Real Measured Negatives as a Substrate for Calibrated, Target-Specific Bioactivity Prediction

A validation of the Nullary negative-results data layer

Nullary Team

nullary.ai · Technical Report · 24 May 2026

Abstract

Public bioactivity resources are biased toward positive findings: compounds that *work* are published; the far larger set of measured failures is mostly discarded. Structure- and ligand-based virtual-screening models therefore train against *computational decoys*—molecules merely assumed inactive—a documented source of inflated benchmark scores and poor prospective generalization. Nullary aggregates experimentally measured negative results across modalities into a single queryable layer (122.3M findings). We report a validation study of whether these real negatives support predictive, well-calibrated activity models. Using ChEMBL single-protein human targets and Nullary's own definition of an inactive, we train per-target classifiers (ECFP4 + gradient boosting) and evaluate them under Bemis–Murcko *scaffold-disjoint* splits. On 25 well-studied kinases, median ROC-AUC is 0.966 (0.950–0.973); the random-split median is only marginally higher (difference of medians 0.006; median of per-target paired differences 0.010; both small, the paired difference statistically significant). We are deliberately careful here: a small random-vs-scaffold gap does *not* by itself establish prospective generalization. In a controlled comparison (Sec 4.5), real-negative training was directionally better than decoy training on all 25 kinases (median paired advantage +0.031 ROC-AUC, Wilcoxon $p<10^{-6}$), and decoy-trained models were markedly overoptimistic when validated on decoys; the per-target advantage cleared our *a priori* 0.03 bar on only 52% of targets — below the 60% threshold we set in advance (and statistically indistinguishable from it at this sample size). Critically, under a *temporal* split (Sec 4.6; train on compounds first characterized ≤ 2018, test on ≥ 2020) — the honest prospective-predictivity proxy — the median falls to 0.775 (from 0.966 scaffold on the same targets, a 0.18-point drop): the scaffold-split headline does not hold prospectively, though performance stays well above chance. One split-validity control (AVE-debiasing) remains future work. Scaled to a registry of 398 kinase and GPCR targets, the median scaffold ROC-AUC is 0.926 (PR-AUC reported against its inactive-prevalence baseline; median Brier 0.081). We also describe a target-exhaustion index summarizing how heavily a target has been queried. We treat this as *preliminary* validation: the scaffold headline is consistent with a competent ECFP4+gradient-boosting model on an optimistic split, and the temporal split is the realistic forward-looking number. Models and indices are reachable through Nullary's MCP and REST interfaces.

1. Introduction

Most machine-learning models of small-molecule bioactivity are trained where the data is: on measured *actives*. Measured *inactives* are comparatively scarce, because negative outcomes are under-reported. To obtain a negative class, the standard practice constructs decoys—property-matched molecules presumed inactive (e.g. DUD-E [2]). Multiple studies have since shown that models can separate such decoys from actives using superficial property differences, producing optimistic retrospective metrics that do not transfer to prospective screening [3,4]. Real, experimentally measured negatives are the missing ingredient. Per-target classifiers trained on measured ChEMBL/PubChem inactives are not new—PIDGIN [11] has done this for years, combining measured ChEMBL/PubChem inactives with sphere-exclusion augmentation—but aggregating measured negatives across modalities at this scale, and testing whether they validate honestly, is the question here.

Nullary is a data layer that aggregates measured negative results—inactive compound–target assays, failed genetic screens, terminated trials, developability failures—and serves them through programmatic interfaces. This report does not argue a product; it asks a falsifiable question: *do the real negatives in this layer carry predictive, well-calibrated signal under an honest evaluation?* We answer it for two target families and describe the resulting model registry and a companion target-exhaustion index.

