Genesis Molecular AI – Technical Report

RUBY: State-of-the-Art Potency Prediction across Discovery Regimes Using Learned Structural Representations

Genesis Research Team

Abstract—Small-molecule drug discovery involves repeated rounds of designing, making, testing, and analyzing compounds. The goal is often to maximize potency while keeping other properties within the ranges required for a viable therapeutic candidate. Compound prioritization places different demands on computational methods as drug discovery progresses from hit identification through hit-to-lead and lead optimization, with changing computational budgets and increasing amounts of on-target experimental data. We introduce RUBY, a Robust Universal Binding affinitY predictor that uses intermediate representations from the structural foundation model PEARL without requiring an explicit binding pose for every compound. RUBY is trained on proprietary data plus `GenesisPotency-2026`, an audited public-source potency training dataset that we are releasing for academic and commercial use under a CC BY-SA 4.0 license with row-level provenance and audit annotations. RUBY achieves the highest aggregate early enrichment across seven public hit-identification targets among evaluated methods, with adjustable inference effort providing a trade-off between enrichment and computational cost over more than two orders of magnitude. RUBY achieves the highest average correlations across four public benchmarks: OpenFE, CASP16, FEP+ 4, and OpenBind. RUBY significantly outperforms other methods in zero-shot prediction across eight biochemical and cellular assays from five ongoing Genesis–Incyte programs, widening its performance advantage relative to the public benchmarks. RUBY continues to improve more than other methods when all are given the opportunity to recalibrate using earlier assay measurements. Together, these contributions provide an annotated training resource for the field and a state-of-the-art potency predictor for broad screening, prioritization within related chemical series, and assay-specific prediction as experimental measurements accumulate.

1. Introduction

1.1. Regimes of drug discovery

The requirements for compound prioritization change with the available experimental data, structural knowledge, and computational budget across hit identification (HitID), hit-to-lead (H2L), and lead optimization (LeadOp).

Hit identification. HitID involves scoring enumerated libraries of tens of thousands to millions of compounds or combinatorial spaces up to 100 billion molecules [1]. Predictors must concentrate diverse and active compounds at the top of a ranking so that a small subset can be experimentally tested. Prioritization at this scale requires computational methods fast enough to operate within practical time and compute budgets.

Hit-to-lead. H2L involves expanding a handful of confirmed active compounds into congeneric series around a common core scaffold, for example through R-group substitutions. Experimental follow-up may involve only tens to hundreds of synthesized compounds, making accurate ranking among closely related candidates important.

Lead optimization. LeadOp usually begins after tens to hundreds of compounds have on-target assay measurements that can guide compound exploration and refinement across multiple parameters such as potency, selectivity, permeability, clearance, and bioavailability, as well as synthetic feasibility. Learning from earlier on-target measurements and refining the feedback loop between prediction and experiment become more important as these properties are balanced against each other.

Assay heterogeneity. The experimental quantities predicted also vary across programs. Binding affinity describes the strength of a protein–ligand interaction, whereas biochemical and cellular potency describe effects measured under particular assay conditions. While related, these quantities are not interchangeable: an IC_{50} can depend on assay format and conditions, and cellular activity additionally reflects factors such as compound permeability and the biological response being measured. Even measurements against the same target can therefore differ across assays [2, 3]. A useful potency predictor must consequently support compound ranking across binding affinity, biochemical potency, and cellular potency, as well as adaptation to the particular assay used to guide a program.

1.2. Computational tools complement experimental measurement

Computational methods help chemists decide which compounds to synthesize and test, with different approaches suited to different combinations of available data, structural knowledge, and computational budget. We briefly summarize the main computational approaches to potency prediction and refer readers to broader reviews for more extensive discussion [4–6].

Quantitative structure–activity relationships. Quantitative structure–activity relationship (QSAR) models learn associations between molecular structure and measured activity. Classical approaches use molecular descriptors or fingerprints, whereas neural models such as Chemprop learn representations directly from molecular graphs [7]. These ligand-based models can make inexpensive predictions and learn from a program's measured structure–activity relationships, but their usefulness depends on how well the available training measurements cover the compounds being considered. Multi-task learning can share information across assays or targets, though transfer to a new target depends on how the model represents that target and on the relevance of the training tasks [8].

Molecular docking. Docking uses an experimental or predicted protein structure to search for ligand conformations and orientations within a binding site. Conventional methods usually hold the receptor fixed or allow limited flexibility, and they rank candidate poses using approximate scoring functions. These scores may combine physical interaction terms, empirically fitted contributions, or statistical patterns learned from known complexes [9]. Docking can prioritize compounds at large scale and provide explicit binding hypotheses for interpreting interactions and designing analogs. Recent large-library studies show that docking scores can contain useful affinity information, though they can also expose errors and artifacts among highly ranked compounds [10, 11].

Binding free energy calculations. Alchemical binding free-energy methods use molecular dynamics simulations and force fields to estimate thermodynamic differences between states. Relative binding free-energy (RBFE) calculations estimate affinity differences by transforming one ligand into another in the bound and unbound states, making them particularly useful for comparisons within a series where chemical transformations are often modest. Absolute binding free-energy (ABFE) calculations estimate the binding free energy of an individual ligand, allowing comparisons without requiring a closely related reference ligand. Both approaches require suitable structural models, careful system preparation, and sufficient conformational sampling, and their very high computational cost generally limits the number of compounds that can be evaluated [12].

Structure prediction. An experimentally determined complex is not always available when compound prioritization begins, and even an available protein structure may not capture the conformation relevant to a candidate ligand. Binding can reorganize side chains and loops, and cryptic pockets may be absent from the observed conformation, complicating the use of a single fixed structure for screening [5]. Cofolding models

such as PEARL predict protein–ligand complexes jointly, providing structural hypotheses for interpreting interactions and guiding compound design [13].

Potency prediction from structural representations. A plausible binding pose does not by itself establish whether a compound binds strongly or how it ranks against other compounds. Boltz-2 addresses this additional task with an affinity module that uses learned pair representations and predicted coordinates [14], while Nesso-1, AQAffinity, and ApherisAffinity also connect learned structural representations with affinity prediction [15–17]. Other approaches include shared pocket–ligand embeddings in LigUnity and coarse structural representations in TerraBind [18, 19], and proprietary systems such as IsoDDE report strong affinity prediction performance [20].

In PEARL and other cofolding models, intermediate pair representations encode relationships among protein and ligand tokens upstream of the module that generates atomic coordinates. These representations offer an opportunity to predict potency directly, avoiding explicit pose generation for every compound while retaining information learned through structural training. Reducing this computation could make structural representations useful across a wider range of screening budgets.

Deploying to partner programs. Public potency benchmarks enable reproducible comparisons, but their selected targets, chemical series, and assay endpoints capture only part of the prediction challenge in pharmaceutical discovery. Ongoing programs pursue novel chemical matter, less-characterized therapeutic targets, or both, while generating measurements under assay conditions specific to their experimental objectives. Models must therefore transfer learned relationships to new compounds and experimental contexts, making evaluation on active pharmaceutical programs an essential complement to public benchmarks.

Genesis has a strategic AI-focused research collaboration with Incyte, building and deploying state-of-the-art AI to accelerate the discovery of novel molecules for collaboration targets selected by Incyte. The collaboration started in 2025 and was expanded in 2026. The expanded agreement is one of the first major Pharma-AI collaborations to power large-scale foundation model training with a partner's proprietary experimental data.

1.3. RUBY: Robust Universal Binding affinitY across drug discovery regimes

We introduce RUBY, a potency prediction model that uses PEARL's intermediate pair representation to predict activity without generating an explicit protein–ligand pose for each compound. We evaluate its use for early enrichment in HitID, zero-shot potency ranking in H2L, and prediction after assay-specific recalibration and fine-tuning in LeadOp. These settings test whether the same pretrained predictor can support both broad screening with limited on-target data and focused prioritization as experimental measurements accumulate.

Contributions. Our primary contributions are an annotated training corpus for the community and a robust, universal model applicable across discovery regimes:

- **An annotated training dataset for community use.** We introduce `GenesisPotency-2026`, an audited public-source training dataset that we are releasing for academic and commercial use under a CC BY-SA 4.0 license with row-level provenance and audit annotations (§2.1).
- **Hit identification.** RUBY achieves the highest aggregate early enrichment among the evaluated methods across experimentally tested compounds from seven targets in four CACHE challenges and three published large-scale virtual screens [10, 21–25]. We examine enrichment together with inference time to characterize its usefulness under different computational budgets (§3.1).

- **Zero-shot potency ranking.** RUBY achieves leading correlations compared with both open models and established affinity prediction methods across the OpenFE, FEP+ 4, CASP16, and OpenBind benchmarks (§3.2). This advantage extends to ongoing Genesis–Incyte drug-discovery programs, where RUBY achieves higher aggregate zero-shot correlations than all evaluated comparison methods across eight biochemical and cellular assays on five targets (§3.3).
- **Adaptation to ongoing discovery programs.** Using earlier measurements from each assay, RUBY achieves higher correlation and lower prediction error after rolling-temporal recalibration than the evaluated comparison methods on Genesis–Incyte programs (§3.3).

2. Methods

2.1. `GenesisPotency-2026` and training data

`GenesisPotency-2026` and training sources. RUBY is trained on a combination of public-source and proprietary activity measurements. The public-source portion, `GenesisPotency-2026`, assembles measurements from ChEMBL 37, BindingDB release 202607, and PubChem HTS (2025) [26–28]. We define $\mathrm{p}X = -\log_{10}(X)$, where X is the numerical value of IC_{50}, K_i, or K_d expressed in mol/L. Higher pX indicates stronger binding or greater potency, depending on the measured endpoint. We restrict potency regression to IC_{50}, K_i, and K_d measurements and retain EC_{50} measurements for binary activity labels. We group measurements by source document, target, and readout, and apply ligand- and assay-eligibility filters described in Appendix A. Figure 1(c) summarizes the resulting training composition: 5.41 million public-source potency data points after the audit, compared with 1.42 million in AQAffinity's released training set. Figure 1(d) breaks them down by drug-target family, and Appendix A.4 compares target coverage with AQAffinity. Following IsoDDE [20], we restrict `GenesisPotency-2026` to measurements dated before 1 January 2023 and apply the benchmark-specific target-exclusion rules detailed in Appendix B.

Source-document audit. Databases can share source records and propagate the same transcription errors, so agreement between databases does not necessarily provide independent confirmation of molecular identities, target assignments, endpoints, or units. We audited available ChEMBL and BindingDB source documents in two passes (Figure 1(a)): assay-level checks of target identity, experimental context, endpoints, and units, followed by checks of up to 15 sampled compounds per document. Source-supported corrections and exclusions were applied to the training corpus, with decision rules and handling of unresolved corrections detailed in Appendix A. We could audit only the subset of records for which source material was openly accessible.

Audit outcomes. Of the 5.90 million pre-audit potency data points, 41.9% passed source-document review and 8.4% were removed or withheld (Figure 1(b)). Appendix A.3 reports the outcomes by source. The audit identified errors that would change the biological meaning of a data point: for instance, the K_i values of 885 FLAP inhibitors from one patent had been assigned to the unrelated protein flap endonuclease 1 in both databases. In another patent, 10 of 15 sampled records depicted a synthetic intermediate in place of the tested compound (Figure 7).

Annotated release. We are releasing `GenesisPotency-2026` for academic and commercial use under a CC BY-SA 4.0 license with row-level provenance and audit annotations[1]. Appendix A describes the audit coverage, decision rules, and distinction between source records and potency data points.

[1]Dataset and release documentation: https://github.com/genesistherapeutics/genesispotency.

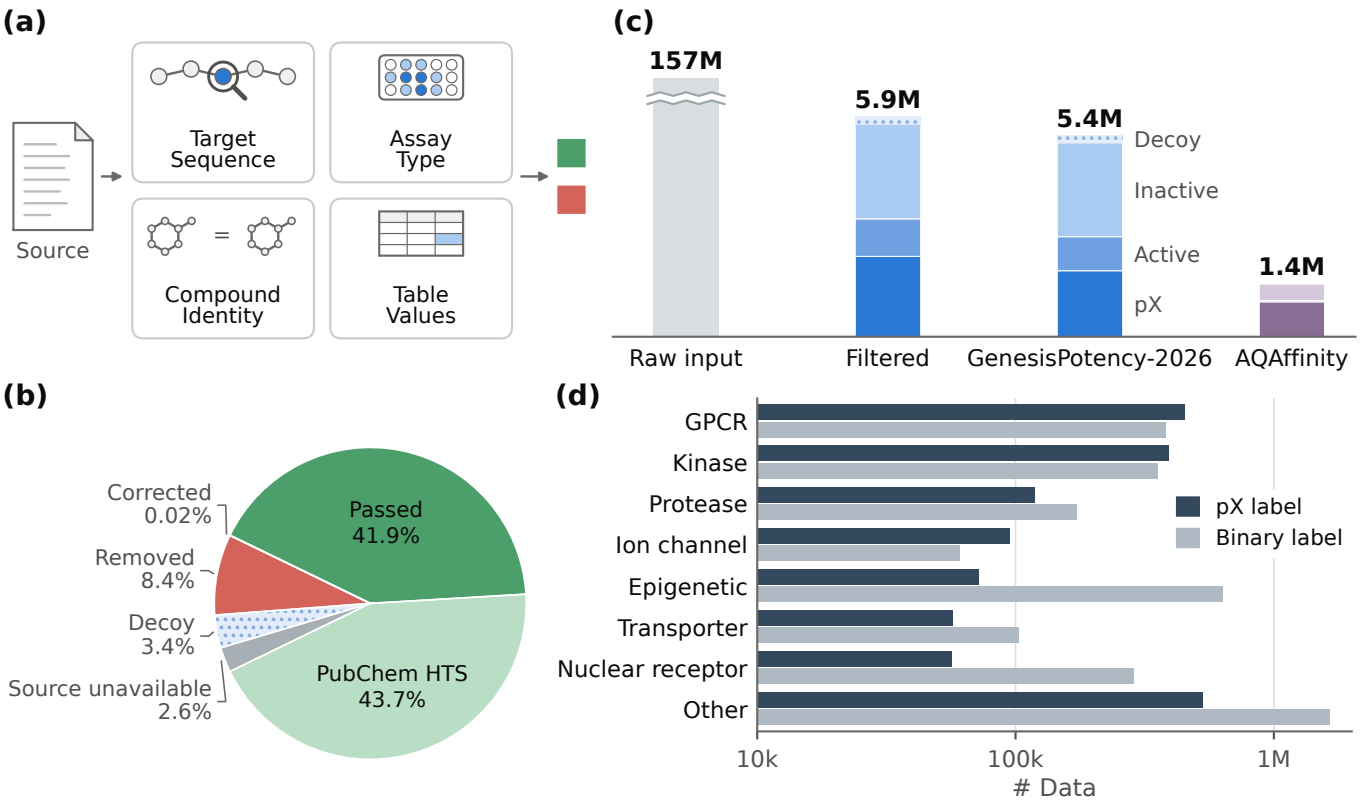


Figure 1. Audit, composition, and target coverage of `GenesisPotency-2026`. **(a)** Two-pass source-document audit. **(b)** Audit status of the 5.90 million pre-audit potency data points. Of the passed data points, 32% were reviewed against the full text and 68% against the abstract or supplement only. PubChem HTS screening calls have no source document and pass without review. **(c)** Raw input rows (bar truncated) and potency data by label type before and after the audit, with AQAffinity's released training set [16]. **(d)** `GenesisPotency-2026`'s potency data points by drug-target family [26], on a log scale. Table 2 lists all target groups.

