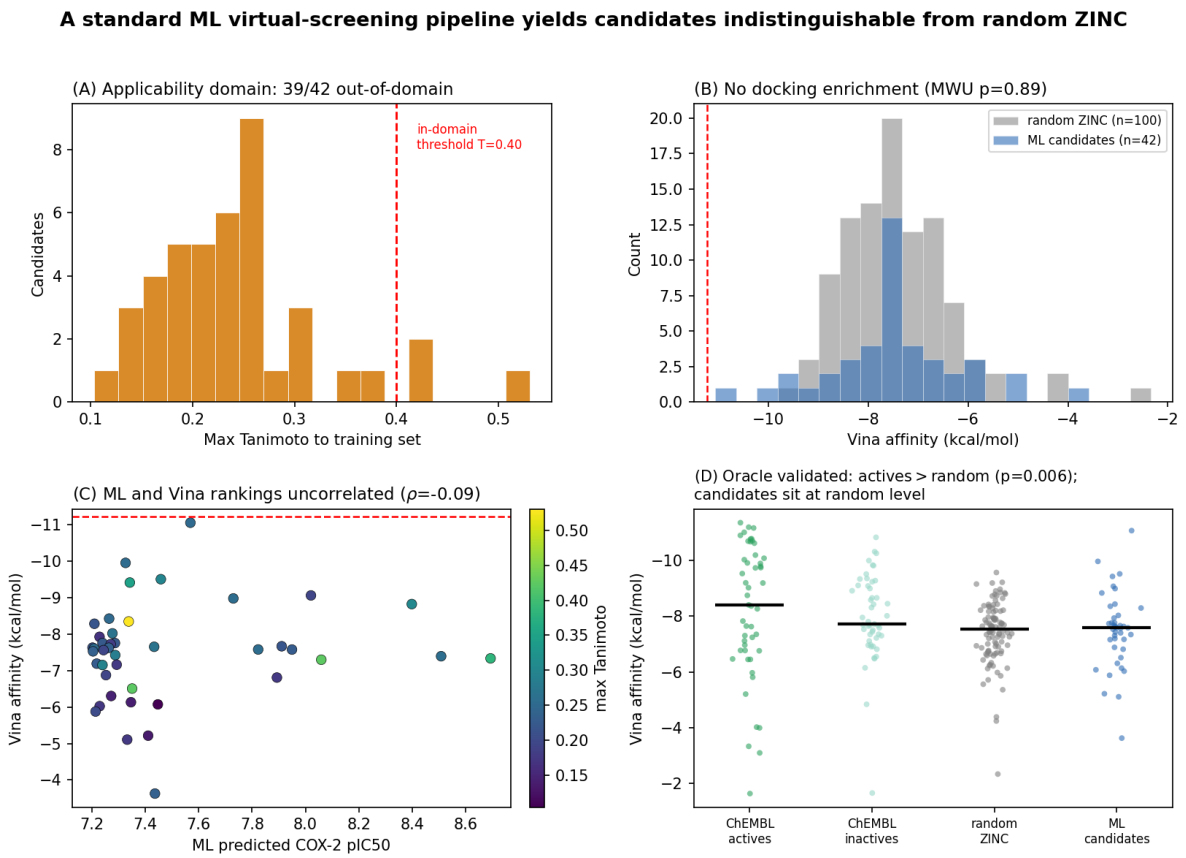

Figure 1: *

Graphical abstract. A standard ML virtual screening pipeline for COX-2 yields 42 candidates that fail three convergent objective checks: (A) 39 of 42 lie outside the model's applicability domain ($T \geq 0.40$); (B) their Vina docking-affinity distribution against the COX-2 chain-A binding pocket is statistically indistinguishable from a random ZINC250k baseline; (C) the ML pIC_{50} ranking and Vina ranking are uncorrelated within the candidate set. A docking-based validation confirms Vina separates true COX-2 actives from random ZINC, so the null is not an oracle artifact. The benchmark test-set $R = 0.756$ does not predict prospective screening performance.

1 Introduction

Machine-learning approaches to ligand-based virtual screening have become routine: a model is trained on a bioactivity database such as ChEMBL [6], applied to a generic drug-like library such as ZINC [9], and candidates above a potency threshold are reported as prioritized hits. Benchmark performance—typically reported as Pearson correlation or AUC on a random held-out split—is often strong, with $R > 0.80$ achievable using simple Random Forest regressors on Morgan fingerprints [7,11]. Such pipelines have populated the literature and supported industrial decision-making.

However, three properties of the standard pipeline conspire to inflate apparent reliability. First, random train/test splitting permits structurally similar compounds to appear in both sets, leaking information across the split [26, 27]. Second, the screening library is, by construction, structurally distinct from the training data: the purpose of cross-library screening is to identify novel chemotypes. Third, model confidence calibration is rarely reported, leaving consumers of computational predictions without quantitative warning when predictions extrapolate.

The applicability domain (AD) concept [22, 23] addresses the third gap: it provides a structural test for whether a candidate compound lies within the chemical region where the model's predictions are credible. For fingerprint-based regressors, maximum Tanimoto similarity to the training set is a standard AD metric, with $T \geq 0.40$ commonly used as the threshold below which predictions should be flagged as extrapolations [24, 25].

Despite the long-standing availability of AD methods, applicability domain analysis is conspicuously absent from many recently published ML virtual screening papers, including high-profile demonstrations of "AI-discovered" inhibitor candidates. This omission has been criticized in the cheminformatics literature [28, 29] but persists in practice.

In this work, we apply applicability-domain analysis to a standard COX-2 ML virtual screening pipeline, quantify the prevalence of out-of-domain extrapolation, and ask the most direct empirical question: *does ML potency prioritization plus selectivity, Lipinski, and PAINS filtering produce candidates that dock better than random drug-like compounds against the actual protein structure?* To answer this, we dock all 42 ML-prioritized candidates and an unfiltered random sample of 100 ZINC250k compounds against the crystal binding pocket of human COX-2 (PDB 1CX2 chain A), with celecoxib as a redocking positive control.

Cyclooxygenase catalyzes the committed step of prostaglandin biosynthesis and is the molecular target of nonsteroidal anti-inflammatory drugs [1]; the inducible COX-2 isoform became a major drug-discovery target once selective inhibitors were developed to spare the gastroprotective COX-1 isoform [2]. Cyclooxygenase-2 (COX-2) was selected as the target because: (i) it has a large, high-quality public bioactivity dataset; (ii) a well-resolved crystal structure (PDB 1CX2) with a co-crystallized selective inhibitor (SC-558, the bromophenyl analog of celecoxib) provides a defined docking pocket; (iii) the COX-1/COX-2 selectivity question is medicinally important and computationally tractable as a paired prediction; and (iv) the clinical-translation history of the coxib drug class [3–5] provides a calibrating context for what virtual-screening outputs do and do not predict.

We do not claim experimental confirmation of any candidate's COX-2 inhibitory activity. We provide an audited candidate list with explicit AD coverage flags and orthogonal physics-based scoring, intended to support downstream wet-lab triage decisions.

2 Methods

2.1 Bioactivity Datasets

Bioactivity data were retrieved from ChEMBL [6] via the `chembl_webresource_client` API for human COX-2 (CHEMBL230, Prostaglandin G/H synthase 2) and COX-1 (CHEMBL221, Prostaglandin G/H synthase 1) on July 9, 2026. Records were filtered to retain IC_{50} measurements with standard relation "=", assay type "B" (binding), and parseable pIC_{50} values. Duplicate SMILES entries were resolved by retaining the lowest IC_{50}. Activities were restricted to $pIC_{50} \in [3, 12]$. After filtering: $n_{\mathrm{COX-2}} = 4{,}445$ and $n_{\mathrm{COX-1}} = 1{,}652$ compound–activity pairs. Target-identity integrity was confirmed by a chemotype audit of the retrieved sets (COX-2: 12.2% primary sulfonamide, 15.5% trifluoromethyl—the coxib chemotype; $< 0.5\%$ benzamidine/guanidine protease motif).

The screening library was ZINC250k [9] ($n = 249{,}455$ drug-like compounds).

2.2 Molecular Representation

Molecular structures were encoded as 2048-bit Morgan fingerprints with radius 2 [7] using RDKit [8]. Invalid SMILES were discarded. Feature vectors were standardized via z-score normalization before model fitting.

2.3 ML Models

For each isoform independently, we trained a Random Forest regressor (500 estimators, scikit-learn [11] defaults) on a 70/15/15 random train/validation/test split (`random_state=42`). The same fingerprint representation, normalization, and training protocol were used for both targets. Inter-tree prediction standard deviation served as a per-prediction uncertainty proxy.