2. The data layer

The layer holds 122,276,636 negative findings over 1,621,294 distinct compounds, normalized into a common schema keyed on *modality* and linked to provenance. Sources span curated bioactivity (ChEMBL), high-throughput screening (PubChem BioAssay), functional-genomics knockout screens (DepMap, BioGRID-ORCS), clinical-trial registries (ClinicalTrials.gov/AACT, EU-CTR), regulatory and safety records, and biologics developability sets. Table 1 summarizes coverage. We are explicit that this 122.3M headline is *repository* scale, not training-ready data: ~39.1M small-molecule findings carry both a structure and a UniProt target (so ~54% lack a clear target and are unusable for per-target modelling), and the experiments below draw on a far smaller curated, target-annotated slice—the 25-kinase proof uses ~148k compound–target pairs. We report the repository size and the modelling size separately and do not conflate them.

Modality	Findings	Principal sources
small molecule	84.5M	PubChem, ChEMBL
CRISPR / functional genomics	37.6M	BioGRID-ORCS, DepMap
clinical trial	104k	AACT, EU-CTR, drugs@FDA
peptide / antibody / PROTAC / other	~51k	THPdb, TheraSAbDab, PROTAC-DB, …

Table 1. Coverage of the negative-results layer by modality (snapshot, 24 May 2026). Counts are exact as of the last bulk load.

A negative is *context-conditional*: “inactive at 10µM in assay Z” is a statement about a measurement, not an intrinsic property of a molecule. The layer preserves the measurement context (assay type, value, units, threshold) so downstream consumers can define labels consistently. Sources are partitioned into a *curated* tier (dose-response or human-curated; e.g. ChEMBL, trial terminations) and a higher-volume, noisier *screening* tier (single-dose HTS, baseline non-essential genes). This study uses only the curated tier.

3. Methods

3.1 Cohort and labels

We select ChEMBL 35 single-protein human targets classified as protein kinases or GPCRs, with dose-type measurements (IC50, Ki, Kd, EC50, potency) in molar units. Labels follow Nullary's production definition: a measurement is **inactive** if value ≥ 10µM (or the activity comment denotes inactivity) and **active** if value ≤ 1µM. The 1–10µM band is discarded to reduce label noise. Measurements are aggregated to one label per (compound, target) pair; pairs with both active and inactive evidence are dropped as conflicting—a small fraction, 1.2–1.4% of labelled pairs. The inactives so defined are exactly the rows the layer serves; the actives are their dose-response complement from the same release. Two consequences of this scheme are revisited in Sec 5: discarding the 1–10µM band inflates class separability, and cross-laboratory IC50 reproducibility is only ~0.68 log units [16], so the one-log active/inactive boundary sits at roughly 1.5 SD of measurement noise.

3.2 Representation and model

Compounds are encoded as 2048-bit ECFP4 (Morgan radius 2) fingerprints [5]. We fit one gradient-boosted tree classifier (LightGBM [6]) per target, predicting P(inactive). Per-target models avoid conflating target identity into a single pooled model and match the way the layer is queried in practice (a known target, a candidate compound).

3.3 Evaluation

Honest generalization is estimated with a *scaffold-disjoint* split: compounds are grouped by Bemis–Murcko scaffold [7] and whole scaffolds are assigned to train or test (80/20), so no test chemotype is seen in training [8]. We report ROC-AUC, PR-AUC (positive = inactive) against its prevalence baseline, and the Brier score, and we test the random-vs-scaffold difference with a paired Wilcoxon signed-rank. We caution up front that scaffold-disjoint splitting is a *weak* generalization test: closely related analogue series fall into different Bemis–Murcko scaffolds yet encode near-identical pharmacophores at the ECFP4 level [4,13], so a small random-vs-scaffold gap is necessary but not sufficient for prospective accuracy. The stronger controls—temporal split [14,17] and AVE-debiasing [12]—are not performed here (Sec 5).