2.2. Model

RUBY predicts protein–ligand activity using learned representations from the structural foundation model PEARL [13]. It takes the same types of inputs: a protein amino acid sequence, a ligand represented as a SMILES string, and optional binding-site information and structural templates. RUBY uses the PEARL trunk's protein–ligand pair representation without generating an explicit binding pose for each compound. The model produces a binary activity score for distinguishing active from inactive compounds and a regression estimate on the pX scale.

2.3. Inference and conditioning

We use inference effort to describe the computational budget allocated to scoring each compound. For hit identification, we evaluate four RUBY settings—Low, Medium, High and Max—using the same trained checkpoint. These settings vary the computation devoted to protein–ligand representations, allowing a trade-off between early enrichment and inference time. All four settings score the same protein–ligand pairs without retraining the model or generating explicit binding poses.

RUBY supports individual-compound and compound-list prediction. In list mode, a compound's predicted potency can depend on the other compounds supplied alongside it, so this distinction affects predictions rather than only computational batching. The hit-identification evaluation uses no binding-site conditioning or structural templates. For public potency benchmarks, RUBY automatically selects among candidate input configurations for each target. Partner-program evaluations use each program's existing pocket conditioning and structural template inputs across all methods. Appendix C.1 describes protein inputs, ligand preparation, and structural conditioning.

2.4. Evaluation sets

We evaluate RUBY across three complementary settings: hit identification on experimentally tested screening compounds, zero-shot potency ranking on public benchmarks, and prediction within ongoing pharmaceutical programs. For the public benchmarks, we pursue aggressive train–test separation through both a pre-2023 measurement cutoff and benchmark-specific target exclusions based on protein sequence similarity. Appendix B specifies the date and sequence-similarity thresholds, their application to each benchmark, and the corresponding policies reported for comparison methods.

Hit identification. We evaluate seven public sets containing 9,364 compounds and 1,008 positive activity labels: D_4, σ_2, and AmpC from large-library docking campaigns, and LRRK2 WD40, NSP13, Mac1, and CBLB TKB from the first four CACHE challenges [10, 21–25]. Our evaluation reranks the experimentally tested compounds from these campaigns and measures enrichment against their observed activity labels. Appendix C discusses other HitID benchmarks.

Public potency ranking. We evaluate 1,100 measurements from 64 source series across OpenFE, FEP+ 4, and CASP16 (autotaxin and chymase), plus 494 compounds from OpenBind [29–33]. Appendix C.3 details cohort membership, label sources and conversion to pX.

Partner programs. We evaluate potency prediction on eight biochemical and cellular potency assays across five Genesis–Incyte targets, totalling 22,036 assay–compound measurements. We report zero-shot performance on the full scored cohort and evaluate assay-specific recalibration using a rolling temporal split, building on prior time-split evaluation of industrial assay models [34]. At each weekly cutoff, recalibration uses earlier exact measurements from the same assay to predict compounds in a seven-day window beginning 28 days later. Recalibrated evaluation requires at least 20 exact measurements in the preceding training history. Base prediction models remain fixed, and comparator scores receive a separate assay-specific affine recalibration at each cutoff. Appendices C.4 and C.5 describe cohort eligibility and the rolling temporal split.

2.5. Metrics and uncertainty

Hit identification. We measure early enrichment using a chance-corrected variant of the normalized enrichment factor of Liu et al. [35]:

$$\mathrm{nEF}_q = \frac{\mathrm{EF}_q - 1}{\mathrm{EF}_{\max,q} - 1},$$

where EF_q is the active fraction among the top-ranked fraction q of compounds divided by the active fraction in the full set, and $\mathrm{EF}_{\max,q}$ is its maximum achievable value. The normalization assigns random ranking an expected score of 0 and the best achievable ranking a score of 1. We average per-target values with equal weight so that larger screens do not dominate the aggregate. Figure 2 uses nEF@1% for the inference-time comparison, and Figure 3b reports mean nEF across selection fractions from 0.2% to 10%. Appendix D

reports per-target nEF from 0.2% to 10% and BEDROC [36].

Public potency. For OpenFE, FEP+ 4, and CASP16, we report Pearson r between predicted and measured potency within each source series, then average within each benchmark weighted by compound count. We keep the 58 OpenFE source series separate. FEP+ 4 and CASP16 each contain one series per target. We use Pearson r to support comparison with published model results. For OpenBind, we report Spearman ρ over the full 494-compound cohort, following its published evaluation [33]. Appendix E provides additional correlation metrics and detailed comparisons with published results.

Partner programs. We evaluate within-assay ranking using Spearman ρ and Kendall τ, linear association using Pearson r, and absolute accuracy after recalibration and fine-tuning using mean absolute error (MAE) in pX units. We report correlations in both zero-shot and recalibrated settings and MAE only after recalibration. Correlations include censored measurements at their reported bounds, whereas MAE uses only uncensored measurements. Aggregate results weight assays equally. Appendix C.4 details eligibility, including the requirement for finite fitted predictions from every method in the recalibrated setting.

Program chemical similarity. For the Genesis–Incyte programs, we examine how prediction performance varies with chemical similarity to prior assay measurements. We define similarity as the maximum Tanimoto similarity between each test compound's nonchiral Morgan fingerprint (radius 3, 2,048 bits) and those of compounds with exact measurements available before its rolling cutoff in the same assay. Within each assay, we pool rolling-temporal recalibrated predictions across all test windows and group compounds by their similarity to measurements available at their respective cutoffs. We compute Spearman ρ and MAE within each similarity bin, using the same eligible compounds across methods for each metric and averaging assay-level metrics equally. This analysis asks how performance varies with chemical familiarity, complementing the whole-assay correlations reported above. Appendix F.3 gives fingerprint parameters, binning rules, and scoring details.

Uncertainty. We estimate uncertainty using 1,000 bootstrap resamples of compounds within each target or assay, using identical resampled compounds across methods for paired comparisons. Error bars show one bootstrap standard error of the reported statistic, reflecting uncertainty from compound sampling with predictions held fixed. Variability from model training or target selection is not captured by these estimates. Appendix C.6 provides resampling and aggregation details.

For each metric in the partner-program figures, we compare RUBY with the comparator having the best aggregate point estimate on the same eligible cohort. Stars summarize one-sided paired bootstrap comparisons, with calculation details in Appendix C.6 and results in Appendix Table 10.

2.6. Baselines

We compare RUBY with models and docking methods evaluated on the same benchmark compounds, published per-compound predictions re-scored under our evaluation protocol, and published aggregate results. For published FEP+ and OpenFE relative binding free-energy (RBFE) predictions, we use the same eligible compounds, metrics and bootstrap procedure as for the methods we run. We identify cited aggregate results separately because they do not support matched uncertainty estimates or paired comparisons. Appendices E.1 and C.7 describe prediction sources, comparator configurations and coverage.

Structure-based comparators. We run Boltz-2 [14], Nesso-1 [15], LigUnity [18] and AQAffinity [16] with their released code and weights, using the evaluation-specific settings documented in Appendix C.7. LigUnity uses its screening checkpoint for hit identification and its ranking checkpoints for potency.

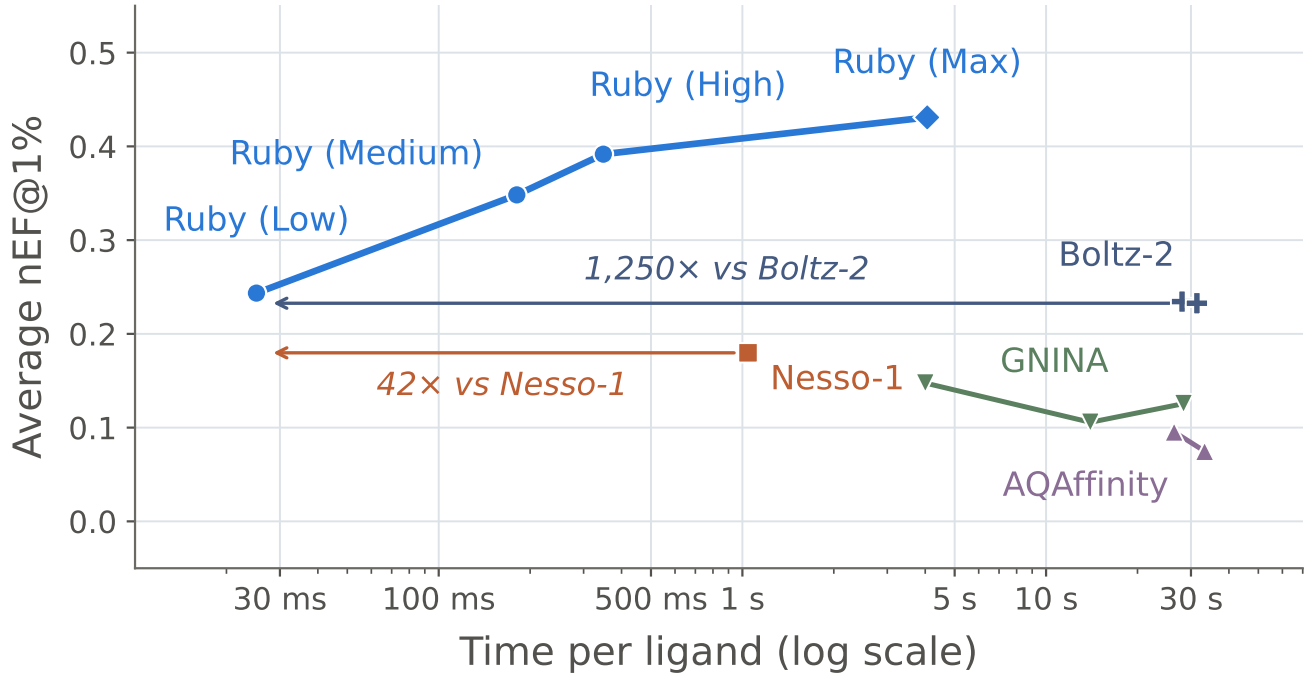


Figure 2. Time per ligand versus equal-target average nEF@1% across seven screening targets. RUBY Low, Medium, High and Max use the same checkpoint, with matched predictions and warmed timing measurements for each setting on Nvidia B200 hardware. The connecting line identifies RUBY's effort settings. AQAffinity points use one and five samples, and GNINA points use one, four and nine conformers. The Boltz-2 line connects 1/1- and 2/5-sample enrichment settings, using RTX PRO 6000 timing for 1/1 and B200 timing measured at 5/5 samples for 2/5.

Docking and trivial baselines. For our docking runs, we use GNINA [37] and rank compounds by negated empirical docking energy. Molecular weight serves as a size-only control. Appendix C.7 specifies receptor preparation, docking settings and coverage.

Cited results. For IsoDDE [20], ApherisAffinity [17], TerraBind [19], and Nesso-1's OpenBind result [15], we use published summary values because per-compound predictions are unavailable to us for these comparisons. We evaluate CASP16 retrospectively and derive TerraBind's displayed aggregate from its published per-target values under our weighting. Appendix E gives the matching, derivation and sensitivity checks.

3. Results

3.1. Inference cost and hit identification on seven public targets

Figure 2 shows increasing early enrichment with RUBY's inference effort across the seven public screening targets. Mean nEF@1% rises from 0.24 at 25.11 ms per ligand for RUBY Low to 0.43 at 4.06 s for RUBY Max. RUBY Medium and High provide intermediate settings, reaching 0.35 at 180.77 ms and 0.39 at 348.48 ms per ligand, respectively.