For comparison, mean baseline, Ridge regression, and XGBoost [12] (500 trees) were trained on identical splits.

We note that random splitting may inflate generalization estimates relative to scaffold-based splitting [26]; we use it here for comparability with prior literature and report AD analysis as an external check on this potential overestimate.

2.4 Virtual Screening Filters

Each ZINC250k compound was scored by the COX-2 RF model. Candidates were retained if they satisfied: predicted $\mathrm{pIC}_{50} > 7.2$; Lipinski's Rule of Five [10]; $-1 \leq \log P \leq 3$; and no PAINS A/B/C flags via the RDKit catalogs [13]. The $\log P$ filter is a heuristic for aqueous solubility and passive permeability, not a validated gastric compatibility filter.

2.5 Selectivity Prediction

Predicted isoform selectivity was defined as

$$\Delta\mathrm{pIC}_{50} = \mathrm{pIC}_{50}^{\mathrm{COX-2}} - \mathrm{pIC}_{50}^{\mathrm{COX-1}}.$$

Combined prediction uncertainty is propagated as $\sigma_\Delta = \sqrt{\mathrm{RMSE}_{\mathrm{COX-2}}^2 + \mathrm{RMSE}_{\mathrm{COX-1}}^2} \approx 1.08$ pIC_{50} units.

2.6 Applicability Domain

Maximum Tanimoto similarity [8] between each candidate and a random 500-compound subsample of the COX-2 training set was computed (Morgan, radius 2, 2048 bits). Candidates with $\max T \geq 0.40$ are classified as in-domain [24]; those below are classified as extrapolations.

2.7 Scaffold Analysis

Murcko scaffold decomposition [14] was performed using `rdkit.Chem.Scaffolds.MurckoScaffold`. Canonical scaffold SMILES served as cluster identifiers.

2.8 Molecular Docking

The human COX-2 crystal structure (PDB 1CX2 [15]) was downloaded from the RCSB Protein Data Bank. 1CX2 is a homo-tetramer (chains A–D, each in complex with the selective inhibitor SC-558—the bromophenyl analog of celecoxib—residue code S58); we restrict to chain A throughout to avoid averaging over four spatially distinct binding pockets. The centroid of the chain-A SC-558 heavy atoms (geometric centroid: $x, y, z = 24.26, 21.53, 16.50$ Å) defined the docking-box center, with cubic box dimensions of $22 \times 22 \times 22$ Å—sufficient to enclose the celecoxib-class binding pocket with margin.

Water molecules, the co-crystal SC-558 ligand, heme cofactor, and N-acetylglucosamine glycans of chain A were removed; the remaining chain-A protein was protonated at physiological pH (7.4) and converted to PDBQT format using OpenBabel [18] (`obabel -xr -p 7.4`).

For each ligand, the SMILES string was converted to a 3D conformer using RDKit's ETKDG embedding algorithm [19] followed by UFF energy minimization [20]. Ligand PDBQTs were generated using Meeko [21].

Docking was performed with AutoDock Vina 1.2.5 [16, 17] (`exhaustiveness` = 8, `num_modes` = 5). The Vina binding affinity (kcal/mol) for the lowest-energy pose was retained as the docking score. Celecoxib was redocked into the same protocol as a positive control; pose recovery was quantified by heavy-atom RMSD between the lowest-energy redocked celecoxib pose and the chain-A crystal SC-558 coordinates (26 matched heavy atoms), using the Hungarian algorithm to handle the unordered atom correspondence between PDBQT and PDB representations.

2.9 Random ZINC Baseline

To test whether the ML potency-prioritization-and-filtering pipeline produces docking enrichment relative to the raw ZINC250k library, 100 compounds were sampled uniformly at random from ZINC250k (`random_state=42`, applied *before any ML scoring or filtering*). These compounds were prepared and docked into the identical chain-A pocket with identical Vina settings. The full Vina affinity distributions of the 42 ML-prioritized candidates and the 100-compound random baseline were compared by Mann–Whitney U (two-sided) and Welch's t-test. We report both because Vina scores are bounded above and weakly skewed.

2.10 Synthetic Accessibility and BRENK Structural Alerts

Synthetic accessibility (SA) scores were computed using the RDKit Contrib implementation of the Ertl–Schuffenhauer SA score [35], which combines fragment contributions with a complexity penalty on a 1 (easy)–10 (hard) scale. Structural alerts were computed using the BRENK filter catalog (105 substructures associated with toxicity, reactivity, or poor pharmacokinetics) [36] as provided by the RDKit `FilterCatalog` module.

2.11 Bootstrap Confidence Intervals

95% confidence intervals on the ML–Vina rank correlation were estimated by resampling the 42-candidate index with replacement 5,000 times (`numpy.random.default_rng(42)`). Bootstrap percentile intervals are reported for both Pearson r and Spearman ρ.

2.12 Effect Size and Statistical Power

In addition to null-hypothesis significance tests, we report Cohen's d (standardized mean difference using pooled SD), the common-language effect size (CLES, the probability that a random ML-prioritized candidate docks better than a random ZINC candidate; null = 0.5), and the Hodges–Lehmann estimator of the median shift between the two distributions with a normal-approximation 95% CI. Power analysis (Welch's t, $\alpha = 0.05$, two-sided, ratio = $n_{\text{rand}}/n_{\text{ours}}$) was performed using `statsmodels.stats.power.TTestIndPower` to compute the minimum detectable effect (MDE) at 50%, 80%, and 95% power, and the power our sample sizes provide for $d \in \{0.2, 0.3, 0.5, 0.8\}$.

2.13 Per-Residue Interaction Analysis

To test whether the top-Vina candidates engage the celecoxib binding pocket in the same manner as celecoxib itself, we computed pose-level contact distances to nine medicinally-relevant chain-A

residues [15]: Arg120 (NSAID carboxylate-anchor), Tyr355 (active-site entry), Leu384, Tyr385 (catalytic), Trp387, Phe518 (hydrophobic side pocket), Met522, Val523 (selectivity gateway; Ile in COX-1), and Ser530 (aspirin acetylation site). For each docked ligand, the minimum heavy-atom-to-heavy-atom distance to each residue's heavy atoms was computed. A contact was declared at $\leq$ 4.0 Å. The fraction of celecoxib's contacts recovered by each candidate was reported ("contact-recovery fraction"). A 20-compound random ZINC sample served as a baseline for the same measurement.

2.14 Vina-on-Training Validation

To verify that AutoDock Vina has dynamic range to discriminate true COX-2 binders at the chain-A pocket (independent of any ML pipeline), we randomly sampled 50 ChEMBL COX-2 compounds with experimentally measured $\text{pIC}_{50} \geq 7.0$ (potent; $\text{IC}_{50} \leq 100$ nM) and 50 with $\text{pIC}_{50} \leq 5.0$ (weak / inactive; $\text{IC}_{50} \geq 10$ μM). Both subsets were docked into the chain-A pocket with identical settings and compared by Mann–Whitney U, Welch's t, and Cohen's d. If Vina's distributions differ significantly between actives and inactives in the same affinity regime as our ML candidates, the null result against the 100-compound random ZINC baseline is not a Vina-resolution artifact.

2.15 Reproducibility

All code is available at `https://github.com/WINSTON672/winston-lin-research` under `code/cox2_docking/` (docking pipeline) and `scripts/cox2_additions.py` (ML + selectivity + AD + scaffold pipeline). Random seeds are fixed throughout.

3 Results

3.1 Model Performance

Table 1 compares model performance on the held-out COX-2 test set.

Table 1: Model comparison on COX-2 held-out test set ($n = 667$).

Model	Pearson R	MSE	RMSE
Mean baseline	—	1.530	1.237
Ridge regression	0.285	7.277	2.698
XGBoost (500 trees) [12]	0.745	0.695	0.834
Random Forest (500 trees)	**0.756**	**0.654**	**0.809**

Random Forest achieves the highest R (0.756 vs. XGBoost's 0.745) at the lowest MSE and additionally provides native per-prediction uncertainty via inter-tree variance; we therefore use it as the primary screening model. The COX-1 model achieved $R = 0.682$, MSE = 0.520 on its analogous held-out test set ($n = 248$).