3.4 Calibration and the served registry

For deployment, each target's model is recalibrated (isotonic with ≥500 examples per class, else Platt/sigmoid) and retrained on all of that target's labelled compounds. We note this 500-point threshold is below the ~1,000-point crossover at which Niculescu-Mizil & Caruana [9] found isotonic regression reliably matches or beats Platt scaling; at 500/class isotonic can overfit the calibration set, so Platt is the safer default and we flag this for revision (Sec 5). Calibration is summarized here by the Brier score only—reliability diagrams and expected calibration error are not yet reported. The registry stores, per target, the model and a card with the scaffold-split metrics above.

3.5 Target-exhaustion index

Independently of the models, we aggregate all negative findings per target into a target-exhaustion ("graveyard") index: the number of distinct compounds tried and failed, broken down by modality and outcome. This is a measurement-*coverage* metric—how heavily a target has been queried—not a tractability score: raw compound counts confound how intensely a target has been prosecuted with how intractable it is, and the index is not calibrated against external ground truth (clinical attrition, or tractability resources such as Open Targets and Pharos). Comparable per-target counts already exist in those resources; our contribution is restricting to the measured-negative slice.

4. Results

4.1 Proof of signal on kinases

On 25 well-studied kinases (148,851 pairs; 66,518 inactive, 82,333 active), the per-target scaffold-split ROC-AUC has median 0.966 (0.950–0.973) and PR-AUC 0.917 (0.901–0.947) (Table 2). The random-split median is 0.972 (Fig. 3). The random–scaffold gap is small (difference of medians 0.006 in Table 2; median of per-target paired differences 0.010) yet statistically significant (paired Wilcoxon $p < 0.0001$): scaffold-disjoint evaluation is measurably—if mildly—harder. A gap this small shows the model is not merely memorizing exact scaffolds, but it does *not* on its own establish prospective generalization (Sec 5). PR-AUC should be read against the inactive-prevalence baseline of 0.45, so 0.92 is well above chance rather than near-perfect. Were the negatives noise, scaffold-split AUC would collapse toward 0.5; it does not.

Split	ROC-AUC	PR-AUC	Brier
Scaffold-disjoint (honest)	0.966 (0.950–0.973)	0.917 (0.901–0.947)	0.060 (0.044–0.081)
Random compound	0.972 (0.964–0.982)	—	—

Table 2. Kinase proof (25 targets). Median (inter-quartile range) of per-target metrics. Positive class = inactive.

4.2 A kinase + GPCR registry

Applying the same recipe across both families yields a registry of 398 calibrated per-target models (249 kinase, 149 GPCR). The median scaffold-split ROC-AUC is 0.926; 254 of 398 targets exceed 0.90 and 348 exceed 0.80 (Fig. 1, Table 3). GPCR models score higher on average than kinase models, partly because the kinase set extends further into small, data-poor targets. Performance rises with the amount of labelled data and the low-AUC tail is concentrated in data-poor targets (Fig. 2)—an expected, and honest, dependence.

Family	Models	Scaffold ROC-AUC	Brier
Kinase	249	0.913 (0.835–0.956)	0.097
GPCR	149	0.955 (0.899–0.979)	0.061
All	398	0.926 (0.855–0.968)	0.081

Table 3. Registry summary. Median (IQR) per-target scaffold-split ROC-AUC and median Brier score.