RUBY Max achieves a mean nEF@1% of 0.43 versus 0.23 for the Boltz-2 reference setting, and it scores compounds more than 7.5 times faster at the recorded inference rates. Similarly, RUBY High reaches 0.39 versus 0.18 for Nesso-1 while being 3 times faster. Even RUBY Low exceeds both methods in mean enrichment

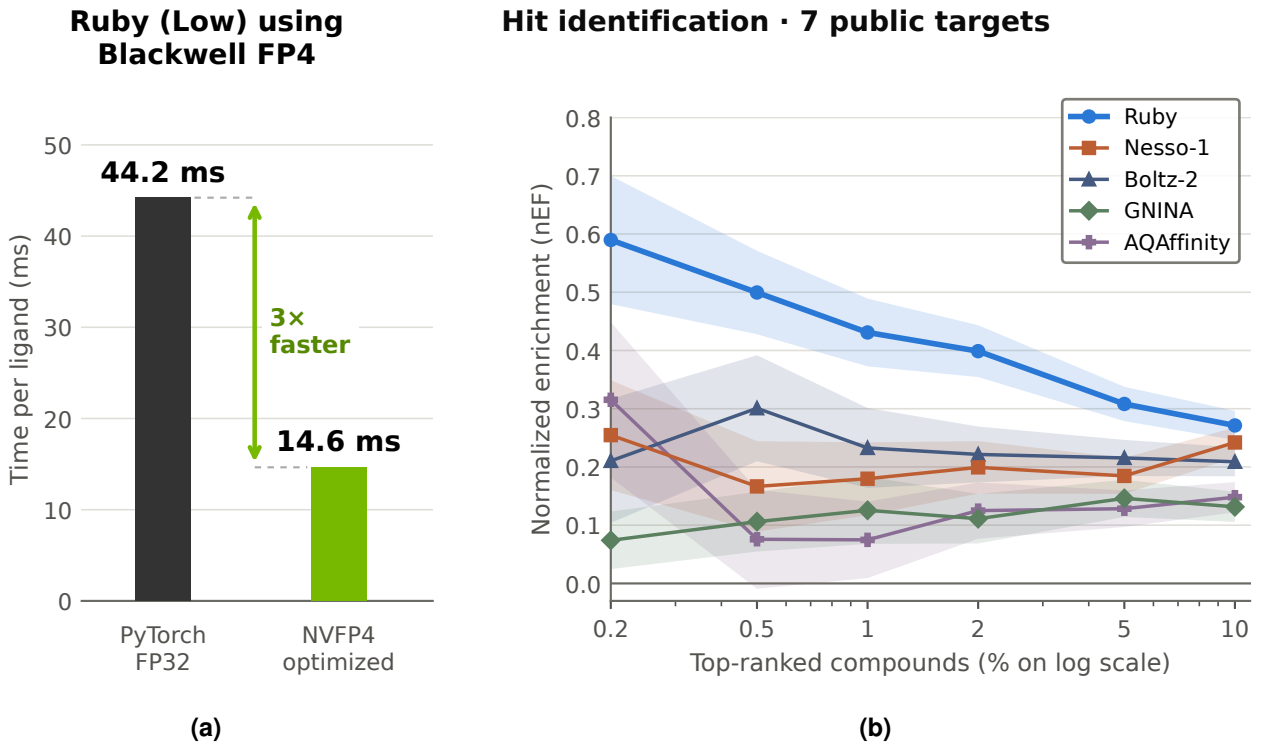


Figure 3. Kernel efficiency and early enrichment. **(a)** Nvidia Blackwell kernel optimization at fixed RUBY Low effort on Nvidia B200 hardware over σ_2 compounds. The optimized stack combines NVFP4 kernels, Transformer Engine, cuBLAS BF16x9, CUDA Graphs and cuEquivariance attention. Bars show mean time per ligand. **(b)** Early enrichment on seven public targets. Lines give the equal-target mean nEF, and shading shows ± one bootstrap SE. Each model uses its highest screening setting evaluated here. Table 6 in Appendix D gives per-target results.

while being 1250 and 42 times faster, respectively.

Figure 3a compares native and optimized GPU execution at fixed RUBY (Low) effort using the same checkpoint on Nvidia B200 hardware. On σ_2, the optimized kernel stack reduces warm inference time from 44.2 to 14.6 ms per ligand, giving approximately threefold higher throughput at the same model effort.

Figure 3b shows nEF across selection fractions from 0.2% to 10% averaged over the seven public screening targets. RUBY (Max) has the highest mean nEF among the plotted methods at every evaluated selection fraction. At 1%, RUBY reaches 0.43 versus 0.23 for Boltz-2 ($p = 0.007$), and at 5%, it reaches 0.31 versus 0.22 ($p = 0.011$). Table 6 in Appendix D reports per-target results across all six selection fractions.

3.2. Zero-shot potency ranking on four public benchmarks

Figure 4 compares RUBY's zero-shot potency ranking in list-inference mode with other methods on the OpenFE, CASP16, FEP+ 4, and OpenBind benchmarks. RUBY has the highest average correlation among the methods. On OpenFE, RUBY reaches $r = 0.74$ across 58 source series spanning 39 targets, compared to IsoDDE at 0.73, FEP+ at 0.72, and surpassing all other methods. On CASP16, RUBY reaches $r = 0.77$, compared to IsoDDE at 0.75, Nesso-1 at 0.65, and surpassing all other methods. On the four-target FEP+ 4 subset, RUBY reaches $r = 0.86$, compared to IsoDDE at 0.85, Nesso-1 at 0.81, and surpassing all other

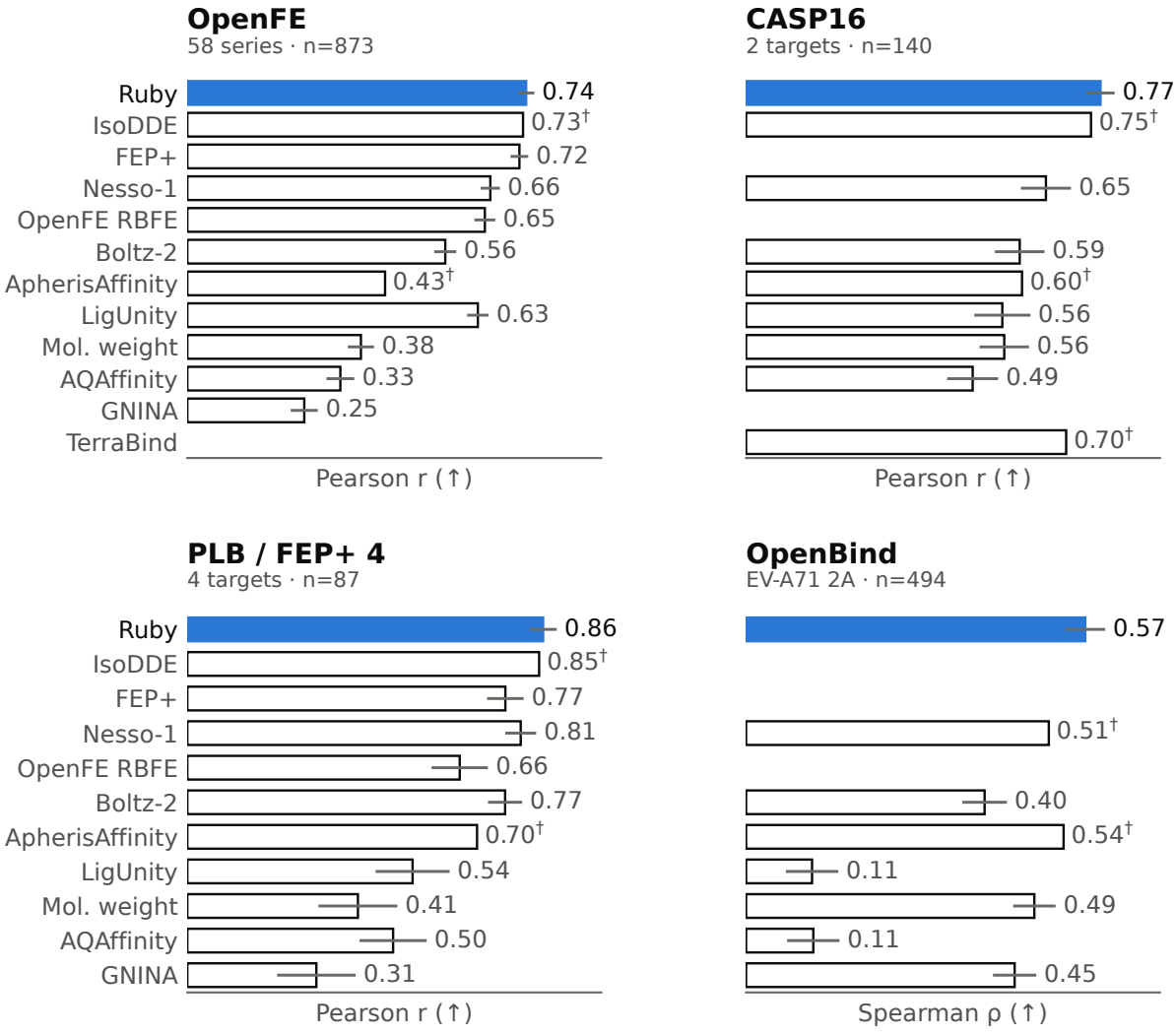


Figure 4. Potency ranking performance over four public benchmarks. For re-scored methods, Pearson correlations are computed within the 58 OpenFE source series and within each CASP16 and FEP+4 target, then averaged within each benchmark using compound-count weights. OpenBind uses Spearman correlation over its full 494-compound cohort. Daggered (†) IsoDDE, ApherisAffinity and TerraBind have no released compound-level predictions here, so they have no matched error bars. Boltz-2 bars show our reruns, and its original report gives $r = 0.62$, 0.65 and 0.66 on OpenFE, CASP16 and FEP+4, respectively [14].

methods. On OpenBind, RUBY reaches $\rho = 0.57$, compared to ApherisAffinity's 0.54, Nesso-1's 0.51, and surpassing all other methods.

Tables 7 and 8 report per-series and per-target results and additional correlation metrics, while Appendix E documents prediction provenance, cohort matching, and sensitivity to aggregation.

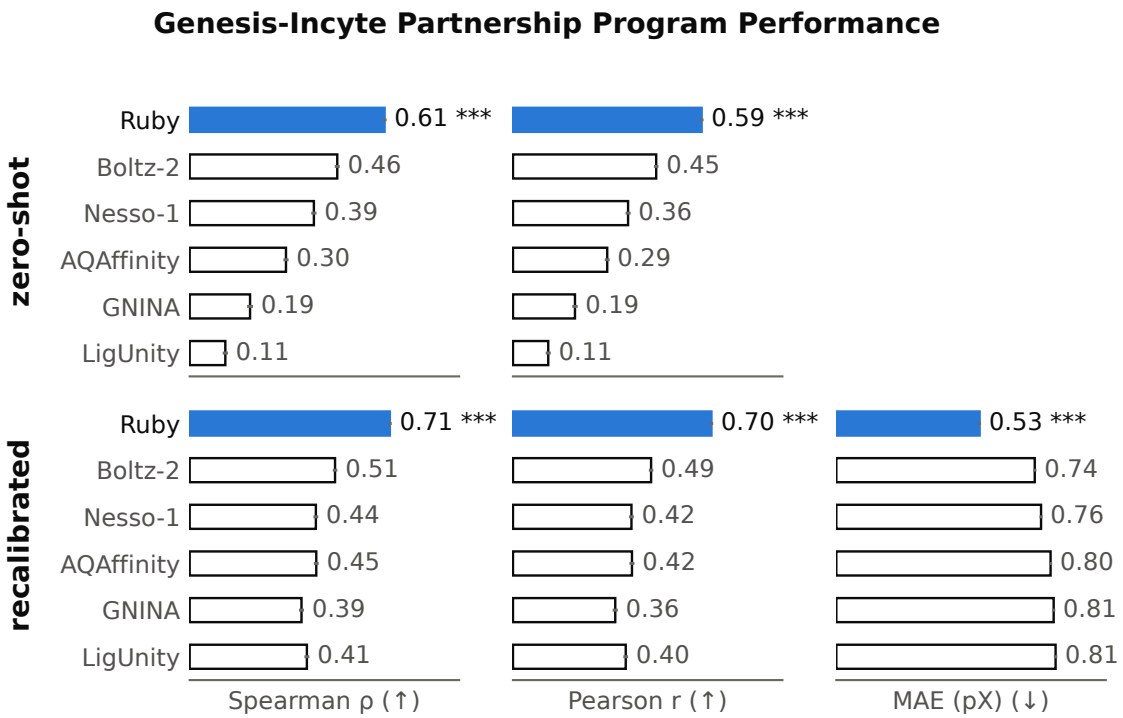


Figure 5. Performance across eight Genesis–Incyte assays on five targets. The top row shows zero-shot Spearman ρ and Pearson r across these assays. The bottom row shows correlations and MAE of rolling-temporal recalibrated predictions. Whiskers show $\pm$ one bootstrap standard error from compound resampling within the eight fixed assays. Three stars on each RUBY bar indicate a one-sided paired bootstrap $p < 0.001$ against the best plotted comparator for that metric (Boltz-2, Appendix Table 10).

3.3. Lead optimization on Genesis–Incyte programs

Figure 5 tests RUBY on experimental measurements from ongoing Genesis–Incyte drug-discovery programs, extending the evaluation to proprietary chemical series and program-specific biochemical and cellular assays. Across these settings, RUBY achieves the highest aggregate zero-shot and recalibrated correlations among the evaluated methods.

The top row of Figure 5 shows zero-shot potency ranking across eight assays on five Genesis–Incyte targets, with each assay weighted equally. RUBY reaches mean Spearman $\rho = 0.61$ and Pearson $r = 0.59$, compared with 0.46 and 0.45 for Boltz-2, the strongest aggregate comparator for both metrics. Both Boltz-2 and Nesso-1 achieve lower Pearson correlations here than on each of the three public benchmarks evaluated with that metric, emphasizing the importance of testing performance directly in active programs.

The bottom row of Figure 5 shows performance over rolling-temporal recalibration within the assay. RUBY's mean MAE is 0.53 pX, compared with 0.74 pX for Boltz-2, the lowest-MAE plotted comparator. On the shared recalibrated cohort, RUBY reaches mean $\rho = 0.71$ and $r = 0.70$, compared with 0.51 and 0.49 for Boltz-2. Appendix Table 9 reports results for each assay.

Figure 6 shows recalibrated ranking and prediction error as a function of maximum Tanimoto similarity to earlier compounds measured in the same assay. Across bins represented in all eight assays, RUBY's MAE

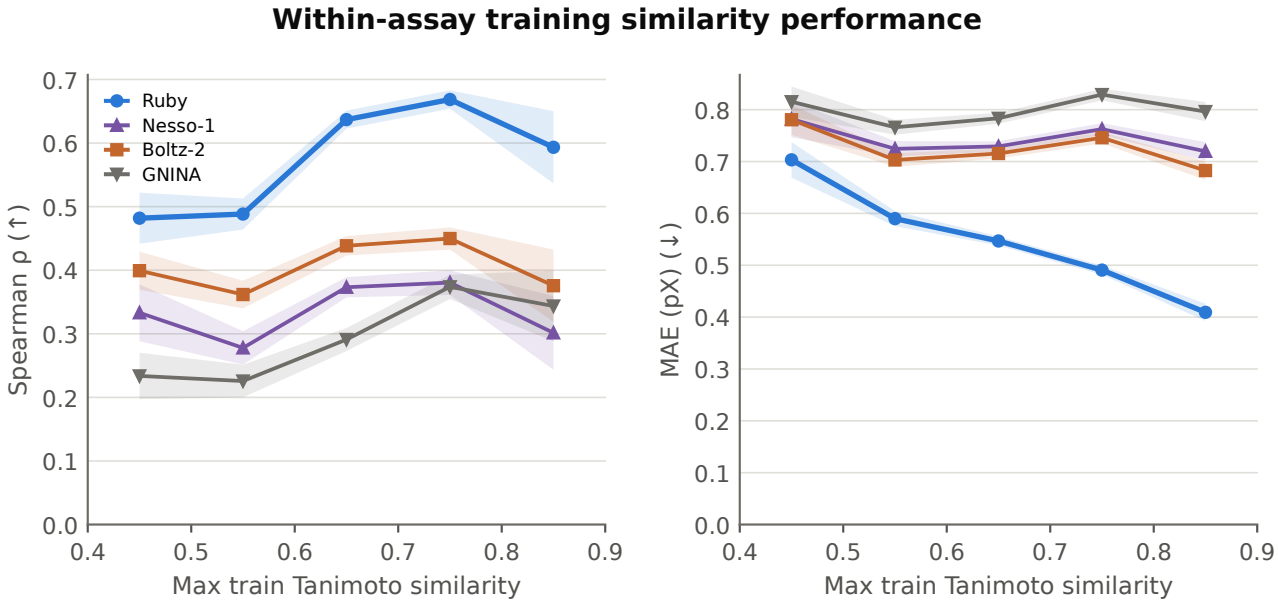


Figure 6. Spearman ρ (left) and MAE (right) in 0.1-wide bins of maximum Tanimoto similarity to earlier exact same-assay measurements, across eight Genesis–Incyte assays on five targets. Both metrics use rolling-temporal recalibrated predictions for every method. Shading shows assay-level bootstrap standard errors. Within-bin rank values omit between-bin comparisons and are not directly comparable with the whole-cohort values in Figure 5.

decreases more dramatically as similarity to prior training data increases. Appendix F.3 gives the binning and rank-score details.