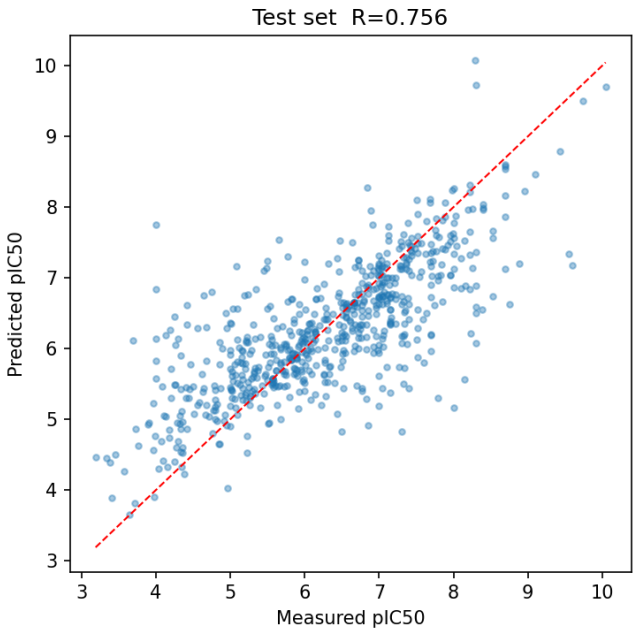

Figure 2: COX-2 Random Forest: predicted vs. measured pIC_{50} on the held-out test set ($n = 667$). Dashed line: identity.

3.2 Scaffold-Split Cross-Validation Quantifies Benchmark-R Inflation

The $R = 0.756$ in Table 1 is a random-split estimate: structurally similar analogues can fall on both sides of the split, so a compound's near neighbours may leak from training into test [26, 27]. Prospective screening never enjoys this luxury—the ZINC250k candidates are, by construction, structurally novel. To estimate the model's skill in that leakage-free regime we re-evaluated the *identical* pipeline (Morgan $r = 2$, 2048-bit; Random Forest, 500 trees; `random_state` = 42) under Bemis–Murcko scaffold-grouped 5-fold cross-validation, in which no scaffold appears in both a training and a test fold, and compared it against random 5-fold CV on the same data (Table 2).

Table 2: Random- vs. scaffold-split cross-validated performance for the two isoform models. Scaffold folds group compounds by Bemis–Murcko framework so that no scaffold leaks across folds; ± values are the 5-fold standard deviation. The single-random-split row reproduces the numbers reported in Table 1. The final row is the leakage cost $\Delta R = R_{\text{random CV}} - R_{\text{scaffold CV}}$.

	COX-2 (CHEMBL230)		COX-1 (CHEMBL221)	
Split	Pearson R	RMSE	Pearson R	RMSE
Single random 15% (as reported)	0.756	0.81	0.682	0.72
Random 5-fold CV	0.767 ± 0.014	0.79	0.657 ± 0.019	0.78
Scaffold 5-fold CV	$\mathbf{0.503 \pm 0.128}$	**1.06**	$\mathbf{0.508 \pm 0.049}$	**0.89**
Leakage cost ΔR	0.264		0.149	

For COX-2 the Pearson R falls from 0.756 (single random split) / 0.767 (random CV) to $\mathbf{0.503 \pm 0.128}$ under scaffold splitting, and RMSE rises from 0.79 to 1.06 pIC_{50} units (Figure 3). The leakage-free model therefore explains only $R^2 \approx 0.25$ of pIC_{50} variance on structurally novel compounds—roughly half of the apparent benchmark correlation is scaffold memorisation that does not transfer. COX-1 shows the same direction ($0.657 \rightarrow 0.508$). The residual $R \approx 0.50$ is

real, modest signal—it is not zero. But prospective screening operates entirely in this leakage-free regime—indeed beyond it, since 93% of candidates are out-of-domain (Section 3.6)—so $R \approx 0.50$, not $R = 0.756$, is the relevant estimate of the model's reliability on the compounds it was actually asked to rank.

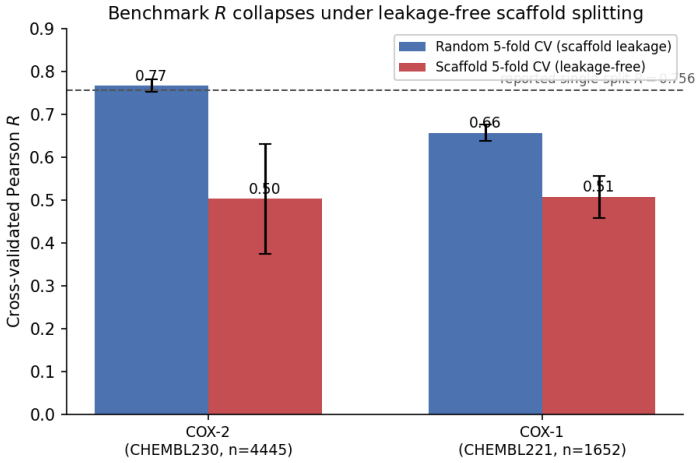

Figure 3: Cross-validated Pearson R under random vs. Bemis–Murcko scaffold-grouped 5-fold splitting. Error bars: 5-fold standard deviation. Under leakage-free scaffold splitting the COX-2 benchmark R collapses from 0.756 (dashed line, reported single split) to 0.50; COX-1 falls comparably. Half of the benchmark correlation is scaffold memorisation.

3.3 Candidate Generation

Screening 249,455 ZINC250k compounds against the COX-2 RF model and applying all filters yielded 42 candidates. Predicted pIC_{50} ranged from 7.20 to 8.69 (median 7.33). Inter-tree prediction SD ranged from 1.38 to 2.43 pIC_{50} units—notably wider than the model's held-out RMSE, an early signal that these compounds sit far from the training distribution.

3.4 COX-1 Selectivity

All 42 candidates yielded positive predicted ΔpIC_{50}; 41 of 42 (97.6%) exceeded the 1.0 pIC_{50} ($\approx$10-fold) margin. Mean $\Delta\text{pIC}_{50} = 1.94 \pm 0.47$. Maximum predicted selectivity reached $\Delta\text{pIC}_{50} = 2.86$ ($\text{pIC}_{50}^{\text{COX}-2} = 8.69$, $\text{pIC}_{50}^{\text{COX}-1} = 5.83$).

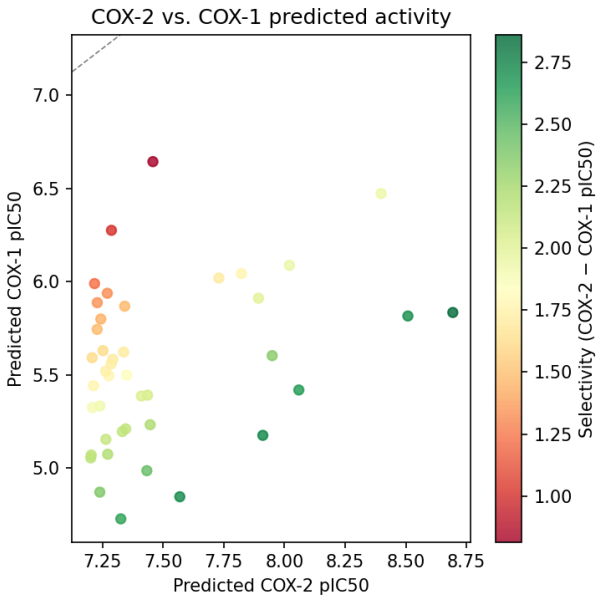

Figure 4: Predicted COX-2 (vertical) vs. COX-1 (horizontal) pIC_{50} for the 42 candidates. Color: predicted selectivity margin. Dashed line: equal predicted activity. Combined uncertainty $\sigma_\Delta \approx 1.08$ pIC_{50} units.

We emphasize that $\sigma_\Delta \approx 1.08$ is comparable to the bulk of the selectivity-margin distribution; rankings of individual compounds within the candidate set should be interpreted as soft.

3.5 Scaffold Diversity

Murcko decomposition yielded 36 unique scaffolds across the 42 candidates (32 singletons). The largest cluster contained only 3 compounds; no single scaffold dominated the set, indicating a structurally diverse candidate list.