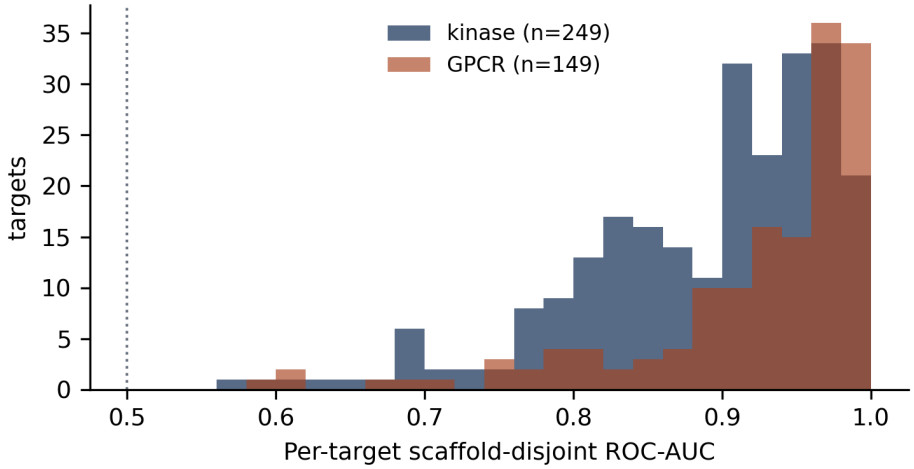

Figure 1. Distribution of per-target scaffold-disjoint ROC-AUC across the 398 registry models, by family. Dotted line marks chance (0.5).

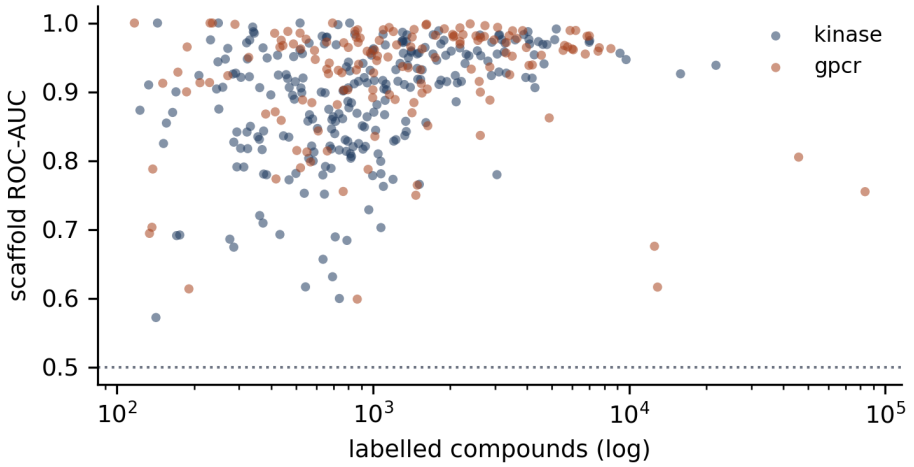

Figure 2. Per-target scaffold ROC-AUC versus the number of labelled compounds (log scale). Accuracy increases with data; the weak tail is data-poor.

4.3 Calibration and example predictions

Median Brier scores (0.06–0.10) summarize calibration, with the caveats in Sec 5 (Brier only; no reliability diagrams or ECE yet). Table 4 is a *face-validity* check only: these scores come from the served model, which is trained on all of a target's labelled compounds, so the listed canonical drugs are in the training set. The table shows scores are not inverted—it is *not* a held-out test. The held-out evidence is Tables 2–3.

Compound	Target	P(inactive)	Verdict
Gefitinib	EGFR	0.009	active
Erlotinib	EGFR	0.004	active
Caffeine	EGFR	0.913	inactive
Haloperidol	DRD2	0.001	active
Aspirin	DRD2	0.995	inactive

Table 4. Face-validity scores from the served (all-data) model—the listed drugs are in training, so this checks sanity, not generalization. Verdict thresholds: active ≤ 0.34, inactive ≥ 0.66.

4.4 Target-exhaustion index

For EGFR, the index records 5,953 distinct compounds tried across 12,585 negative findings (small molecules, PROTACs and peptides), of which 8,276 derive from curated sources. Such summaries quantify how picked-over a target is before a program commits to it.