4. Discussion

The discovery regimes introduced above place different demands on a potency predictor: early enrichment over large candidate sets, discrimination among related compounds, and adaptation to the assays that guide an ongoing program. The `GenesisPotency-2026` training set addresses the experimental supervision on which such predictors depend. RUBY uses learned structural representations to make predictions across these settings. Across these evaluations, RUBY's strong performance supports the use of a common pretrained predictor for HitID, H2L, and LeadOp, with computational effort and assay-specific recalibration adjusted to the available resources and experimental history. Together, these results suggest that learned structural representations can support compound prioritization both before substantial on-target measurements are available and as those measurements accumulate.

Beyond supporting RUBY, `GenesisPotency-2026` provides a large, curated training resource that is publicly available to the community. By combining public-source measurements with standardized annotations and source-level auditing, the release makes this experimental supervision more accessible for model development.

Screening at scale. For HitID, prediction quality and computational cost jointly determine how broadly a method can be applied. Under a fixed computational budget, lower inference cost allows more candidates to be scored, while stronger early enrichment helps concentrate active compounds within the subset selected for experimental testing. RUBY's effort sweep exposes this trade-off: lower-effort settings retain useful enrichment

at substantially lower inference cost, whereas higher-effort settings provide stronger enrichment when additional computation is available. At the recorded inference rates, the GPU time required to score 10,000 compounds with the Boltz-2 reference setting would support scoring approximately 12.5 million compounds with RUBY Low. This throughput advantage could substantially expand the chemical space explored within a fixed inference budget. Nvidia kernel optimization provides a complementary improvement, increasing warm inference throughput approximately threefold at fixed model effort on the example workload. These retrospective results motivate prospective evaluation of how RUBY's combination of speed and enrichment translates into hit recovery under fixed computational and experimental budgets.

Ranking within chemical series. The public potency benchmarks in Figure 4 test discrimination among compounds within a target, including the congeneric series relevant to H2L. In this setting, a predictor must resolve activity differences among related molecules to help prioritize which analogs to synthesize and test. On OpenFE, RUBY achieves an aggregate Pearson correlation of 0.74, compared with 0.72 for FEP+, demonstrating strong ranking even before any assay-specific recalibration. This result suggests that learned structural representations can capture information useful for compound prioritization alongside established free-energy methods. Learning from the physical inductive biases of free-energy methods could further improve AI-based potency prediction while retaining the breadth and speed of learned models.

Prediction and adaptation in active programs. The Genesis–Incyte results demonstrate why evaluation on active pharmaceutical programs matters alongside performance on established public benchmarks. Both Boltz-2 and Nesso-1 show lower aggregate Pearson correlations on these programs than on the public potency benchmarks, indicating that their public benchmark performance does not fully transfer to these program settings. RUBY's stronger ranking on the same measurements supports its relevance to the compound-prioritization decisions faced in ongoing discovery. With mean zero-shot Spearman $\rho = 0.61$ and Pearson $r = 0.59$, RUBY provides a stronger starting point for ranking compounds before assay-specific recalibration.

As a program advances into LeadOp, its experimental history becomes increasingly valuable for compound prioritization because earlier measurements describe local structure–activity relationships under the assay conditions used to guide the program. This context matters because biochemical and cellular potency reflect different experimental quantities, and even assays against the same target can yield different measurements [2, 3]. Learning from these observations is well established in QSAR [7] and is also pursued by pretrained models such as ActFound [38] and recent work on project-specific fine-tuning of Boltz-2 [39]. Beyond zero-shot accuracy, usefulness in LeadOp depends on combining knowledge learned across targets with earlier assay measurements to predict subsequent compounds.

Figure 5 evaluates this setting across eight biochemical and cellular assays on five Genesis–Incyte targets. At successive temporal cutoffs, recalibration uses earlier measurements from the same assay to predict compounds measured later, testing performance at multiple stages of each program. RUBY's higher aggregate correlations and lower prediction error support its usefulness for both relative prioritization and estimation of assay-specific potency. In ongoing programs, these capabilities can help prioritize compounds within a series and assess whether their predicted potency, alongside other properties, warrants further testing.

Figure 6 examines how recalibrated performance varies with chemical similarity to earlier compounds measured in the same assay. As expected, RUBY's prediction error is lower in bins with greater similarity to those earlier measurements, and its error declines more across the reported bins than that of the comparison methods. This pattern is consistent with the practical value of local experimental history during LeadOp, where successive compounds often refine an existing series. Evaluating performance within familiar chemistry is therefore important alongside testing generalization to new scaffolds, and rolling temporal evaluation

retains these relationships while restricting recalibration to earlier measurements.

Learning from structural representations. RUBY's strong performance across these regimes supports the use of intermediate cofolding representations for potency prediction. By using information learned through structural training without generating an explicit pose for every compound, RUBY avoids the major portion of inference cost, helping make the approach practical for larger candidate sets. The potency benchmarks demonstrate that useful information for ranking compounds is retained in this setting. Explicit structural models remain valuable for interpreting interactions and developing hypotheses about how chemical modifications affect binding. We therefore expect models like PEARL to increasingly complement potency predictors like RUBY that share intermediate representations, optionally providing structural hypotheses for interpreting structure–activity relationships only for the most interesting compounds.

Limitations. These evaluations are retrospective and measure predictive performance on compounds selected through existing experimental workflows. The HitID benchmarks re-rank experimentally tested selections, so they do not directly establish enrichment over the very large libraries from which those compounds originated. Similarly, rolling temporal evaluation preserves the ordering of assay measurements but does not establish how RUBY-guided choices would change the compounds synthesized or the subsequent course of a program. The training-data audit was not able to access sources for every assay and so has partial source coverage. It samples compounds within documents, so retained records could still contain some errors that were not detected.

Contributions

Writing: Drausin Wulsin, Frank Pickard, Dominique Beaini, Ken Leidal, Steven Kearnes.

Experiments: Michael Atkin, Drausin Wulsin, Jules Berman, Dominique Beaini, Nina Jovic, Gabriel Monteiro da Silva, Allison Keys.

RUBY model: Han Yoon, Jules Berman, Rebecca Boiarsky, Drausin Wulsin, Michael Atkin, Wesley Qian, Will Harvey, William Wang, Ansh Sharma, Michael V. Levine.

PEARL model: Nate Gruver, Jules Berman, Richard Strong Bowen, Alejandro Dobles, Pranav Murugan, Gianluca Scarpellini, David Charles Williams.

`GenesisPotency-2026` **curation:** Han Yoon, Rebecca Boiarsky, Wesley Qian, Adi Nagle, Deidre Johns, Gerald Lelais, Indrawan McAlpine.

Compute infrastructure: Igor Morozov, Han Wang, Cyr Vetcher, Rohit Kundu, Jamin Chen, Wojtek Swiderski.

Leadership: Sergey Edunov, Carl Tilbury, Laura Tapias, Evan Feinberg.

Correspondence may be sent to `papers@genesis.ml`.

Acknowledgements

We are grateful to Nvidia, including Roy Tal Dew and Kushal S. Shah, for their support in optimizing PEARL training and inference, since RUBY benefits from PEARL's structural foundation. Nvidia's kernel optimizations and libraries such as cuEquivariance have also helped us improve RUBY's performance on B200 GPUs, allowing larger screens within a given computational budget.

We are also grateful to Incyte for our ongoing AI-focused research collaboration and for generating and providing the experimental data associated with the results in Section 3.3.

We thank Pat Walters for providing feedback on an early draft of this report.

References

1. Enamine. *Enamine REAL Space* https://enamine.net/compound-collections/real-compounds/real-space-navigator (2026).
2. Kalliokoski, T., Kramer, C., Vulpetti, A. & Gedeck, P. Comparability of Mixed IC50 Data: A Statistical Analysis. *PLOS ONE* **8,** e61007 (2013).
3. Postnikova, E., Cong, Y., DeWald, L. E., *et al.* Testing therapeutics in cell-based assays: Factors that influence the apparent potency of drugs. *PLOS ONE* **13,** e0194880 (2018).
4. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616,** 673–685 (2023).
5. Lam, J. H. & Katritch, V. Navigating structure-based drug discovery with emerging innovations in physics- and knowledge-based approaches. *npj Drug Discovery* **2,** 29 (2025).
6. Wang, Y., Li, Y., Chen, J. & Lai, L. Modeling protein–ligand interactions for drug discovery in the era of deep learning. *Chem. Soc. Rev.* **54,** 11141–11183 (2025).
7. Heid, E. *et al.* Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **64,** 9–17 (2024).
8. Allenspach, S., Hiss, J. A. & Schneider, G. Neural multi-task learning in drug design. *Nat. Mach. Intell.* **6,** 124–137 (2024).
9. Bender, B. J. *et al.* A practical guide to large-scale docking. *Nat. Protoc.* **16,** 4799–4832 (2021).
10. Liu, F. *et al.* The impact of library size and scale of testing on virtual screening. *Nature Chemical Biology* **21,** 1039–1045 (2025).
11. Moesgaard, L., Shoichet, B. K. & Mailhot, O. Modeling the Sensitivity of Large-Scale Virtual Screening to Scoring Function Accuracy, Artifacts, and Library Composition. *J. Chem. Inf. Model.* **66,** 8325–8335 (2026).
12. Muegge, I. & Hu, Y. Recent Advances in Alchemical Binding Free Energy Calculations for Drug Discovery. *ACS Med. Chem. Lett.* **14,** 244–250 (2023).
13. Genesis Research Team. Pearl: A Foundation Model for Placing Every Atom in the Right Location. *arXiv preprint arXiv:2510.24670* (2025).
14. Passaro, S. *et al.* Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv,* 2025.06.14.659707 (2025).
15. Shenoy, N. *et al.* Nesso-1: Accelerating Open-Source Binding Affinity Predictions. *bioRxiv.* Preprint, posted 2026-08-03. https://www.biorxiv.org/content/10.64898/2026.08.01.742196v1 (2026).
16. SandboxAQ. *AQAffinity: Open-Source Structure-to-Affinity Built on OpenFold3 by SandboxAQ* SandboxAQ Blog. Model and code: https://huggingface.co/SandboxAQ/AQAffinity. Jan. 2026. https://www.sandboxaq.com/post/aqaffinity-open-source-structure-to-affinity-built-on-openfold3-by-sandboxaq.
17. Bucher, A. O., Javer, A., Triendl, H. & Hales, I. *Grounding ApherisAffinity predictions in binding structure* Updated 29 October 2026. https://www.apheris.com/resources/grounding-apherisaffinity-predictions-in-binding-structure.

18. Feng, B. *et al.* Hierarchical affinity landscape navigation through learning a shared pocket-ligand space. *Patterns* **6.** The model is named LigUnity. Preprint bioRxiv doi:10.1101/2025.02.17.638554, 101371 (2025).
19. Rossi, M., Pederson, R., *et al.* TerraBind: fast and accurate binding affinity prediction through coarse structural representations. *arXiv,* 2602.07735 (2026).
20. Isomorphic Labs Team. *Accurate Predictions of Novel Biomolecular Interactions with IsoDDE* tech. rep. (Isomorphic Labs, Feb. 2026). https://storage.googleapis.com/isomorphiclabs-website-public-artifacts/isodde_technical_report.pdf.
21. Ackloo, S. *et al.* CACHE (Critical Assessment of Computational Hit-finding Experiments): a public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nature Reviews Chemistry* **6,** 287–295 (2022).
22. Herasymenko, O., Silva, M., Correy, G. J., *et al.* CACHE Challenge #3: Targeting the Nsp3 Macrodomain of SARS-CoV-2. *Journal of Chemical Information and Modeling* **66,** 1566–1581 (2026).
23. CACHE Challenge. *CACHE challenges and results* https://cache-challenge.org/node/ (2026).
24. Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566,** 224–229 (2019).
25. Alon, A. *et al.* Structures of the σ_2 receptor enable docking for bioactive ligand discovery. *Nature* **600,** 759–764 (2021).
26. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research* **52,** D1180–D1192 (2024).
27. Gilson, M. K. *et al.* BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research* **44,** D1045–D1053 (2016).
28. Kim, S. *et al.* PubChem 2025 update. *Nucleic Acids Research* **53,** D1516–D1525 (2025).
29. Baumann, H. M. *et al.* Industry-scale benchmarking of relative binding free energy calculations. *J. Chem. Inf. Model.* **66,** 6429–6452 (2026).
30. Wang, L. *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,** 2695–2703 (2015).
31. Chen, W. *et al.* Enhancing Hit Discovery in Virtual Screening through Absolute Protein–Ligand Binding Free-Energy Calculations. *Journal of Chemical Information and Modeling* **63,** 3171–3185 (2023).
32. Gilson, M. K. *et al.* Assessment of Pharmaceutical Protein–Ligand Pose and Affinity Predictions in CASP16. *Proteins: Structure, Function, and Bioinformatics* **94,** 249–266 (2026).
33. Nelen, J., Khan, O., Adams, E., *et al.* The first OpenBind release: An open experimental structure–affinity dataset and benchmark for structure-based AI. *bioRxiv* (2026).
34. Kearnes, S. & Riley, P. Ordinal Confidence Level Assignments for Regression Model Predictions. *Journal of Chemical Information and Modeling* **64,** 9299–9305 (2024).
35. Liu, S. *et al.* Practical Model Selection for Prospective Virtual Screening. *Journal of Chemical Information and Modeling* **59,** 282–293 (2019).
36. Truchon, J.-F. & Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *Journal of Chemical Information and Modeling* **47,** 488–508 (2007).