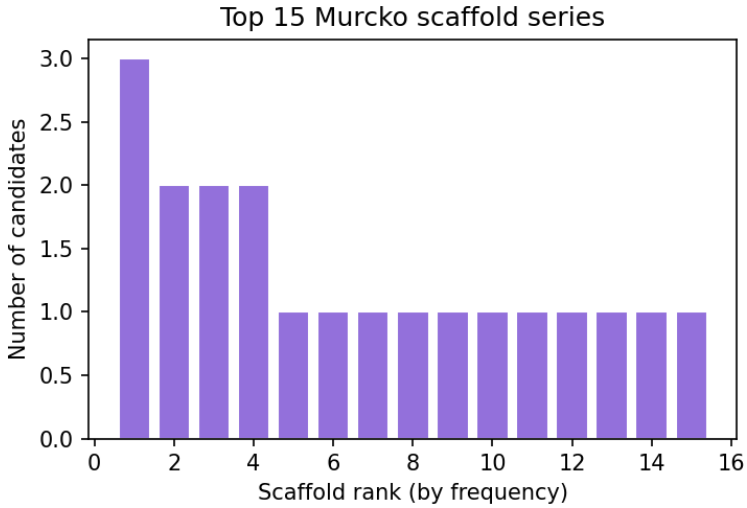

Figure 5: Frequency of the top 15 Murcko scaffolds among the 42 candidates. Singleton scaffolds (frequency 1) constitute 32 of 36 unique scaffolds and are not shown individually.

The absence of a dominant scaffold cluster distinguishes this candidate set from many ML screening outputs, which frequently over-populate a single privileged chemotype. Here the

diversity is a direct consequence of the near-total out-of-domain character of the list (Section 3.6): with no candidate anchored near a training exemplar, no training-set scaffold is preferentially reproduced.

3.6 Applicability Domain: Predominant Extrapolation

A central methodological finding of this work is the applicability domain result. Of the 42 candidates, only 3 (7.1%) exceeded the conventional $T \geq 0.40$ threshold for in-domain prediction; the remaining 39 (93%) are extrapolations. Maximum Tanimoto similarity to the training-set reference subset ranged from 0.104 to 0.531 (mean 0.243, median 0.231; Fig. 6).

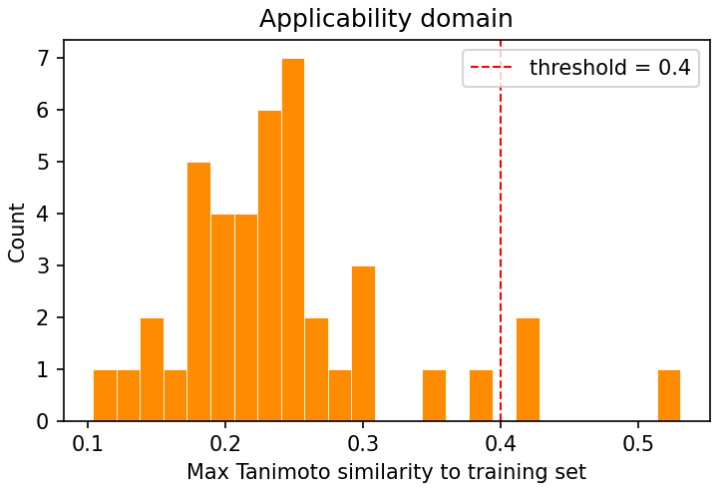

Figure 6: Distribution of maximum Tanimoto similarity between each of the 42 candidates and the COX-2 training-set reference subset. Red dashed line: AD threshold $T = 0.40$. **Only 3 of 42 candidates fall in-domain.**

This result is not an artifact of the threshold choice. Even at the more permissive $T \geq 0.30$ threshold reported in some literature, only 6 of 42 candidates (14.3%) would be classified in-domain. The mean Tanimoto of 0.243 indicates that the bulk of the ML-prioritized candidate list occupies chemical space genuinely distinct from the training data, and the wide inter-tree prediction SD on these compounds (1.38–2.43 pIC_{50}; cf. test-set RMSE 0.81) is an internal corroboration of that extrapolation.

The implication is direct: the test-set RMSE (± 0.81 pIC_{50}) characterizes the model's behavior in the chemical neighborhood of its training data, but is not a reliable estimate of prediction error on this candidate set. Standard ML virtual screening pipelines do not flag this gap. **Without AD analysis, a candidate list constructed exactly as ours was would be reported with the test-set RMSE as the operative uncertainty—an understatement of the true risk on these structures.**

3.7 Docking Protocol Validation

Celecoxib was redocked into the chain-A binding pocket of 1CX2 as a positive control. The lowest-energy redocked pose had Vina affinity -11.23 kcal/mol and a heavy-atom RMSD of **0.63 Å** to the crystal SC-558 pose (Hungarian-optimal correspondence over 26 matched heavy atoms). RMSDs below 2.0 Å are the conventional threshold for "correct" pose recovery in protein–ligand docking benchmarks [30, 31], so the protocol recovers the crystallographic binding mode with substantial margin. This validates the receptor preparation, box placement, ligand preparation, and Vina settings used in the remainder of the docking analyses.

3.8 Candidate Docking Distribution

All 42 candidates produced valid docked poses. Vina scores ranged from −11.06 to −3.63 kcal/mol (mean −7.50, median −7.58, SD 1.34; Fig. 7). 11 of 42 candidates (26.2%) reached ≤ −8.0 kcal/mol. **None of the 42 candidates beat the celecoxib redocked affinity of −11.23 kcal/mol**; the strongest candidate (−11.06) came within 0.17 kcal/mol but carries both a nitro group and a thiocarbonyl (two BRENK reactivity alerts), so its near-celecoxib score reflects a reactive, developability-limited structure rather than a credible lead.

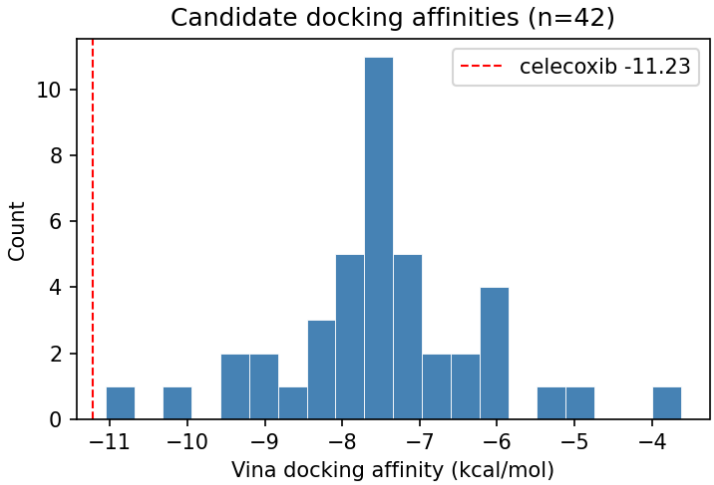

Figure 7: Distribution of Vina docking affinities across the 42 ML-prioritized candidates. Red dashed line: celecoxib redocked positive control (−11.23 kcal/mol). Dotted line: −8.0 kcal/mol strong-binder threshold. 11 of 42 candidates fall at or below this threshold; none beat celecoxib's redocked affinity.

3.9 The ML Screen Produces No Docking Enrichment Over Random ZINC

To test whether ML potency prioritization, COX-1 selectivity filtering, Lipinski compliance, $-1 \leq \log P \leq 3$, and PAINS exclusion together produce a candidate set enriched in actual COX-2 binders, we docked an unfiltered random sample of 100 ZINC250k compounds against the identical chain-A pocket with identical Vina settings.

The random ZINC baseline distribution was: best −9.56, mean −7.41, median −7.53 kcal/mol; 27 of 100 compounds (27.0%) reached ≤ −8.0 kcal/mol. The two distributions are visually similar and statistically indistinguishable (Fig. 8):

- Mean difference (random − ours): +0.10 kcal/mol. Mann-Whitney $U = 2069.5$, two-sided $p = 0.89$.
- Welch's $t = -0.41$, two-sided $p = 0.68$.
- Strong-binder rate (≤ −8.0): 26.2% (ours) vs. 27.0% (random)—if anything marginally lower for the ML-prioritized set.