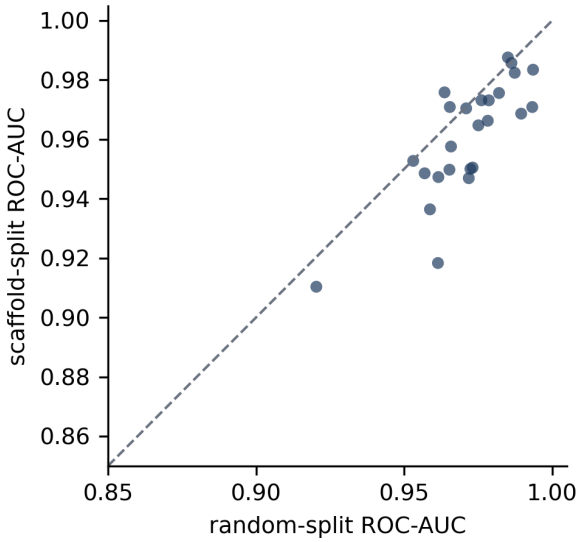

Figure 3. Scaffold- versus random-split ROC-AUC for the 25 kinase proof targets. Points lie near, but mostly below, the diagonal: a small yet statistically significant drop on unseen scaffolds (paired Wilcoxon $p < 0.0001$).

4.5 Real negatives versus decoys

The motivating contrast—measured negatives versus computational decoys—is testable directly. For the 25-kinase cohort, holding architecture and scaffold split fixed, we swap only the negative training class: (R) the real measured inactives, or (D) property-matched, topologically-dissimilar decoys (DUD-E style: matched on MW / logP / HBD / HBA / rotatable bonds to the actives, ECFP4 Tanimoto < 0.35 to any active, drawn from ChEMBL background), at a 1:1 ratio to the real inactives (not DUD-E's 50:1, so the two conditions have equal-sized negative classes). Both models are judged on the *same* held-out real test (actives vs real inactives, scaffold-disjoint). We set an *a priori* decision rule (specified before the run, not publicly deposited): the 'real negatives' framing is supported only if R beats D by ≥ 0.03 ROC-AUC on ≥ 60% of targets. Caveat: decoys were filtered for dissimilarity to the actives but not to the held-out real inactives, and the Tanimoto < 0.35 filter makes our decoys *harder* than stock DUD-E (which does not strictly enforce it) — likely why our decoy-validation inflation below (~0.055) is smaller than the 0.10–0.15 often reported.

Training negatives	Test set	Median ROC-AUC
Real inactives (R)	real inactives	0.962
Decoys (D)	real inactives	0.932
Decoys (D)	decoys	0.987

Table 5. Real negatives vs decoys, 25 kinases. Real-negative training wins on the realistic test for all 25 targets (median +0.031); the decoy model is overoptimistic when validated on decoys (0.987) rather than real negatives (0.932).

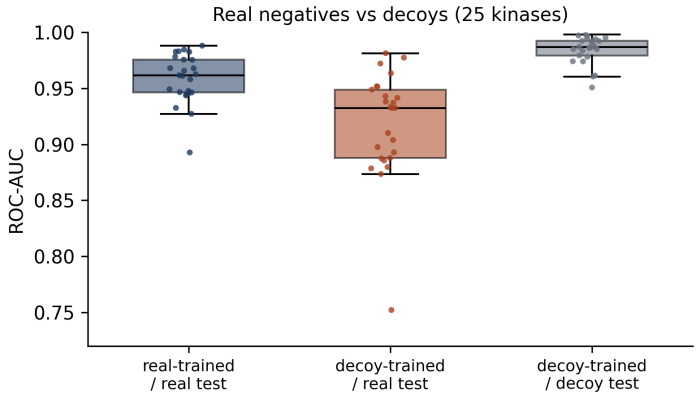

Figure 4. ROC-AUC by training-negative and test set. The decoy model scores highest on decoys (the illusion) yet lowest on the realistic real-negative test.