37. McNutt, A. T., Li, Y., Meli, R., Aggarwal, R. & Koes, D. R. GNINA 1.3: the next increment in molecular docking with deep learning. *Journal of Cheminformatics* **17,** 28 (2025).

38. Feng, B. *et al.* A bioactivity foundation model using pairwise meta-learning. *Nature Machine Intelligence* **6,** 962–974 (2024).

39. Amini, S., Sciabola, S. & Wang, Y. Affinity Fine-Tuning of Boltz-2: An Open Framework for Protein-Ligand Potency Prediction in Drug Discovery. *bioRxiv.* Preprint (2026).

40. Bento, A. P. *et al.* An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics* **12,** 51 (2020).

41. CASP Prediction Center. *CASP16 Pharmaceutical Ligand Data Archive* https://predictioncenter.org/download_area/CASP16/extra_experiments/pharma_ligands/ (2026).

42. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry* **55,** 6582–6594 (2012).

43. 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* **14,** e0220113 (2019).

44. Buterez, D. *et al.* MF-PCBA: multifidelity high-throughput screening benchmarks for drug discovery and machine learning. *Journal of Chemical Information and Modeling* (2023).

45. Ross, G. A. *et al.* The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Communications Chemistry* **6,** 222 (2023).

46. Anew Labs Team. AnewDDE: An Agentic Drug Discovery Engine for Biomolecular Interaction Modelling and Closed-Loop Design. *bioRxiv* (2026).

Appendix contents

A `GenesisPotency-2026`: source curation and audit **20**

A.1 Sources, eligibility, and label preparation . . . 20

A.2 Source-document audit protocol . . . 20

A.3 Audit outcomes and remaining work . . . 20

A.4 Coverage comparison with AQAffinity . . . 22

A.5 Annotated dataset release . . . 23

B Benchmark-specific training exclusions **23**

C Supporting evaluation methods **25**

C.1 Input preparation and scientific identity . . . 25

C.2 Hit-identification benchmark choice . . . 25

C.3 Public potency cohorts and labels . . . 25

C.4 Partner-program cohorts and recalibration . . . 26

C.5 Rolling temporal evaluation window . . . 26

C.6 Bootstrap details . . . 26

C.7 Comparator recipes and coverage . . . 28

D Hit-identification and runtime results **28**

D.1 Per-target enrichment . . . 29

E Per-series and per-target potency results **30**

E.1 Published per-compound predictions . . . 31

F Partner-program analyses **35**

F.1 Per-assay performance . . . 35

F.2 Paired comparisons . . . 36

F.3 Genesis–Incyte chemical similarity . . . 37

F.4 Data scaling for fine-tuning . . . 37

A. `GenesisPotency-2026`: source curation and audit

A.1. Sources, eligibility, and label preparation

`GenesisPotency-2026` covers the public-source portion of the potency-data curation, drawing on ChEMBL 37, BindingDB release 202607, and PubChem HTS (2025) [26–28]. The full RUBY training index additionally includes proprietary measurements outside this public-source corpus. For the training build used here, we retain measurements dated before 2023-01-01 and group them by document, target, and readout. RUBY and IsoDDE both restrict supervised affinity training to measurements dated before 2023, imposing an explicit temporal restriction absent from the reported training criteria of the other comparison methods in Table 3 [20]. Those methods do not report an affinity-data date cutoff, so their temporal training restrictions cannot be established from the available descriptions. We convert IC_{50}, K_i, and K_d to pX, while EC_{50} contributes binary activity labels only. For numerical potency measurements used for binary supervision, we assign activity at $p\text{X} \geq 6.70$. Eligibility filters exclude ligands with more than 50 heavy atoms, PAINS-flagged compounds, covalent assays, and assays with fewer than 10 measurements. Quality control and sampling reduce the 157 million input rows (Table 1) to 5.9 million potency data points. Most of the reduction is in PubChem HTS, where we keep every active and nine inactives per active, following Boltz-2 [14]. Source records retain their reported relation qualifiers, and censored regression labels are used at their reported bounds in model training. Benchmark-specific target exclusions are described in Appendix B. Measurements may also supply more than one task, so row counts do not represent independent experiments.

A.2. Source-document audit protocol

We reviewed source documents in two passes. The document-level pass checked each assay group's target, species, construct, origin of reported values, endpoint, and units. The compound-level pass examined up to 15 compounds per document, comparing structures with drawings, systematic names, and substituent tables, and comparing values and qualifiers with the reported measurements. In total, this pass examined 115,235 sampled compounds from 8,995 documents. The two passes took approximately 160 and 700 agent-hours, respectively.

Full text was available for 5,329 patents and 3,666 articles. For 19,746 paywalled articles, only abstracts were available. Rejection decisions required a contradictory source passage. Code checked the quotation against the document, and an independent reviewer assessed the decision without seeing the first review. Structured reports recorded the fields checked and the proposed actions. We removed the affected assay group if we found a document-level contradiction or errors in at least 20% of sampled compounds. Records were marked for correction when the source supported a specific replacement and for removal when it did not. The audit does not cover PubChem HTS records, and it does not imply that every compound within an audited document was inspected.

A.3. Audit outcomes and remaining work

Of 2.51 million ChEMBL and BindingDB source records in scope, 1.24 million belonged to auditable documents. The audit retained 699,742 unchanged, marked 52,575 for correction, and removed 486,600 (Table 1). Of the 5.90 million pre-audit potency data points (Figure 1(b)), 8.4% were removed or withheld, 0.02% were corrected, and 41.9% passed source-document review, including partial passes: 13.6% against the full text and 28.3% against the abstract or supplement alone. Another 2.6% are ChEMBL and BindingDB data points without an accessible source or with an unresolved review. PubChem HTS screening calls (43.7%), which

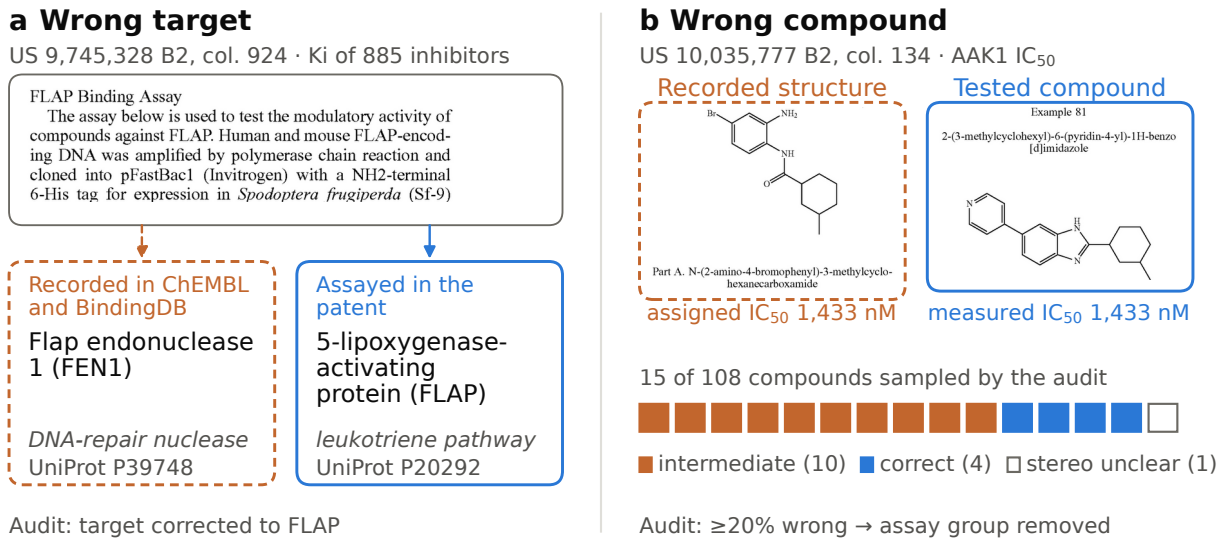


Figure 7. Source-document evidence for the two audit examples, cropped from the issued patents. In each panel, the database record (orange) is shown left of the source (blue). (a) Patent US9745328 assays FLAP, but ChEMBL and BindingDB assigned the K_i values of its 885 inhibitors to flap endonuclease 1 (FEN1). (b) For Example 81 of patent US10035777, both databases record the synthetic intermediate drawn below the tested compound. Ten of the 15 compounds sampled from this patent had the same error.

screening centers deposit directly, and synthetic decoys (3.4%) have no source document and pass without review. Of 486,600 removed records, 338,718 belonged to assay groups excluded because at least 20% of sampled compounds had errors. Other reasons included values not measured in the cited document, activity classes or thresholds treated as measured values, and mismatched targets, constructs, endpoints, or units. Figure 7 shows the two examples from the main text in the issued patents. In patent US9745328, the binding assay measures 5-lipoxygenase-activating protein (FLAP), but both databases assigned the K_i values of its 885 inhibitors to flap endonuclease 1. The audit remapped these records to FLAP. These records are withheld from the frozen training build because the correction requires a new protein input. In patent US10035777, the record for Example 81 depicts the synthetic intermediate drawn below the tested title compound and carries the title compound's AAK1 IC_{50}. Ten of the 15 sampled records had the same error, so the audit removed the assay group. Both errors originated in BindingDB and reached ChEMBL through its import of BindingDB patent data.

The audit covered the ChEMBL and BindingDB records of a broader curation build without the 2023 date cutoff, so its scope includes records outside the training eligibility filters.

Of 52,575 source records marked for correction, 49,575 require new protein or ligand inputs: 31,585 target remappings, 17,384 sequence changes for stated point mutations, and 606 ligand-structure corrections. The frozen training build applies supported corrections where the required inputs are available and withholds records requiring new inputs pending their preparation. Records from inaccessible documents, partial reviews, and documents that were ambiguous on a deciding point were retained, so the training corpus

Table 1. Public-source inputs and source-document audit outcomes for `GenesisPotency-2026`. The middle block counts the potency data shown in Figure 1(c). Synthetic decoys count toward the source of their assay. The lower block counts ChEMBL and BindingDB source-record audit decisions, a different population and unit from the input rows and potency data. The audit covered the ChEMBL and BindingDB records of a broader curation build without the 2023 date cutoff, so its scope includes records outside the training eligibility filters. Partial or unavailable review includes records for which only abstract-level checks were possible. Retained unchanged includes 21,844 records whose source document was read but was ambiguous on a deciding point. PubChem HTS was outside the source-document audit (—), so audit totals cover ChEMBL and BindingDB only.

Stage	ChEMBL	BindingDB	PubChem HTS	Total
Available input rows	24,527,044	3,228,554	129,334,526	157,090,124
Potency data before audit	1,356,596	1,965,221	2,583,180	5,904,997
Potency data after audit	1,167,510	1,659,926	2,583,180	5,410,616
Records in audit scope	1,004,926	1,504,456	—	2,509,382
Partial or unavailable review	576,473	693,992	—	1,270,465
Auditable source records	428,453	810,464	—	1,238,917
Retained unchanged	240,362	459,380	—	699,742
Marked for correction	17,516	35,059	—	52,575
Removed	170,575	316,025	—	486,600

contains both source-checked and unverified records. The public-source corpus is distinct from proprietary training records and from the papers and patents consulted during the audit.

A.4. Coverage comparison with AQAffinity

We compare `GenesisPotency-2026` with the 1,415,051 training rows of AQAffinity's release, excluding its validation rows [16], and process both sets identically. Compounds are identified by the InChIKey of their ChEMBL Structure Pipeline parent [40], and proteins are identified by the UniProt accession of their sequence. PubChem HTS constructs without a full-length match are resolved through UniParc, PDB chain mappings, or UniProt identifier mapping. Each (protein, compound) pair is counted once across assays, sources, and label types, with PubChem HTS included in both sets. Target groups follow the ChEMBL protein classification [26], with our own family rollup for the 13% of proteins it does not classify. An AQAffinity row matches an audit outcome when its compound, protein, and label agree with a pre-audit `GenesisPotency-2026` data point within 0.01 log units. Because the audit removes whole assay groups, a match marks a measurement from a rejected group, not a verified error.

Compared with AQAffinity, `GenesisPotency-2026` has 1.16 times as many proteins (3,110 versus 2,691), 1.23 times as many compounds (1.00 versus 0.82 million), and 2.78 times as many (protein, compound) pairs, mostly from PubChem HTS. For pairs with a pX label, the ratio is 1.24. Ion channels gain the most proteins (2.5 times), and kinases have 1.5 times as many pairs (Table 2).

Neither set contains the other: 655 proteins, 0.25 million compounds, and 0.32 million pX pairs appear only in AQAffinity, partly because `GenesisPotency-2026` excludes benchmark targets and data dated after 2022. About 5% of AQAffinity's data points (68,678) match a measurement our agentic audit removed or marked for correction, so auditing AQAffinity the same way may be worthwhile.

Table 2. Proteins and (protein, compound) pairs per ChEMBL target group in AQAffinity and `GenesisPotency-2026`. Rows follow Figure 1(d), whose "Other" bar combines the last five target groups listed here. The last column counts proteins present in both datasets.

	AQAffinity		`GenesisPotency-2026`		Both
Target group	Proteins	Potency data	Proteins	Potency data	Proteins
GPCR	456	300,203	481	315,592	381
Kinase	439	216,064	475	324,379	387
Protease	250	117,614	276	170,472	194
Ion channel	68	14,078	169	62,766	62
Epigenetic	103	35,646	102	638,164	80
Transporter	93	36,437	122	101,345	76
Nuclear receptor	65	54,577	63	241,087	46
Other enzyme	237	95,807	336	248,811	184
Hydrolase & PDE	334	99,718	323	270,245	205
Oxidoreductase	237	47,927	274	622,846	180
Transferase	206	79,641	253	200,767	147
Other	203	152,051	236	274,150	94
All groups	2,691	1,249,763	3,110	3,470,624	2,036

A.5. Annotated dataset release

We are releasing the audited public-source training dataset as `GenesisPotency-2026` for academic and commercial use under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license, with annotations attached to its rows. The license permits reuse, redistribution, and adaptation, subject to attribution, a link to the license, and identification of changes. Adapted material that is shared must be distributed under the same or a compatible license. Release documentation and version history will be maintained in the GenesisPotency GitHub repository, with dataset artifacts hosted on Hugging Face. Each released row identifies its source provenance, audit status, and applied corrections, distinguishing document-level review from direct checking of a sampled compound and identifying partial reviews or unavailable source evidence. The annotations link rows to their source records and audit decisions, since a source record can contribute to more than one task. Original endpoints, values, units, relation qualifiers, and normalized labels preserve the context needed to interpret the supervision. Removed rows and corrections awaiting input preparation are not represented as usable, corrected data points. The release contains the redistributable public-source portion of the corpus with source attribution.