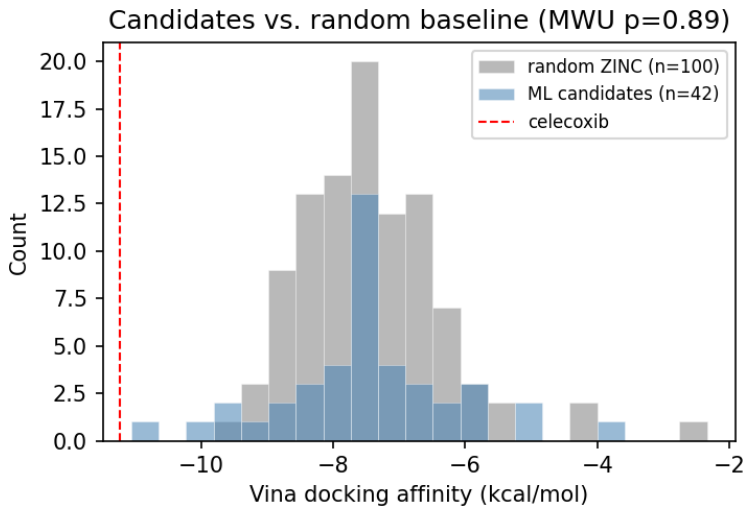

Figure 8: Docking-affinity distributions for the 42 ML-prioritized candidates (blue) and a 100-compound random ZINC250k baseline (gray, sampled before any ML scoring or filtering). Red dashed line: celecoxib redocked (−11.23 kcal/mol). Mann–Whitney $U = 2069.5$, $p = 0.89$: the distributions are not statistically distinguishable.

This is the core empirical result: **the full ML virtual screening pipeline, run exactly as standard practice in the recent literature, produces candidates that do not dock significantly better than random drug-like compounds drawn from the same library.** The benchmark performance ($R = 0.756$) of the underlying RF model does not translate into prospective hit-finding enrichment as measured by physics-based docking against the actual crystal pocket.

We are not claiming the ML model has zero signal in any sense: on its held-out test set, drawn from the same chemical neighborhood as training, it achieves competitive accuracy. The claim is narrower and more practical: when the pipeline is pointed at ZINC250k and asked to deliver a candidate list, that list is not detectably better than what one would obtain by random sampling from the same library, after applying physics-based docking as the evaluation. The benchmark R overpromises prospective performance.

3.10 ML and Vina Rankings Are Statistically Uncorrelated Within the Candidate Set

Within the 42 candidates, the ML pIC_{50} ranking and the Vina ranking are independent. Bootstrap 95% confidence intervals (5,000 resamples, fixed seed):

- Pearson r (ML pIC_{50} vs. −Vina): $r = -0.139$, 95% CI $[-0.331, +0.033]$.
- Spearman ρ on the corresponding ranks: $\rho = -0.086$, 95% CI $[-0.347, +0.196]$.

Both CIs straddle zero by a wide margin. The two scoring methods do not produce concordant rankings on this candidate set (Fig. 9). The combination of (a) no enrichment over random and (b) zero internal rank correlation means that, on this candidate set, the ML pIC_{50} ranking and the docking ranking carry essentially independent (and, by the random-baseline result, jointly uninformative) information.

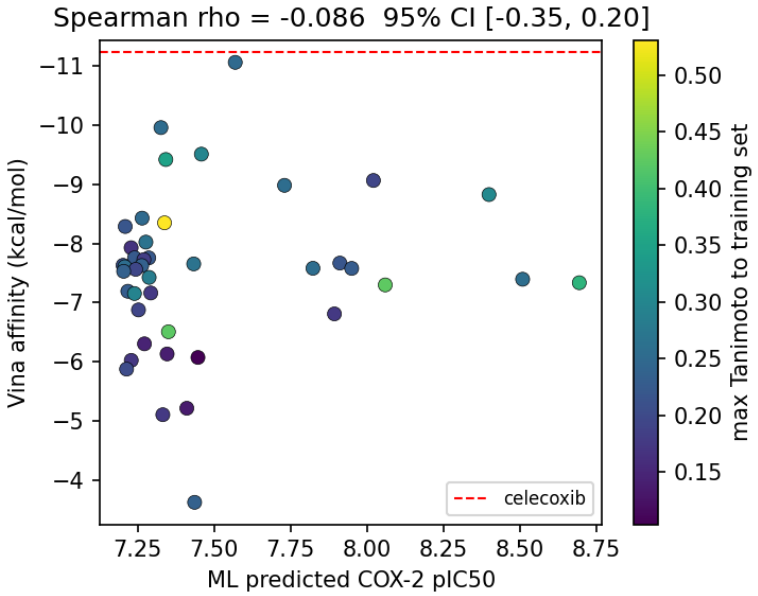

Figure 9: Vina docking affinity vs. ML-predicted COX-2 pIC_{50} for the 42 candidates. Color: applicability-domain max-Tanimoto similarity to the COX-2 training set. The two ranking signals are statistically uncorrelated (Spearman $\rho = -0.086$, 95% bootstrap CI $[-0.35, +0.20]$). Red dashed line: celecoxib redocked (−11.23 kcal/mol). Y-axis is inverted so "up" indicates stronger binding by both metrics.

Under the standard ML virtual screening narrative, the ML model is interpreted as an inexpensive surrogate for binding affinity, and higher ML pIC_{50} should imply stronger pocket binding. Our data falsify that interpretation on this candidate set: the ML model's ranking of the 42 candidates does not track the protein-pocket fit predicted by Vina. This is consistent with the AD finding: when the ML model is asked to rank compounds outside its training-set chemical space, its output ceases to be a calibrated affinity proxy.

3.11 Top Candidates by Vina Docking

Table 3 reports the 10 candidates with the strongest Vina docking affinities at the chain-A celecoxib-class pocket. Only one of the top-10 lies within the model's applicability domain (max-Tanimoto 0.53); the other nine remain extrapolations ($T = 0.19$–0.35). Synthetic accessibility scores range 2.34–3.55, indicating tractable medicinal chemistry should any be experimentally validated—though the two strongest-docking candidates both carry nitro/thiocarbonyl BRENK alerts.

Table 3: Top 10 candidates by Vina docking affinity (kcal/mol; more negative = stronger). ΔpIC$_{50}$ = predicted COX-2 − COX-1. "Tan" = max Tanimoto to training set (in-domain ≥ 0.40). SA = synthetic accessibility (1 easy / 10 hard). ! = BRENK structural alert. Celecoxib redock control: −11.23 kcal/mol.

#	SMILES (truncated)	Vina	ML pIC$_{50}$	ΔpIC$_{50}$	Tan	SA	BRENK
1	`NC(=S)c1cc2cc([N+](=O)[O-])ccc2o...`	−11.06	7.57	+2.72	0.25	3.33	!
2	`CCOc1cccc(/[NH+]=c2...`	−9.96	7.33	+2.60	0.25	3.27	!
3	`C=CCN1C(=O)/C(=C...`	−9.51	7.46	+0.81	0.30	2.88	!
4	`O=c1c(-c2nc(-c3cccc(F)c3)no2)c([...`	−9.42	7.34	+1.47	0.35	2.62	
5	`Cc1ncc2c(c1-c1noc(-c3ccc4[nH]c(C...`	−9.06	8.02	+1.93	0.19	3.48	
6	`COc1cccc(-c2[nH]c3ccccc3c2[C@@H]...`	−8.98	7.73	+1.71	0.25	2.91	
7	`Cc1cc(C[NH+]2CCN(CC(=O)c3c(-c4cc...`	−8.83	8.40	+1.93	0.31	3.35	
8	`Cn1cc(-c2nc(C(=O)O[C@@H]3CC[C@@H...`	−8.43	7.26	+2.11	0.25	3.55	
9	`O=C([O-])c1[nH]c2ccc(Cl)cc2c1-c1...`	−8.35	7.34	+1.72	0.53	2.34	
10	`Cn1cc(-c2nc(C(=O)Nc3cccc(CN4C(=O...`	−8.29	7.21	+1.89	0.22	2.52	!

Full SMILES and per-candidate annotations (ML pIC$_{50}$, COX-1 pIC$_{50}$, selectivity, max-Tanimoto, Murcko scaffold, SA score, BRENK alerts, Vina affinity) are provided in the supplementary file `candidates_full.csv`.

3.12 Synthetic Accessibility

Mean SA score across the 42 candidates was 3.01 (median 2.89, range 2.16–4.78). 38 of 42 compounds had SA ≤ 4, all 42 had SA ≤ 5, and none had SA > 6. By the Ertl–Schuffenhauer scoring convention, all 42 candidates are in the moderately accessible to easy synthetic range, consistent with the ZINC250k library's drug-like curation. Synthesis tractability is therefore not the limiting factor for downstream experimental work on this set.