Real-negative training beats decoy training on the realistic test for *every* one of the 25 targets (median paired advantage +0.031 ROC-AUC; paired Wilcoxon $p < 10^{-6}$), so the directional claim is unambiguous. Decoy-trained models are overoptimistic by ~0.055 when validated on decoys rather than real negatives—the documented decoy-bias failure mode [3,4], reproduced here. The *effect size*, however, is borderline against our a priori bar: the advantage clears 0.03 on 13/25 targets (52%; Clopper–Pearson 95% CI 31–72%, which spans the 60% threshold — i.e. indistinguishable from it at n=25), and the rate is bar-sensitive (72% at 0.02, 32% at 0.04). We therefore report a *consistent but modest* advantage for real negatives, plus the avoidance of decoy-validation optimism—not a blanket claim of superiority. (PR-AUC and Brier per condition would test whether real negatives also improve calibration; not yet computed.)

4.6 Temporal split: does the headline hold prospectively?

Section 4.5 asked *which negatives* to train on; the remaining question is whether the split itself flatters the result. Scaffold-disjoint AUC can still be optimistic if train and test share analogue series [4,12,13]. The honest stress test is temporal [14,17]: train on what was known by a cutoff, predict what gets characterized later. We assign each (compound, target) pair its first ChEMBL publication year and, for the 25-kinase cohort, train on pairs first characterized ≤ 2018 and test on those ≥ 2020 (2019 a buffer). All 25 targets had sufficient pre- and post-cutoff data to evaluate, with no per-target exclusions; per target, training sets ranged 938–7,368 pairs and the prospective test 34–1,366 pairs (median 378), each retaining both classes (≥15 active and inactive in train, ≥5 in test). The drop is large (Table 6, Fig. 5).

Split	Median ROC-AUC (IQR)
Scaffold-disjoint (same 25 targets)	0.966
Temporal (train ≤2018 / test ≥2020)	0.775 (0.734–0.850)

Table 6. Temporal vs scaffold split, 25 kinases (median per-target ROC-AUC). Median paired drop -0.179.

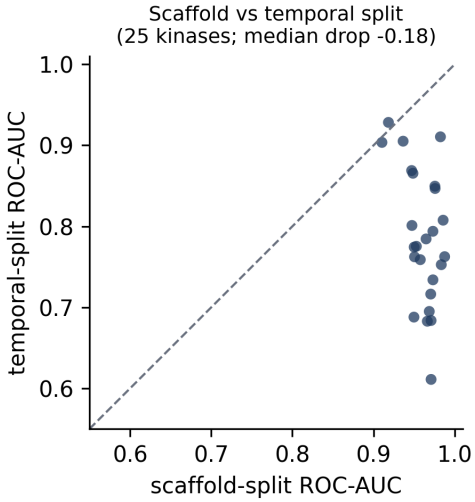

Figure 5. Per-target scaffold vs temporal-split ROC-AUC. Points fall well below the diagonal: the model degrades substantially on compounds characterized after the training cutoff.

The scaffold-split headline (0.966) falls to 0.775 under the temporal split — a 0.18-point drop, mirroring the degradation Lenselink reported (MCC 0.49 to 0.17 [14]). Two honest readings follow: (i) the realistic forward-looking number for a screening anti-filter is ~0.78, not ~0.97 — the scaffold figure overstates prospective accuracy; and (ii) 0.78 is still well above chance, so the measured negatives carry real, if more modest, prospective signal. We report the temporal number as the honest headline and the scaffold number as an upper bound.

Finally, we re-ran the Sec 4.5 decoy comparison *under* this temporal split (train-era negatives, prospective real test). The clean real-over-decoy advantage does not survive the harder regime: real and decoy training both fall to ~0.775 (median real 0.775, decoy 0.774), the median advantage shrinks from +0.031 (scaffold, 25/25 targets) to +0.014 (temporal, 17/25 targets). So the evidence that measured negatives beat decoys is robust under scaffold splitting but only marginal, directionally-preserved, under prospective evaluation — another reason we frame the contribution as preliminary.