B. Benchmark-specific training exclusions

Following Boltz-2 [14] and IsoDDE [20], we exclude a training target when its sequence aligns to an OpenFE, FEP+ 4 or hit-identification target with at least 90% identity and the alignment covers at least 50% of either sequence. CASP16 and OpenBind are deleaked implicitly by the date cutoff. We remove no program targets.

The published descriptions behind Table 3 state:

> **IsoDDE:** "cutoff date of 01-January-2023" and "we removed any target with a sequence identity of 90% or greater to either of the sets." [20, Secs. 2.2 and 4.5]

Table 3. Reported benchmark-specific eligibility of supervised affinity training observations. Protein < 90% means no qualifying high-identity match under that method's coverage rule. "No exclusion" means no panel-specific target exclusion, not confirmed training exposure. Label dates refer to supervised affinity or activity records, not protein structures. A dash means the criterion was not specified.

Model	Label date	Target similarity				
		HitID (7)	OpenFE / FEP+ 4	CASP16	OpenBind	Genesis–Incyte
RUBY	Before 2023	Protein < 90%	Protein < 90%	No exclusion	No exclusion	No exclusion
IsoDDE [20]	Before 2023	—	Protein < 90%	No exclusion	—	—
Boltz-2 [14]	—	—	Outside ≥ 90% test clusters	No exclusion	—	—
Nesso-1 [15]	—	—	Protein < 90%	—	—	—
TerraBind [19]	—	—	—	—	—	—
AQAffinity [16]	—	—	—	—	—	—
LigUnity [18]	—	—	—	—	—	—

Boltz-2: "training protein that falls into a cluster shared with a validation or test protein" and "except for CASP16 and the Recursion internal assays" [14, App. D.2.1].

Nesso-1: "removing any training datapoint with ≥ 90% protein sequence similarity to the FEP+ benchmark" [15, Sec. 2.1]

TerraBind (structure data): "All structures from the Protein Data Bank (PDB)" and "released before 2021-09-30" [19, Sec. 2.2.2]

AQAffinity release: "the L3000 target has high similarity with at least one protein target in the training set (i.e. single point mutation)." [16]

LigUnity: "assays similar to those included in the FEP benchmarks are excluded in the pre-training stage" and "Ligands in the FEP benchmarks are also removed" [18]

CASP16 released the L1000 and L3000 ligand targets in May 2024, and the archive dates their experimental affinity files to November–December 2024 [41]. This post-cutoff benchmark release does not preclude earlier measurements on the same protein or compounds in training.

For Boltz-2, MMseqs clusters training and test proteins using at least 90% identity over at least 1% of the longer sequence, then removes training proteins in a cluster containing a test protein [14, App. D.2.1]. The authors exempt CASP16 from this protein filter, stating that its data were released after training, but they do not give a complete affinity-label date cutoff. RUBY instead removes direct alignment matches covering at least half of either sequence. TerraBind's quoted cutoff applies to structural PDB data, but its paper does not give a corresponding cutoff or CASP16 target-exclusion rule for affinity labels [19]. LigUnity's benchmark-ligand and matched-assay holdouts apply to its narrower JACS/Merck panel [18]. No global benchmark-ligand holdout is reported for the other methods. AQAffinity reports a training target one residue away from CASP16 L3000 [16].

C. Supporting evaluation methods

C.1. Input preparation and scientific identity

We key benchmark identity to the source compound's InChIKey and carry the prepared molecular state for inference separately. This avoids treating two protonation states as independent experimental compounds. Frozen prepared-source records bind each scored observation to its protein input, ligand state, and cached representation. Joins require exact identity coverage and reject conflicting structures or nonfinite scores. Protein sequences and any selected pocket constraints or structural templates come from each benchmark's prepared inputs, and programs use their production input configurations.

C.2. Hit-identification benchmark choice

DUD-E pairs known actives with computationally selected decoys, whose selection can introduce chemical biases that models exploit without learning target-specific interactions [42, 43]. MF-PCBA uses experimental HTS measurements [44], but compounds recurring across public screening datasets may also occur in a model's training data, even after target-level exclusions, motivating an explicit audit of compound overlap. Enrichment within a fixed HTS library can also have limited practical value when the collection is screened in bulk, since ranking compounds does not reduce the number tested. Prioritization is more directly actionable when experimental capacity limits which compounds can be tested or advanced to follow-up.

C.3. Public potency cohorts and labels

Cohort membership. OpenFE contains 873 observations from 58 source series spanning 39 targets, and each series is scored separately. Compound identity is unique within each source series, and observations of the same compound in different source series are retained. FEP+ 4 contains 87 observations across CDK2, JNK1, p38 and TYK2, with one series per target. CASP16 contains 123 autotaxin and 17 chymase observations, giving 1,100 observations across 64 source series in total [29–32]. Correlations are computed within source series and averaged within each benchmark using compound-count weights, as specified in Section 2.5. Table 7 gives series-level counts and results.

Label sources and conversion. For OpenFE and FEP+ 4, we use the 873 and 87 `pchembl_value` labels, respectively, from the pinned IsoDDE benchmark extraction [20]. The extraction is fixed at revision `7e6077bd` of the benchmark data repository, and source-series and compound identities are matched before scoring. These supplied pX values are retained without an additional energy-to-concentration conversion. The FEP+ 4 values are verified against $pX = -\log_{10}(C)$ using the recorded concentrations and units. For CASP16, we instead use the organizers' original IC_{50} measurements, with $pX = 6 - \log_{10}(IC_{50}/\mu M)$. The nine autotaxin observations at the 10 µM assay ceiling are marked as censored and enter correlation calculations at the boundary value $pX = 5$. The autotaxin protein input corresponds to the rat construct.

OpenBind. We retain the full 494-entry affinity benchmark after the source's measurement quality control, using the released experimental pK_D values, where $pK_D = -\log_{10}(K_D)$ [33]. These measurements were obtained against CVA16 2A protease, the experimental surrogate for EV-A71 2A protease. Released comparator predictions are matched by structure identifier to the same 494 benchmark entries, and their experimental labels are checked against our frozen cohort. For methods with per-compound predictions, Spearman ρ and its bootstrap uncertainty use this full cohort. Cited aggregate results are identified separately in Table 8.

C.4. Partner-program cohorts and recalibration

The Genesis–Incyte evaluation uses shared scored rolling cohorts totalling 22,036 assay–compound records across eight assays and five targets. Biochemical and cellular measurements remain separate tasks, and a compound measured in both can contribute to each task. Aggregate results give every assay equal weight.

Correlations include censored measurements at their reported pX bounds. Zero-shot correlations use the full scored cohorts. Recalibrated correlations use the intersection of finite rolling predictions across all methods. RUBY and comparison methods are scored on the same subset within each setting.

At each rolling cutoff, assay-specific recalibration uses only earlier exact measurements from that assay, and the base prediction models remain fixed. We give comparison methods a separate affine fit at each cutoff. For every method, recalibrated Pearson r, Spearman ρ and Kendall τ_b are computed after pooling the eligible rolling predictions within each assay. Kendall τ_b accounts for ties in both measured and predicted values. Fold-specific slopes and intercepts can change the pooled correlations, even when each individual fit has a positive slope.

C.5. Rolling temporal evaluation window

Related time-split validation has been used to evaluate prediction confidence on internal pharmaceutical assay data [34]. We construct folds separately for each of the eight Genesis–Incyte assays. Figure 8 illustrates the split and per-window recalibration. Let t_k be a fold cutoff, advancing in seven-day steps from the assay's earliest exact compound date. We assign each compound the date of its earliest exact measurement, or its earliest censored measurement if none is exact. The fold's training history contains compounds measured by t_k, and its scored window contains compounds whose assigned date falls in $[t_k + 28\,\mathrm{d},\ t_k + 35\,\mathrm{d})$. Thus each prediction uses a 28-day gap after its training cutoff, and the seven-day windows do not overlap. We begin scoring only once at least 20 compounds have dated, uncensored measurements in the training history. The fold grid is determined from uncensored dates. Compounds with only censored measurements may enter a scored window, but they do not enter the exact-label recalibration fits.

We pool the eligible scored windows for each assay and evaluate every method on a shared eligible cohort for each metric and setting. Each compound contributes once, at its first eligible window. At each cutoff, models are recalibrated using only the assay history available by that date. Zero-shot correlations use the full shared scored cohort.

C.6. Bootstrap details

For each partner-program metric, we select the comparator with the best aggregate point estimate and retain that comparator throughout the paired bootstrap. We draw 1,000 bootstrap samples with replacement within each target or assay, retaining its observed sample size. For paired comparisons, we share compound draws across methods. For public targets containing several source series, we resample compounds within each source series before pooling, keeping series sizes fixed. This preserves pairing with historical source-series summaries. Predictions, including rolling recalibration fits, remain fixed throughout resampling. The resulting uncertainty is conditional on the evaluated targets, assays and fitted models. The RUBY (Max) setting used for the hit-identification enrichment curve was selected after inspecting its nEF@1% score. The associated paired bootstrap p-values condition on this selection and do not adjust for it.

Within each paired bootstrap draw, we define RUBY's advantage as RUBY minus comparator for correlations

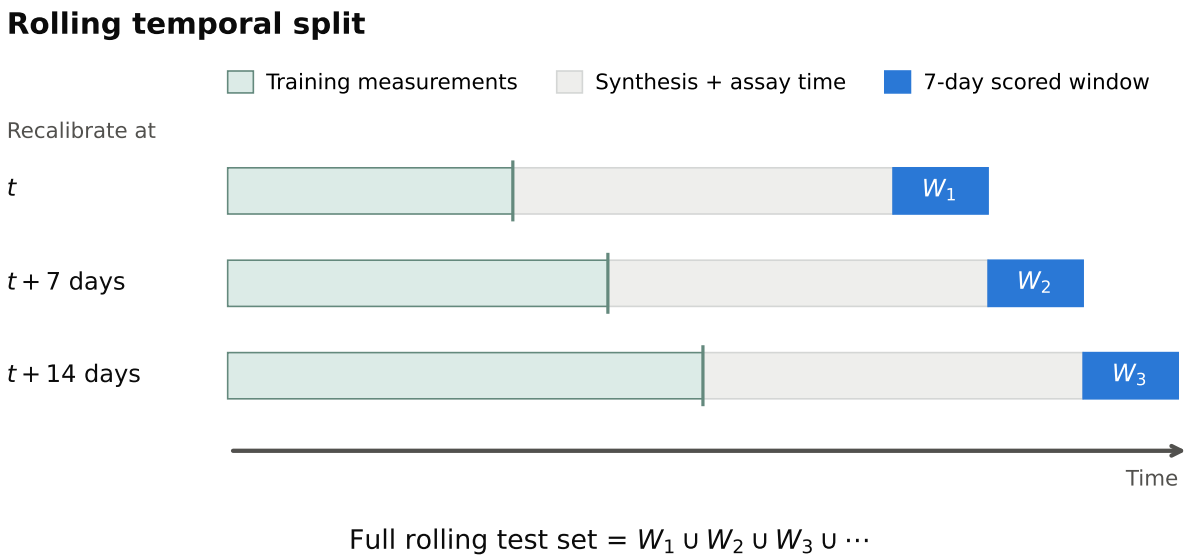


Figure 8. Rolling temporal evaluation within one assay. Each row shows a fold: recalibration is refit using only exact training measurements available at its cutoff, while the base prediction model remains fixed. A 28-day synthesis and assay interval separates the cutoff from the seven-day scored window. The cutoff advances by seven days, expanding the training history. The full rolling temporal test set is the union of the scored windows, $\bigcup_k W_k$. Each window uses its own recalibration fit, and each compound contributes once. Three consecutive eligible folds are shown schematically, with earlier history truncated. Bar widths indicate elapsed time, not compound counts.

and comparator minus RUBY for MAE. We report the one-sided bootstrap tail fraction p as the fraction of valid draws with negative advantage, counting ties as half. Stars indicate $p \leq 0.05$ (*), $p \leq 0.01$ (**), and $p \leq 0.001$ (***). Appendix Table 10 reports the paired comparisons.

For aggregates computed from bootstrap draws, we apply the point estimate's fixed membership and weights to each draw: equal target weights for hit identification, compound-count weights for public-potency correlations, and equal assay weights for partner programs. The bootstrap standard error is the sample standard deviation of the valid aggregate draws. An aggregate draw is invalid whenever any member correlation is undefined. We change neither membership nor weights to rescue a draw. Where aggregate error bars are propagated from independently bootstrapped assay estimates, we use $\mathrm{SE}_{\mathrm{mean}} = \sqrt{\sum_{a=1}^{A} \mathrm{SE}_a^2}/A$ for A equally weighted assays. This propagation assumes independence across assays and does not account for dependence when assays share compounds.

For the chemical-similarity analysis in Figure 6, we resample compounds separately within each assay and similarity bin, using the metric-specific eligible cohort. These draws are not paired across methods. For each assay and bin, we estimate one standard error as the width of the central 95% bootstrap percentile interval divided by 2×1.96. Undefined correlation draws are excluded, and an uncertainty estimate requires at least 500 valid draws. We propagate these assay-level estimates to the equal-assay mean using the formula above, and the shading shows the mean plus or minus this estimated standard error.

C.7. Comparator recipes and coverage

Where every comparator number comes from. GNINA has no complete CASP16 aggregate because the frozen benchmark carries no receptor geometry for its 17 chymase compounds. We omit that aggregate rather than report autotaxin alone as the full benchmark. FEP+ and OpenFE RBFE appear only on the two benchmarks for which published predictions are available. We assume that IsoDDE's CASP16 cohort matches ours from its stated use of the Boltz-2 protocol, but the available report does not establish an exact compound match (Appendix E.1). TerraBind's CASP16 aggregate is our compound-weighted derivation from its two published target values. ApherisAffinity contributes cited results for all four public-potency benchmarks, and Nesso-1's OpenBind value is also cited rather than re-scored. OpenBind's other comparator prediction sources are detailed in Table 8. All partner-program comparator predictions were generated within the partner evaluations.