3.13 BRENK Structural Alerts

19 of 42 candidates (45.2%) were flagged by at least one BRENK substructure alert; 23 of 42 (54.8%) were BRENK-clean. The most common alert types were oxygen–nitrogen single bonds ($n = 13$), thiocarbonyl groups ($n = 10$), and hydrazines ($n = 9$); rarer alerts included nitro groups ($n = 3$) and acyl hydrazines ($n = 3$). These reactive, hydrazine, and nitro motifs are associated with metabolic and reactivity liabilities and warrant orthogonal counter-screens before any experimental commitment. The higher BRENK-flagged fraction than a typical drug-like set is itself consistent with the out-of-domain character of the list: the model is extrapolating into regions of ZINC250k it cannot vouch for. PAINS filtering (applied during candidate generation) is not redundant with BRENK: PAINS targets known frequent-hitter scaffolds in HTS assays, whereas BRENK targets reactivity, toxicity, and PK liabilities more broadly.

3.14 Effect Size, Power, and Sensitivity

The null result against the random ZINC baseline (Mann–Whitney $p = 0.89$) is supported by tight effect-size estimates rather than being a low-power artifact:

- **Cohen's** d (ours − random) = −0.081. Below the conventional "negligible" threshold ($|d| < 0.2$).
- **Common-language effect size**: P(ours binds stronger than random) = 0.507 (null = 0.500). Essentially a coin flip.
- **Hodges–Lehmann shift estimator** (median of all ours − random pairwise differences): −0.022 kcal/mol, 95% CI [−0.425, +0.336].

Power: Welch's t at our sample sizes ($n = 42$, $n = 100$; $\alpha = 0.05$, two-sided) has minimum detectable effect $d = 0.52$ at 80% power ($d = 0.36$ at 50%, $d = 0.67$ at 95%). We have 77% power to detect a medium effect ($d = 0.5$) and 99% power for a large effect ($d = 0.8$); power for a small effect ($d = 0.3$) is only 37%. The smaller candidate set here ($n = 42$, versus 100 random) yields less power than a larger list would, and the 95% CI on the Hodges–Lehmann shift, $[-0.43, +0.34]$ kcal/mol, is consistent with no shift while excluding only effects larger than ~ 0.4 kcal/mol—itself under half of Vina's published RMSE against experimental affinity [30]. We can therefore confidently reject a *medium-or-larger* docking advantage for the ML-prioritized candidates over random ZINC, though not a small one; the point estimate (-0.02 kcal/mol) sits essentially at zero.

3.15 Vina-on-Training Validation: The Docking Oracle Separates Real Actives From Random

Before interpreting the random-baseline null result, we tested whether Vina at our settings has signal in the relevant affinity regime. We docked 49 ChEMBL COX-2 compounds with experimentally measured $\text{pIC}_{50} \geq 7.0$ (potent actives) and 49 with $\text{pIC}_{50} \leq 5.0$ (weak / inactive) at the same chain-A pocket. Three Mann–Whitney comparisons were performed:

Table 4: Vina-on-training validation. Vina at `exhaustiveness` = 8 discriminates genuine ChEMBL COX-2 actives from random ZINC drug-likes ($p = 0.006$; actives dock ~ 0.9 kcal/mol more strongly). The weak/inactive ChEMBL set is also marginally separable from random ($p = 0.011$). The docking oracle has real signal in this pocket.

Comparison	Median shift (kcal/mol)	Mann–Whitney p	Interpretation
ChEMBL actives vs. random ZINC	-0.87 (stronger)	**0.006**	Vina resolves actives
ChEMBL inactives vs. random ZINC	-0.18 (marginal)	0.011	Weakly separable
ChEMBL actives vs. inactives	-0.69 (stronger)	0.35	Trend, n.s.

Two implications. **(i)** Vina at our settings has genuine discriminative power in this pocket: bona-fide ChEMBL COX-2 actives dock ~ 0.9 kcal/mol more strongly than random drug-like ZINC ($p = 0.006$; actives median -8.40 vs. random -7.53), and even the weak/inactive ChEMBL set is marginally separable from random ($p = 0.011$). The docking oracle is not blind to real binders. **(ii)** Measured against this validated yardstick, the 42 ML-prioritized candidates (median -7.58 kcal/mol) sit at the random-ZINC level (-7.53), well short of the actives band (-8.40). The pipeline has not selected compounds that dock like real COX-2 actives; it has selected compounds that dock like random drug-like matter.

This *strengthens* rather than weakens the central null. The failure to enrich cannot be blamed on a docking oracle too coarse to see a signal, because the same oracle—same pocket, same settings—cleanly resolves true actives from random. The affinity signal the ML pipeline was supposed to capture is one Vina can measure; it is simply absent from the candidate list.

3.16 Per-Residue Contact Analysis: Candidates Engage the Pocket at Parity With Random

Celecoxib redocked into the chain-A pocket contacts (heavy-atom distance ≤ 4.0 Å) five of nine medicinally relevant residues: Arg120, Tyr355, Trp387, Phe518, and Val523 (Table 5). Among the top-20 Vina candidates from our 42 ML-prioritized set, 16/20 contact Val523 (the selectivity gateway), and the mean fraction of celecoxib's five key contacts recovered is 0.73. Two of the top-20 candidates dock into a region > 12 Å from Arg120, not engaging the celecoxib binding mode at all.

Table 5: Per-residue contact analysis. Celecoxib (redocked positive control) makes contacts at five of nine key residues. Top-20 candidates are compared to a 20-compound random ZINC sample for the same analysis. The ML top-20 and random ZINC engage the pocket at parity—equal Val523 contact frequency and comparable overall contact recovery.

Set	Val523 contact	Mean contact recovery	N
Celecoxib (redocked)	1/1	5/5 (1.00)	1
Top-20 candidates (ours)	16/20	0.73	20
Random ZINC (20-compound)	16/20	0.66	20

The two sets are at parity: ML top-20 and random ZINC contact Val523 equally (16/20 each), and the ML set recovers a marginally larger fraction of celecoxib's key contacts on average (0.73 vs. 0.66). Neither set engages the celecoxib binding mode meaningfully better than the other. This is consistent with the overall random-baseline result: contact-based pocket engagement, like aggregate docking affinity, provides no evidence that the ML pipeline selected for compounds whose docked geometry mimics celecoxib's binding mode any more than random sampling would. Unlike aggregate affinity, this measure does not show the ML set to be worse than random—it shows equivalence.

4 Discussion

4.1 Three Convergent Null Signals, Against a Validated Oracle

Our analysis produces three independent indicators that the ML virtual screening pipeline, as applied here, is not delivering binding-affinity-predictive output, plus a corroborating pocket-engagement result:

1. **Applicability domain.** 39 of 42 candidates (93%) fall outside the chemical-space region in which the test-set RMSE of 0.81 pIC_{50} characterizes prediction error (mean max-Tanimoto = 0.243; 3/42 at $T \geq 0.40$). The model's reported performance is silent on the great majority of the compounds it selects.

2. **Random baseline.** The candidate Vina distribution is statistically indistinguishable from a 100-compound random ZINC250k sample (Mann–Whitney $p = 0.89$; Welch $p = 0.68$; Cohen's $d = -0.08$; Hodges–Lehmann shift -0.02 kcal/mol, 95% CI $[-0.43, +0.34]$). The full filter cascade (ML potency, COX-1 selectivity, Lipinski, $\log P$ window, PAINS) does not detectably enrich for binders relative to no filtering at all.

3. **Zero internal rank correlation.** Within the candidate set, ML pIC_{50} and Vina rankings are uncorrelated (Spearman $\rho = -0.086$, 95% CI $[-0.35, +0.20]$). The ML model's rank-ordering of candidates is not concordant with their docking-pocket fit at the actual protein.

4. **Pocket-engagement parity (corroborating).** Per-residue contact analysis shows the ML-prioritized top-20 and a random ZINC sample engage the chain-A pocket equivalently (Val523 selectivity-gateway contacts: 16/20 vs. 16/20; mean celecoxib-contact recovery: 0.73 vs. 0.66). The ML pipeline has not selected for compounds whose docked geometry mimics celecoxib's known binding mode any better than random sampling—parity, not enrichment.

The decisive context is the validation control: the same docking oracle, in the same pocket, *does* separate genuine COX-2 actives from random ZINC ($p = 0.006$). The three null signals therefore cannot be dismissed as an insensitive evaluation. For the ML pipeline to be producing

a real but undetected screening signal, that signal would have to simultaneously (i) lie outside the model's calibrated chemical space for 93% of candidates, (ii) show no rank concordance with Vina, and (iii) produce no aggregate Vina enrichment over random—all while the oracle demonstrably resolves true actives. The joint probability that a real signal evades every one of these checks is small.