5. Limitations

This study establishes signal and internal consistency, and now a prospective lower bound; the two framing-critical experiments (decoy comparison Sec 4.5, temporal split Sec 4.6) are both performed. **(1) Split validity.** Bemis–Murcko scaffold splits leave analogue-series leakage [4,13]; our temporal split (Sec 4.6) confirms the scaffold figure is optimistic — AUC drops 0.18 to 0.775, mirroring Lenselink's MCC 0.49-to-0.17 collapse [14] (and cluster-CV baselines are similarly low, ECFP+RF ~0.68 [15]). The one split-validity control still un-run is AVE-debiasing [12], which would quantify train/test near-neighbour redundancy directly. **(2) Statistics.** A Y-randomization (label-permutation) control on the 25-kinase cohort gives median ROC-AUC 0.467 (IQR 0.45–0.50, straddling chance; the deviation from 0.50 is sampling variance over 25 targets), confirming the model learns label structure rather than fingerprint structure; we report paired Wilcoxon tests for the random-vs-scaffold gap but not yet bootstrap CIs, multiple-testing correction, or per-target reliability diagrams / ECE. **(3) Prior art.** Per-target classifiers on measured inactives are established (PIDGIN [11]); our contribution is scale and the negative-results framing, not the method.

Data and modelling caveats. **(4) Label engineering.** Discarding the 1–10µM band raises class separability, and ~0.68-log cross-laboratory IC50 noise [16] places the one-log active/inactive boundary at ~1.5 SD of noise. **(5) Assay heterogeneity.** Inactives are pooled across binding, cellular, and panel assays; dropping intra-target conflicts (1.2–1.4%) may preferentially remove promiscuous / frequent-hitter compounds and sanitize the inactive class. **(6) Calibration.** Brier only (no ECE or reliability diagrams); isotonic at ≥500/class is below the ~1,000-point crossover [9], where Platt is safer. **(7) Scope & provenance.** Kinases and GPCRs only; curated tier only; single-protein human filter (which discards informative ortholog and complex assays); findings are auto-extracted, not manually verified. We therefore present this as preliminary validation, not a prospective benchmark.

6. Availability

The negative-results layer, the target-exhaustion index (*get_target_landscape*) and the scoring registry are accessible through Nullary's MCP server and REST API. The validation and training pipelines are deterministic and reproducible from the cited public ChEMBL release. Per-target results (UniProt, n_active, n_inactive, ROC-AUC, PR-AUC, Brier, scaffold/random gap) are provided as supplementary CSVs. Contact: nullary.ai.

References

[1] Mendez D, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* 2019;47(D1):D930–D940.

[2] Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E). *J Med Chem.* 2012;55(14):6582–6594.

[3] Chen L, et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS ONE* 2019;14(8):e0220113.

[4] Sieg J, Flachsenberg F, Rarey M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J Chem Inf Model.* 2019;59(3):947–961.

[5] Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model.* 2010;50(5):742–754.

[6] Ke G, et al. LightGBM: a highly efficient gradient boosting decision tree. *NeurIPS* 2017.

[7] Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem.* 1996;39(15):2887–2893.

[8] Wu Z, et al. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci.* 2018;9(2):513–530.

[9] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. *ICML* 2005.

[10] RDKit: open-source cheminformatics. https://www.rdkit.org

[11] Mervin LH, et al. Target prediction utilising negative bioactivity data covering large chemical space (PIDGIN). *J Cheminform.* 2015;7:51.

[12] Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. *J Chem Inf Model.* 2018;58(5):916–932.

[13] van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. *J Chem Inf Model.* 2022;62(23):5938–5951.

[14] Lenselink EB, et al. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J Cheminform.* 2017;9:45.

[15] Mayr A, et al. Large-scale comparison of ML methods for drug target prediction on ChEMBL. *Chem Sci.* 2018;9:5441–5451.

[16] Kalliokoski T, Kramer C, Vulpetti A, Gedeck P. Comparability of mixed IC50 data – a statistical analysis. *PLoS ONE* 2013;8(4):e61007.

[17] Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J Chem Inf Model.* 2013;53(4):783–790.