Configuration. Table 4 gives the versions and settings of our public screening and OpenFE, FEP+ 4 and CASP16 runs. The OpenBind results re-scored from published predictions retain their source inference settings. Partner-specific preparation and inference records are linked from `paper/RUN_PROVENANCE.md`.

Table 4. Comparator versions and inference settings for runs generated here on the seven screening targets and the OpenFE, FEP+ 4 and CASP16 benchmarks. A dash in the Method column continues the method above. Settings may differ from those used for published OpenBind predictions or partner evaluations.

Method	Version	Setting	Value
Boltz-2 [14]	2025 release	Models	Structure and affinity checkpoints
–		Hit identification	3 recycles and 2 structure samples
–		Public potency	5 recycles and 5 structure samples
–		Shared sampling	200 structure steps and 5 affinity samples
–		Output	Molar pX conversion and molecular-weight correction
Nesso-1 [15]	v1.0.0	Model assets	Released affinity ensemble
AQAffinity [16]	v0.0.1 (2026)	Models	OpenFold3 and affinity checkpoints
–		Sampling	1 seed and 5 diffusion samples
–		Conditioning	Templates enabled
LigUnity [18]	2025 release	Hit identification	Dedicated screening checkpoint
–		Potency and programs	5 pocket-ranking and 5 protein-ranking checkpoints
GNINA [37]	v1.3.3, locally modified	Search	9 starting conformers, 10 poses, exhaustiveness 1, seed 42
–		Receptor box	Reference-ligand autobox with 6 Å padding
–		Scoring	Negative empirical docking energy, best pose/state per compound
–		Limits	600-s docking and 150-s conformer timeouts and 0.25 Å pose deduplication
–		Clash handling	Weight 1, severe penetration 0.2, multiplier 3, maximum 0.5 and polar-hydrogen refinement

D. Hit-identification and runtime results

Table 5 gives relevant published reference timings for different workloads and hardware, with a citation for each claim.

Table 5. Published timing claims for other workloads and hardware. These values are provided as context and are not points in Figure 2.

Method	Reported timing	Workload and source
Boltz-2	20 GPU-s / compound	FEP+ 4-target subset. GPU unstated. Boltz-2 paper, Supp. Table 12 [14]
Nesso-1	1.0-2.7 s / compound	One H100, 300-1000 residues. Embedding load excluded. Nesso-1 paper
LigUnity	2.4 ms / compound	RTX A6000, 10k ligands. Ligand embeddings included. LigUnity paper, Fig. S7
AnewAffinity	3.4 s / ligand	Amortized over the FEP+ 4-target set. GPU unstated. AnewAffinity paper
GNINA 1.3	16.8 s / compound	RTX 3080, default CNN ensemble, PDBbind core. GNINA 1.3 paper
OpenFE / FEP+	3-12 GPU-h / edge	one RBFE transformation. OpenFE / FEP+ published per-edge costs

D.1. Per-target enrichment

Table 6. Per-target hit identification: n and n_{active} count scored compounds and positive labels. nEF at 0.2%, 0.5%, 1%, 2%, 5%, and 10%, and BEDROC at $\alpha = 80.5$ and $\alpha = 20$ show point estimates ± one bootstrap standard error. The opening block gives equal-target means of the per-target nEF values, and BEDROC is reported per target only. Bold marks the best method in each metric row, compared as printed.

Target	n	n_{active}	Metric	*RUBY*	*Nesso-1*	*Boltz-2*	*GNINA*	*LigUnity*	*AQAffinity*	*Mol. weight*
equal-target mean	9,364	1,008	nEF@0.2%	**0.59**±0.11	0.25±0.09	0.21±0.11	0.07±0.05	-0.16	0.32±0.13	-0.01
			nEF@0.5%	**0.50**±0.07	0.17±0.08	0.30±0.09	0.11±0.05	-0.08	0.08±0.08	0.15
			nEF@1%	**0.43**±0.06	0.18±0.06	0.23±0.07	0.13±0.06	-0.03±0.06	0.07±0.07	0.15±0.06
			nEF@2%	**0.40**±0.04	0.20±0.05	0.22±0.05	0.11±0.04	0.01	0.13±0.05	0.15
			nEF@5%	**0.31**±0.03	0.18±0.03	0.22±0.03	0.15±0.03	0.00±0.03	0.13±0.03	0.19±0.03
			nEF@10%	**0.27**±0.03	0.24±0.03	0.21±0.02	0.13±0.03	0.01	0.15±0.03	0.16
			$\text{BEDROC}_{80.5}$	–	–	–	–	–	–	–
			BEDROC_{20}	–	–	–	–	–	–	–
`ampc`	1,447	298	nEF@0.2%	0.58±0.40	**1.00**±0.26	0.58±0.36	-0.26±0.02	-0.26±0.07	0.16±0.39	-0.26±0.02
			nEF@0.5%	**0.82**±0.19	0.60±0.20	0.46±0.24	-0.26±0.05	-0.26±0.12	-0.08±0.18	-0.26±0.08
			nEF@1%	**0.73**±0.15	0.62±0.15	0.46±0.17	-0.17±0.08	-0.17±0.09	-0.08±0.11	-0.08±0.15
			nEF@2%	0.48±0.12	0.50±0.13	**0.57**±0.11	-0.22±0.06	-0.09±0.07	-0.04±0.10	0.17±0.11
			nEF@5%	**0.49**±0.07	0.43±0.07	0.41±0.08	-0.03±0.06	-0.15±0.04	0.06±0.07	0.14±0.07
			nEF@10%	0.44±0.05	**0.46**±0.05	0.29±0.05	-0.02±0.04	-0.07±0.04	0.08±0.04	0.11±0.05
			$\text{BEDROC}_{80.5}$	**0.67**±0.08	0.63±0.08	0.58±0.08	0.08±0.04	0.09±0.04	0.19±0.06	0.25±0.06
			BEDROC_{20}	**0.57**±0.04	0.56±0.04	0.48±0.04	0.17±0.03	0.14±0.03	0.25±0.03	0.32±0.04
`d4_drd4`	548	80	nEF@0.2%	**1.00**±0.50	**1.00**±0.34	-0.17±0.36	-0.17±0.14	-0.17±0.49	**1.00**±0.55	-0.17±0.35
			nEF@0.5%	**0.61**±0.32	**0.61**±0.31	0.22±0.33	-0.17±0.18	0.22±0.33	0.22±0.35	0.22±0.29
			nEF@1%	**0.77**±0.24	0.59±0.25	0.30±0.27	0.06±0.18	0.30±0.25	0.30±0.26	0.06±0.22
			nEF@2%	**0.68**±0.21	0.41±0.17	0.15±0.17	-0.06±0.12	0.15±0.16	0.25±0.17	0.04±0.14
			nEF@5%	**0.31**±0.11	0.26±0.10	0.26±0.11	0.00±0.09	0.09±0.09	0.09±0.10	0.09±0.09
			nEF@10%	**0.25**±0.07	**0.25**±0.07	0.23±0.07	0.00±0.05	0.11±0.06	0.06±0.06	0.04±0.06
			$\text{BEDROC}_{80.5}$	**0.62**±0.13	0.52±0.13	0.31±0.13	0.11±0.08	0.27±0.12	0.33±0.13	0.18±0.11
			BEDROC_{20}	**0.45**±0.07	0.39±0.07	0.34±0.07	0.15±0.05	0.24±0.06	0.24±0.06	0.19±0.05
`sigma2_tmem97`	495	128	nEF@0.2%	**1.00**±0.15	-0.35±0.39	**1.00**±0.42	**1.00**±0.14	-0.35±0.18	-0.35±0.24	-0.35±0.61
			nEF@0.5%	**1.00**±0.16	-0.01±0.36	**1.00**±0.42	**1.00**±0.20	-0.35±0.23	-0.35±0.30	0.33±0.49
			nEF@1%	**0.73**±0.21	0.12±0.30	0.46±0.32	**0.73**±0.28	-0.08±0.31	-0.08±0.31	0.46±0.31
			nEF@2%	**0.87**±0.13	0.39±0.21	0.33±0.22	0.60±0.20	0.19±0.22	0.33±0.23	0.19±0.22
			nEF@5%	**0.78**±0.11	0.37±0.14	0.46±0.12	0.46±0.14	0.14±0.13	0.41±0.13	0.35±0.13
			nEF@10%	**0.70**±0.08	0.45±0.09	0.46±0.09	0.27±0.09	0.19±0.09	0.30±0.09	0.35±0.09
			$\text{BEDROC}_{80.5}$	**0.89**±0.08	0.45±0.14	0.59±0.14	0.73±0.12	0.30±0.12	0.39±0.14	0.46±0.14
			BEDROC_{20}	**0.79**±0.05	0.56±0.07	0.60±0.07	0.53±0.07	0.38±0.07	0.49±0.07	0.50±0.07

continued

Target	n	n_{active}	Metric	*RUBY*	*Nesso-1*	*Boltz-2*	*GNINA*	*LigUnity*	*AQAffinity*	*Mol. weight*
cache1_lrrk2_wd40	1,618	101	nEF@0.2%	**0.64**±0.35	-0.07±0.01	-0.07±0.12	-0.07±0.15	-0.07±0.01	-0.07±0.18	-0.07±0.20
			nEF@0.5%	**0.20**±0.19	-0.03±0.04	0.07±0.12	0.07±0.12	-0.07±0.01	**0.20**±0.16	0.07±0.13
			nEF@1%	**0.13**±0.11	0.00±0.06	0.00±0.07	0.07±0.09	-0.07±0.01	**0.13**±0.11	0.07±0.10
			nEF@2%	**0.17**±0.08	0.03±0.05	0.07±0.06	0.13±0.08	-0.03±0.03	0.13±0.07	0.07±0.07
			nEF@5%	**0.16**±0.05	0.04±0.03	0.07±0.04	0.09±0.04	-0.05±0.01	0.08±0.04	0.14±0.05
			nEF@10%	**0.16**±0.04	0.06±0.04	0.14±0.04	0.14±0.05	-0.09±0.02	0.14±0.04	0.15±0.05
			$\text{BEDROC}_{80.5}$	**0.22**±0.07	0.08±0.03	0.11±0.05	0.16±0.06	0.01±0.01	0.18±0.06	0.16±0.05
			BEDROC_{20}	**0.23**±0.04	0.14±0.03	0.19±0.03	0.20±0.04	0.02±0.01	0.20±0.04	0.21±0.04
cache2_nsp13	1,954	155	nEF@0.2%	-0.09±0.11	-0.02±0.08	-0.09±0.14	**0.19**±0.23	-0.09±0.01	**0.19**±0.21	**0.19**±0.22
			nEF@0.5%	0.02±0.12	0.00±0.06	**0.24**±0.16	0.13±0.13	-0.09±0.01	0.02±0.11	0.13±0.13
			nEF@1%	0.08±0.09	0.00±0.06	**0.24**±0.11	0.02±0.08	-0.09±0.01	0.02±0.08	0.13±0.11
			nEF@2%	0.05±0.06	0.03±0.05	**0.25**±0.08	0.11±0.07	-0.09±0.01	0.05±0.06	**0.25**±0.08
			nEF@5%	0.00±0.03	0.12±0.04	0.18±0.05	0.16±0.04	-0.08±0.01	0.06±0.04	**0.20**±0.05
			nEF@10%	0.01±0.03	0.17±0.03	0.18±0.03	0.14±0.04	-0.10±0.01	0.11±0.03	**0.24**±0.04
			$\text{BEDROC}_{80.5}$	0.11±0.05	0.12±0.04	**0.26**±0.06	0.18±0.05	0.00±0.00	0.13±0.05	0.25±0.06
			BEDROC_{20}	0.11±0.03	0.22±0.03	0.27±0.03	0.24±0.03	0.02±0.01	0.18±0.03	**0.31**±0.04
cache3_mac1	1,661	229	nEF@0.2%	**1.00**±0.20	0.23±0.31	0.23±0.32	-0.16±0.06	-0.16±0.13	0.61±0.36	0.61±0.35
			nEF@0.5%	**0.86**±0.19	0.02±0.14	0.13±0.20	-0.01±0.12	-0.01±0.14	0.28±0.22	0.57±0.22
			nEF@1%	**0.59**±0.15	-0.06±0.08	0.18±0.14	0.11±0.13	-0.09±0.08	0.11±0.14	0.39±0.14
			nEF@2%	**0.51**±0.10	0.04±0.07	0.16±0.09	0.12±0.09	-0.02±0.07	0.05±0.09	0.37±0.10
			nEF@5%	**0.34**±0.06	0.02±0.05	0.06±0.05	0.15±0.06	0.08±0.05	0.08±0.05	0.30±0.07
			nEF@10%	**0.25**±0.04	0.10±0.03	0.07±0.04	0.17±0.04	0.04±0.03	0.13±0.04	0.13±0.04
			$\text{BEDROC}_{80.5}$	**0.57**±0.08	0.14±0.05	0.25±0.07	0.23±0.06	0.13±0.05	0.24±0.07	0.46±0.08
			BEDROC_{20}	**0.40**±0.04	0.21±0.03	0.22±0.03	0.28±0.04	0.18±0.03	0.24±0.03	0.33±0.04
cache4_cblb_tkb	1,641	17	nEF@0.2%	-0.01±0.00	-0.01±0.00	-0.01±0.00	-0.01±0.02	-0.01±0.00	**0.66**±0.35	-0.01±0.00
			nEF@0.5%	-0.01±0.00	-0.01±0.00	-0.01±0.00	-0.01±0.06	-0.01±0.00	**0.24**±0.17	-0.01±0.00
			nEF@1%	-0.01±0.00	-0.01±0.00	-0.01±0.03	0.05±0.07	-0.01±0.00	**0.12**±0.09	-0.01±0.00
			nEF@2%	0.04±0.05	-0.01±0.02	0.04±0.07	**0.10**±0.09	-0.02±0.00	**0.10**±0.08	-0.02±0.03
			nEF@5%	0.07±0.09	0.06±0.07	0.07±0.08	**0.20**±0.11	-0.05±0.00	0.13±0.10	0.07±0.09
			nEF@10%	0.09±0.11	**0.22**±0.11	0.09±0.10	**0.22**±0.12	0.02±0.09	**0.22**±0.12	0.09±0.11
			$\text{BEDROC}_{80.5}$	0.02±0.02	0.02±0.01	0.05±0.03	0.10±0.06	0.00±0.00	**0.18**±0.10	0.02±0.01
			BEDROC_{20}	0.10±0.05	0.12±0.05	0.11±0.06	**0.20**±0.08	0.04±0.02	**0.20**±0.09	0.10±0.05

LigUnity's screening checkpoint is retained in Table 6: its seven-target equal-target nEF@1% is -0.029 ± 0.058 and nEF@5% is -0.005 ± 0.025. Its ensemble inference averaged 0.000829 seconds per ligand across the seven seeded 32-ligand timing samples, weighted by the full target cohorts, and this includes ligand embedding time.