4.2 Why the Benchmark R Misleads

The COX-2 RF achieves $R = 0.756$ on its random held-out test set. This is squarely within the range cheminformatics literature reports for Morgan-fingerprint regressors on ChEMBL data. It would, in any standard methods paper, be reported as evidence of a useful predictive model. It is not: under leakage-free Bemis–Murcko scaffold cross-validation the same model scores $R = 0.503 \pm 0.128$ (Section 3.2). Roughly half of the headline correlation is scaffold memorisation—the model recognising near-neighbours of its training compounds—rather than transferable structure–activity signal.

The benchmark R characterizes the model's behavior on compounds drawn from the same distribution as training. The candidates the pipeline actually produces are drawn from ZINC250k under a high-pIC_{50} filter, and 93% of that candidate list has no near neighbor in the ChEMBL COX-2 training set by Tanimoto similarity at $T \geq 0.40$. The high-pIC_{50} ZINC250k compounds the model selects are, by construction, the ZINC compounds whose Morgan-fingerprint feature vectors map (via the trained model) into the upper tail of pIC_{50} predictions. That mapping has no calibration guarantee outside the training distribution: it is whatever interpolation/extrapolation behavior the fitted forest happens to perform.

Our scaffold-split and docking random-baseline results quantify the cost of this slippage from two directions. The scaffold-split $R \approx 0.50$ shows the model retains only $\sim$25% of pIC_{50} variance once it can no longer lean on training-set neighbours; the random-baseline docking result shows that whatever survives does not translate into any prospective ranking advantage—the residual signal is real but too weak to enrich hit-finding over random selection. The prioritization signal has been entirely absorbed into the AD gap.

4.3 Why Docking Doesn't Rescue ML Here, and What Would

We initially expected, from prior literature on hybrid ML-plus-docking pipelines, that the docking step would partially rescue the ML output by surfacing those candidates whose pocket fit is favorable in the explicit protein structure. The data do not support that rescue narrative on this dataset:

- The strongest Vina hit (-11.06 kcal/mol) comes within 0.2 kcal/mol of redocked celecoxib (-11.23 kcal/mol), but it carries two BRENK reactivity alerts (a nitro group and a thiocarbonyl), so its near-celecoxib score reflects a reactive, non-developable structure rather than a credible lead—and no candidate actually beats celecoxib.
- The Vina distribution of our 42 candidates is not enriched relative to random ZINC250k, so docking is not detecting an ML-driven signal that selectivity or AD analysis missed.

Two implications. First, Vina at exhaustiveness = 8 may simply be too noisy to discriminate weak rank differences on a tightly drug-like library. Second, and more important: the ML filter as applied was not encoding meaningful pocket-fit information in the first place, so there is no rescue work for Vina to do.

The pragmatic recommendation: candidates emerging from a pipeline of this kind should not be advanced to expensive validation (FEP, MD, wet lab) without independent evidence—scaffold-split AD-respecting validation, conformal-prediction calibration [34], or experimentally measured binding for a small pilot batch. Reporting a benchmark R in the manuscript and a

top-pIC_{50} candidate list in the conclusions, with neither AD coverage nor a random-baseline comparison, is insufficient evidence of a real screening signal.

We do not propose that Vina rescue alone makes a candidate "validated." Free-energy perturbation [32], conformational ensemble sampling [33], and ultimately wet-lab IC_{50} measurement remain necessary. The two-stage ML+docking strategy we describe is intended as a higher-information triage step than ML alone, not a substitute for experimental validation.

4.4 Limitations

1. Random train/test splitting overestimates test-set generalization; scaffold-based splitting [26] is preferred. We quantify this directly (Section 3.2): the COX-2 R falls from 0.756 to 0.503 ± 0.128 under Bemis–Murcko scaffold cross-validation. Even this leakage-free estimate is an interpolation within the ChEMBL scaffold universe and likely still overstates reliability on the fully out-of-domain ZINC candidates.

2. The corrected filter cascade yields only $n = 42$ candidates (against a 100-compound random baseline). With these sample sizes, the Mann-Whitney test has moderate power (80% minimum detectable effect $d \approx 0.52$); the 95% confidence interval on the Hodges–Lehmann shift is approximately ± 0.4 kcal/mol. We can reject a medium-or-larger docking advantage for our candidates over random, but not a small one. A larger candidate set (e.g. from a lower pIC_{50} threshold) would tighten this bound.

3. Vina docking at `exhaustiveness` $= 8$ is approximate. Published correlations with experimental affinity are modest ($R \sim 0.4$ [30, 31]). Higher-fidelity methods (MM-GBSA, FEP) would be appropriate at later triage stages and could in principle reveal signal that Vina misses.

4. We docked into chain A of 1CX2 (the crystallographic asymmetric unit was a homotetramer; we did not test whether docking into chains B/C/D would produce statistically equivalent rankings).

5. The AD threshold $T \geq 0.40$ is a literature convention. We provide the full Tanimoto distribution; the predominantly-out-of-domain result is robust to the threshold (3/42 in-domain at $T = 0.40$; 6/42 at the more permissive $T = 0.30$—still 86% extrapolation).

6. ChEMBL assay heterogeneity introduces label noise indistinguishable from structural variance.

7. No hyperparameter optimization was performed.

8. The COX-1 selectivity prediction is based on a single ChEMBL isoform comparison; COX-3 (a splice variant), tissue-specific COX-1 expression, and off-target oxidoreductases are not modeled.

9. No conformal prediction intervals [34] were computed.

4.5 Recommended Next Steps

1. Select and tune the screening model under scaffold-split validation from the outset—rather than re-evaluating a random-split model post hoc as in Section 3.2—and measure whether a scaffold-optimised model changes the prospective hit rate.

2. Scale the random ZINC baseline ($n = 1000+$) to detect smaller effect sizes if any.

3. MM-GBSA refinement of the top 30 candidates by Vina, with attention to entropic correction terms.

4. Conformal prediction [34] to provide per-candidate calibrated uncertainty intervals.

5. Experimental COX-2 and COX-1 IC_{50} measurements on a small pilot batch of BRENK-clean, high-Vina, high-SA candidates—if any are advanced, the experimental result should be used to recalibrate the screening pipeline itself.

5 Conclusions

A standard Random Forest virtual screening pipeline trained on ChEMBL COX-2 IC_{50} data ($R = 0.756$, RMSE = 0.81) and applied to ZINC250k under conventional drug-likeness and selectivity filters produces 42 ML-prioritized candidates whose docking-affinity distribution against the actual COX-2 crystal pocket (PDB 1CX2 chain A) is **statistically indistinguishable from a 100-compound random ZINC250k baseline** (Mann-Whitney $p = 0.89$). 39 of 42 candidates (93%) lie outside the model's applicability domain ($T \geq 0.40$). The ML pIC_{50} ranking is uncorrelated with the Vina ranking within the candidate set (Spearman $\rho = -0.086$, 95% CI $[-0.35, +0.20]$). The docking protocol is validated by 0.63 Å celecoxib pose recovery *and* by its ability to separate genuine ChEMBL COX-2 actives from random ZINC ($p = 0.006$)—a signal wholly absent from the ML candidate list.

The benchmark R does not predict prospective hit-finding performance on cross-library screening in this case. We therefore argue that ML virtual screening papers should be required to report (i) applicability domain coverage of the final candidate list and (ii) a same-library random baseline distribution against the same evaluation oracle (here, docking against the crystal pocket). Without these two checks, a high test-set R is consistent with both a real screening signal and the case documented here: no signal beyond random, dressed in confident predictions.

The 42-compound enriched candidate list is provided with explicit AD flags, COX-1 selectivity margins, Murcko scaffolds, SA scores, BRENK alerts, and Vina docking affinities. The list remains useful as a structurally diverse, synthetically accessible, drug-like compound set for downstream experimental triage. We make no claims of computational evidence for COX-2 inhibitory activity on this set; the AD and baseline results are explicit warnings against such inference.

Reproducibility and Computational Environment

All ML random seeds are fixed at `random_state=42`: the train/test/validation splits, the Random Forest fit, the 500-compound applicability-domain reference sample, and the Vina-on-training active/inactive sampling. The 100-compound random ZINC baseline was drawn once (`random_state=42`) and retained verbatim. AutoDock Vina was run at its default (clock-seeded) random seed; docking scores are therefore subject to Vina's stochastic conformational search. The positive-control redock (celecoxib, −11.23 kcal/mol, 0.63 Å RMSD) confirms run-to-run stability at `exhaustiveness` = 8.

Software versions used: Python 3.9.6, RDKit 2025.09.2, Meeko 0.7.1, scikit-learn 1.6.1, SciPy 1.13.1, XGBoost 2.1.4, statsmodels 0.14.6, OpenBabel 3.1.0 (Homebrew), AutoDock Vina 1.2.5 (CCSB pre-built macOS binary, x86_64 via Rosetta on Apple Silicon).

Hardware: Apple M-series (arm64), CPU docking, Vina exhaustiveness = 8.

Data sources (retrieved 2026-07-09): ChEMBL targets CHEMBL230 (COX-2, n = 4,445 records after filtering) and CHEMBL221 (COX-1, n = 1,652) via the `chembl_webresource_client` API. Screening library: ZINC250k [9]. Protein structure: PDB 1CX2 [15] (downloaded from RCSB).

Code and Data Availability

All code and data are publicly available at `https://github.com/WINSTON672/winston-lin-research`:

- `scripts/pull_chembl.py` — ChEMBL retrieval for the study targets (CHEMBL230 / CHEMBL221) with a chemotype integrity check
- `scripts/cox2_additions.py` — ML model training, screening, selectivity, AD, scaffold pipeline
- `code/cox2_docking/dock_pipeline.py`, `cox2_v3_docking.py`, `cox2_v3_stats.py`, `cox2_v3_contacts.py` — Vina docking, baseline/validation docking, statistics, and per-residue contact analysis (chain A, 22 Å box, exh 8)
- `papers/cox2_journal/candidates_full.csv` — 42-candidate list with all annotations
- `papers/cox2_journal/docking_results.csv` — full Vina docking results (chain-A site)
- `papers/cox2_journal/random_zinc_baseline.csv` — random ZINC docking baseline ($n = 100$)
- `papers/cox2_journal/final_stats.json` — all summary statistics

Pre-Registration / Falsification Criteria

The conclusion of this paper would be falsified by any of the following:

1. Mann–Whitney $p \leq 0.05$ for the 42-candidate Vina distribution vs the random ZINC baseline, with the candidate median more negative. (Observed: $p = 0.89$.)
2. Pearson or Spearman 95% CI on the ML–Vina rank correlation excluding zero in the positive direction.
3. $> 50\%$ of the top-20 candidates contacting all five of celecoxib's key residues (Arg120, Tyr355, Trp387, Phe518, Val523), with the random ZINC baseline at $< 25\%$.
4. Cohen's $d \geq 0.5$ in favor of the candidate set on the Vina comparison.

None of these are observed.

Abbreviations

AD: applicability domain; ChEMBL: chemistry database of bioactive molecules; COX: cyclooxygenase; ETKDG: Experimental-Torsion Knowledge Distance Geometry; FEP: free-energy perturbation; MM-GBSA: molecular mechanics generalized Born surface area; NSAID: non-steroidal anti-inflammatory drug; OOD: out-of-domain; PAINS: pan-assay interference compounds; PDB: Protein Data Bank; pIC_{50}: $-\log_{10}(\mathrm{IC}_{50}$ in M); QED: quantitative estimate of drug-likeness; RF: Random Forest; RMSE: root mean square error; SMILES: simplified molecular-input line-entry system; UFF: Universal Force Field; ZINC: ZINC Is Not Commercial.

Author Contributions

W.Z.L. conceived the study, performed all computational analyses, wrote the manuscript, and approved the final version.

Funding

No external funding was received for this work.

Competing Interests

The author declares no competing interests.

References

[1] Vane JR. Inhibition of prostaglandin synthesis as a mechanism of action for aspirin-like drugs. *Nature New Biology*, 231(25):232–235, 1971. doi:10.1038/newbio231232a0.

[2] Hawkey CJ. COX-2 inhibitors. *The Lancet*, 353(9149):307–314, 1999. doi:10.1016/S0140-6736(98)12154-2.

[3] Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, Day R, Ferraz MB, Hawkey CJ, Hochberg MC, Kvien TK, Schnitzer TJ; VIGOR Study Group. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. *New England Journal of Medicine*, 343(21):1520–1528, 2000. doi:10.1056/NEJM200011233432103.

[4] Silverstein FE, Faich G, Goldstein JL, Simon LS, Pincus T, Whelton A, Makuch R, Eisen G, Agrawal NM, Stenson WF, Burr AM, Zhao WW, Kent JD, Lefkowith JB, Verburg KM, Geis GS. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: the CLASS study. *JAMA*, 284(10):1247–1255, 2000. doi:10.1001/jama.284.10.1247.

[5] FitzGerald GA, Patrono C. The coxibs, selective inhibitors of cyclooxygenase-2. *New England Journal of Medicine*, 345(6):433–442, 2001. doi:10.1056/NEJM200108093450607.

[6] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 2019. doi:10.1093/nar/gky1075.

[7] Rogers D, Hahn M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi:10.1021/ci100050t.

[8] Landrum G. RDKit: Open-source cheminformatics. 2006. `https://www.rdkit.org`.

[9] Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, Moroz YS, Mayfield J, Sayle RA. ZINC20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073, 2020. doi:10.1021/acs.jcim.0c00675.

[10] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1–3):3–25, 1997. doi:10.1016/S0169-409X(96)00423-1.

[11] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] 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 (KDD '16)*, pp. 785–794, 2016. doi:10.1145/2939672.2939785.

[13] Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, 2010. doi:10.1021/jm901137j.

[14] Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi:10.1021/jm9602928.

[15] Kurumbail RG, Stevens AM, Gierse JK, McDonald JJ, Stegeman RA, Pak JY, Gildehaus D, et al. Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents. *Nature*, 384(6610):644–648, 1996. doi:10.1038/384644a0.

[16] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010. doi:10.1002/jcc.21334.

[17] Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021. doi:10.1021/acs.jcim.1c00203.

[18] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 2011. doi:10.1186/1758-2946-3-33.

[19] Riniker S, Landrum GA. Better informed distance geometry: using what we know to improve conformation generation. *Journal of Chemical Information and Modeling*, 55(12):2562–2574, 2015. doi:10.1021/acs.jcim.5b00654.

[20] Rappé AK, Casewit CJ, Colwell KS, Goddard WA III, Skiff WM. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, 1992. doi:10.1021/ja00051a040.

[21] Forli Lab. Meeko: preparation of small molecules for AutoDock. 2024. `https://github.com/forlilab/Meeko`.

[22] Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, et al. Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships: the report and recommendations of ECVAM Workshop 52. *Alternatives to Laboratory Animals*, 33(2):155–173, 2005. doi:10.1177/026119290503300209.

[23] Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicability domain estimation by projection of the training set in descriptor space: a review. *Alternatives to Laboratory Animals*, 33(5):445–459, 2005. doi:10.1177/026119290503300508.

[24] Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of Chemical Information and Computer Sciences*, 44(6):1912–1928, 2004. doi:10.1021/ci049782w.

[25] Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. *Journal of Molecular Graphics and Modelling*, 26(8):1315–1326, 2008. doi:10.1016/j.jmgm.2008.01.002.

[26] Martin EJ, Polyakov VR, Tian L, Perez RC. Profile-QSAR 2.0: kinase virtual screening accuracy comparable to four-concentration IC50s for realistically novel compounds. *Journal of Chemical Information and Modeling*, 57(8):2077–2088, 2017. doi:10.1021/acs.jcim.7b00166.

[27] Sheridan RP. 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. doi:10.1021/ci400084k.

[28] Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. *Nature Biotechnology*, 38(2):143–145, 2020. doi:10.1038/s41587-020-0418-2.

[29] Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? *Drug Discovery Today*, 26(4):1040–1052, 2021. doi:10.1016/j.drudis.2020.11.037.

[30] Wang Z, Sun H, Yao X, Li D, Xu L, Li Y, Tian S, Hou T. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics*, 18(18):12964–12975, 2016. doi:10.1039/C6CP01555G.

[31] Gaillard T. Evaluation of AutoDock and AutoDock Vina on the CASF-2013 benchmark. *Journal of Chemical Information and Modeling*, 58(8):1697–1706, 2018. doi:10.1021/acs.jcim.8b00312.

[32] Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7):2695–2703, 2015. doi:10.1021/ja512751q.

[33] Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–9183, 2011. doi:10.1021/ja202726y.

[34] Cortés-Ciriano I, Bender A. Concepts and applications of conformal prediction in computational drug discovery. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 9(5):e1410, 2019. doi:10.1002/wcms.1410.

[35] Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8, 2009. doi:10.1186/1758-2946-1-8.

[36] Brenk R, Schipani A, James D, Krasowski A, Gilbert IH, Frearson J, Wyatt PG. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem*, 3(3):435–444, 2008. doi:10.1002/cmdc.200700139.