E. Per-series and per-target potency results

Choice of correlation metric. Figure 4 uses Pearson r for OpenFE, FEP+ 4, and CASP16 to support comparison with published results: IsoDDE reports Pearson correlations for all three benchmarks, ApherisAffinity plots Pearson correlations for all three, and TerraBind reports per-target Pearson correlations for CASP16 [17, 19, 20]. The CASP16 organizers instead assess affinity using a compound-count-weighted mean of per-target Kendall τ [32], and we report this metric alongside the individual target values in Table 8. For OpenBind, Figure 4 and Table 8 use Spearman ρ, following the published ranking benchmark [33]. We do not use OpenBind's RMSE metric for this zero-shot comparison because the evaluated methods do not all produce calibrated predictions on a common $\text{p}K_D$ scale. Absolute errors would therefore reflect differences in

score scale and offset as well as predictive performance. Spearman ρ assesses ranking without requiring assay-specific calibration.

E.1. Published per-compound predictions

We join FEP+ [45] and OpenFE RBFE [29] predictions to frozen rows by system, series and ligand name, then re-score them with the same paired bootstrap as re-run methods. Neither covers CASP16. The source authors shift their relative networks to each series' experimental mean. Within-series correlations are invariant to these shifts, whereas pooling independently anchored series can credit experimental offsets as predicted between-series differences. We therefore compute correlations within the original source series for every re-scored method and report no MAE for the relative free-energy methods. The OpenFE means are $r = 0.72$ for FEP+ (published 0.72) and 0.65 for RBFE (published 0.63). Our pinned RBFE predictions include the September 2025 upstream correction to the 40 PFKFB3 predictions. The earlier release reproduces $r = 0.63$ under source-series aggregation.

Boltz-2 reruns. Our Boltz-2 runs give $r = 0.56, 0.59$ and 0.77 on OpenFE, CASP16 and FEP+ 4, respectively, compared with the original report's 0.62, 0.65 and 0.66 [14]. These are separate inference runs scored on our frozen cohorts, and we have not reconciled the remaining differences with the published aggregates. Changing the OpenFE aggregation alone does not explain them.

Boltz-2 sampling check. On the three public-potency benchmarks, increasing structure recycles from three to five and structure samples from two to five produced small changes in compound-weighted Pearson r. The paired differences (higher minus lower sampling) were -0.004 on OpenFE (95% bootstrap interval $[-0.017, 0.010]$), -0.001 on FEP+ 4 ($[-0.023, 0.020]$), and -0.006 on CASP16 ($[-0.037, 0.023]$). Inputs, checkpoints, affinity sampling and molecular-weight correction were unchanged. This sensitivity analysis retains the historical 39-target grouping for OpenFE. The main evaluation instead computes correlations within its 58 source series. We used 1,000 paired compound bootstrap draws, of which 842 were valid for OpenFE and all 1,000 for the other two benchmarks. Each setting had one inference run, so these intervals reflect compound resampling rather than inference-seed variability. We detected no difference on these benchmarks, but the comparison does not establish equivalence or test enrichment on the screening targets. The current public-potency figures use five recycles and five structure samples, and the screening figures retain three recycles and two structure samples.

OpenFE bootstrap validity. Fourteen source series contain only three to five compounds, so resampling can produce constant labels or predictions and an undefined correlation. An aggregate draw is retained only when every constituent series has a defined correlation. We do not drop series or renormalize their weights within a draw. For Pearson r, 455 of 1,000 draws are valid for RUBY and all re-scored comparators except molecular weight, which retains 426. The reported standard errors are conditional on these valid draws.

Cited and derived results. IsoDDE [20] reports 873 observations over 39 OpenFE targets, matching our counts, but supplies no target manifest. It says these public sets follow the Boltz-2 protocol [14], which weights cross-assay regression averages by compound count, but does not state the correlation units or weights used for IsoDDE's own reported Pearson means. The count of 39 unique proteins does not establish whether source series sharing a protein were pooled before computing correlations. Its FEP+ 4 set names our four targets and 87 compounds. It does not name its CASP16 compounds. Boltz-2 calls the set 140 pairs across two targets, and the CASP16 assessment lists autotaxin (123) and chymase (17) as the affinity targets [32], so we assume the same cohort. IsoDDE takes the other methods' plotted values from Boltz-2, so agreement between those bars and Boltz-2's reported aggregates cannot establish IsoDDE's own weighting.

No per-compound IsoDDE predictions are available here for matched bootstrap errors or an equal-target sensitivity check.

TerraBind [19] reports Pearson 0.73 on autotaxin (123 compounds) and 0.45 on chymase (17), read from figure labels, but no pooled CASP16 value or pooling rule. Our compound-weighted derivation, $(0.73 \times 123 + 0.45 \times 17)/140 = 0.696$, is not a quoted aggregate.

AnewAffinity [46] explicitly weights system-level metrics by ligand count on the eight-system, 199-ligand JACS subset of OpenFE, but does not report on the full OpenFE panel or the other two series sources. Its reported Pearson value is a weighted mean of within-system r^2, which cannot be converted to our mean Pearson r by taking its square root.

Table 7. Per-series Pearson r for each method with per-compound predictions, ± one bootstrap standard error, with each benchmark's compound-weighted mean first. OpenFE rows are its 58 original source series, grouped by source collection. FEP+ 4 and CASP16 carry one series per target. Pearson is undefined for a series with fewer than three compounds or constant labels or predictions. Bold marks the best method in each row, compared as printed. Significance marks are omitted from the public-potency figure and this table.

Series / target	n	*RUBY*	*FEP+*‡	*Nesso-1*	*OpenFE RBFE*‡	*Boltz-2*	*LigUnity*	*Mol. weight*	*AQAffinity*	*GNINA*
OpenFE										
compound-weighted	873	**0.74**±0.02	0.72±0.02	0.66±0.02	0.65±0.02	0.56±0.02	0.63±0.02	0.38±0.03	0.33±0.03	0.25±0.03
Charge annihilation										
cdk2	3	0.97±0.01	0.98±0.01	0.92±0.04	**1.00**±0.00	-0.10±0.81	-0.70±0.13	-0.47±0.88	0.91±0.04	0.57±0.19
dlk	5	0.86±0.26	**0.99**±0.01	0.45±0.49	0.64±0.51	0.44±0.51	0.95±0.08	0.26±0.57	0.83±0.31	0.16±0.62
egfr	5	**-0.13**±0.52	-0.47±0.42	-0.42±0.67	-0.37±0.47	-0.20±0.54	-0.51±0.68	-0.18±0.63	-0.40±0.50	-0.23±0.55
ephx2	4	0.89±0.47	0.93±0.47	**1.00**±0.00	0.82±0.48	**1.00**±0.00	0.39±0.65	0.94±0.03	0.96±0.47	0.93±0.04
irak4_s2	5	0.82±0.31	0.85±0.41	0.88±0.34	0.83±0.36	0.91±0.45	**0.93**±0.29	0.63±0.75	0.22±0.57	0.76±0.71
irak4_s3	4	0.75±0.45	0.42±0.50	-0.22±0.67	0.67±0.82	0.53±0.48	**0.89**±0.07	-0.86±0.09	0.64±0.46	-0.78±0.45
itk	4	0.63±0.58	0.37±0.65	**0.99**±0.07	0.23±0.59	0.65±0.43	0.46±0.65	0.12±0.73	0.13±0.65	0.49±0.55
jak1	6	0.51±0.37	0.88±0.25	**0.94**±0.21	0.55±0.34	0.47±0.54	0.73±0.58	0.86±0.11	0.36±0.45	0.54±0.27
jnk1	3	0.86±0.84	0.92±0.04	0.08±0.83	0.69±0.83	0.11±0.84	**0.97**±0.01	-0.97±0.01	-0.97±0.01	-0.64±0.82
ptp1b	3	0.93±0.03	0.92±0.03	**0.99**±0.00	0.92±0.03	0.72±0.83	0.93±0.03	0.98±0.01	0.95±0.02	0.95±0.02
thrombin	6	0.94±0.47	**0.97**±0.12	0.70±0.49	0.94±0.08	0.90±0.39	0.90±0.35	0.41±0.45	0.86±0.70	0.76±0.61
tyk2	5	0.69±0.39	0.89±0.53	0.94±0.30	**0.97**±0.53	0.89±0.25	0.03±0.57	0.29±0.53	0.12±0.60	-0.29±0.52
Fragments										
hsp90_2rings	6	0.29±0.56	0.50±0.62	0.43±0.43	0.66±0.45	0.60±0.51	0.37±0.41	0.37±0.41	**0.72**±0.33	0.19±0.51
hsp90_single_ring	7	0.54±0.26	0.73±0.23	0.45±0.31	0.42±0.40	**0.91**±0.12	0.62±0.21	0.78±0.25	-0.13±0.42	0.73±0.27
jak2_set1	10	0.88±0.09	0.87±0.09	0.64±0.19	0.86±0.13	0.72±0.16	**0.95**±0.03	0.57±0.21	0.68±0.22	0.35±0.33
jak2_set2	8	**0.85**±0.09	0.64±0.21	0.48±0.22	0.71±0.20	0.76±0.21	0.83±0.09	0.45±0.27	0.77±0.16	-0.53±0.23
liga	11	**0.98**±0.01	0.97±0.04	0.95±0.02	0.96±0.05	**0.98**±0.02	0.90±0.08	0.96±0.04	0.96±0.07	0.80±0.14
mcl1	12	0.52±0.18	**0.56**±0.23	0.23±0.25	0.52±0.22	0.42±0.17	0.39±0.23	0.26±0.27	0.20±0.29	-0.38±0.17
mup1	6	0.81±0.18	0.96±0.13	-0.49±0.28	**0.97**±0.13	-0.59±0.29	0.51±0.37	**0.97**±0.05	-0.26±0.41	0.96±0.13
p38	6	0.82±0.37	**0.95**±0.07	0.82±0.27	0.70±0.40	0.72±0.36	0.89±0.11	0.80±0.32	0.66±0.29	0.77±0.19
t4_lysozyme	12	0.64±0.23	0.65±0.17	0.46±0.29	0.79±0.09	0.57±0.20	**0.94**±0.04	0.49±0.23	-0.05±0.44	0.44±0.25
JACS										
bace	36	0.75±0.08	0.62±0.11	**0.85**±0.05	0.55±0.12	0.26±0.15	0.80±0.05	-0.15±0.13	-0.14±0.12	-0.40±0.11
cdk2	16	**0.94**±0.02	0.65±0.14	0.77±0.08	0.77±0.08	0.86±0.05	0.79±0.09	0.48±0.19	0.71±0.11	0.62±0.15
jnk1	21	**0.92**±0.04	0.84±0.05	0.70±0.11	0.76±0.07	0.76±0.05	0.63±0.15	0.39±0.25	0.52±0.11	-0.54±0.18
mcl1	42	0.88±0.03	0.70±0.09	**0.92**±0.02	0.70±0.07	0.89±0.03	0.65±0.07	0.55±0.11	0.61±0.10	0.26±0.14

continued

Series / target	*n*	*RUBY*	*FEP+*‡	*Nesso-1*	*OpenFE RBFE*‡	*Boltz-2*	*LigUnity*	*Mol. weight*	*AQAffinity*	*GNINA*
p38	34	0.79±0.07	0.70±0.08	**0.87**±0.04	0.43±0.16	0.67±0.08	0.34±0.18	0.57±0.11	0.53±0.16	0.05±0.22
ptp1b	23	0.83±0.07	**0.92**±0.02	0.84±0.05	0.78±0.09	0.66±0.11	0.74±0.07	0.74±0.10	0.77±0.10	0.67±0.18
thrombin	11	**0.92**±0.06	0.82±0.14	0.83±0.10	0.76±0.18	0.74±0.13	0.76±0.17	0.48±0.25	**0.92**±0.07	0.71±0.17
tyk2	16	0.85±0.09	**0.93**±0.03	0.86±0.08	0.89±0.04	0.86±0.08	0.70±0.15	0.06±0.24	0.19±0.26	-0.08±0.17
Janssen BACE										
bace_ciordia_prospective	9	0.70±0.23	0.77±0.16	0.78±0.14	0.76±0.15	-0.05±0.30	**0.80**±0.15	0.57±0.29	0.02±0.30	0.35±0.29
bace_p3_arg368_in	21	0.54±0.12	0.79±0.08	**0.86**±0.05	0.63±0.14	0.46±0.14	0.69±0.12	0.26±0.22	0.16±0.22	0.55±0.18
ciordia_retro	32	0.81±0.06	0.75±0.09	**0.95**±0.02	0.83±0.05	0.80±0.08	0.83±0.05	0.52±0.12	0.20±0.19	0.56±0.15
keranen_p2	12	0.47±0.17	0.26±0.34	0.17±0.23	0.46±0.24	0.51±0.25	**0.67**±0.25	-0.03±0.31	0.37±0.25	-0.41±0.24
MCS docking										
hne	17	**0.92**±0.07	0.87±0.14	0.91±0.07	0.74±0.10	0.83±0.17	0.76±0.24	0.06±0.35	0.79±0.16	-0.03±0.18
renin	29	0.74±0.09	0.68±0.09	0.83±0.08	0.74±0.10	0.75±0.10	**0.84**±0.06	0.33±0.17	0.59±0.12	0.53±0.14
Merck										
cdk8	32	**0.83**±0.08	0.75±0.10	0.81±0.06	0.80±0.06	0.68±0.09	0.72±0.07	0.57±0.14	0.27±0.18	0.71±0.11
cmet	24	0.87±0.06	**0.91**±0.03	0.86±0.06	0.79±0.09	0.80±0.07	0.81±0.06	0.71±0.11	0.13±0.15	-0.27±0.14
eg5	28	**0.77**±0.11	0.59±0.10	0.29±0.14	0.61±0.16	0.51±0.16	0.68±0.10	0.59±0.16	0.06±0.20	-0.22±0.19
hif2a	41	0.39±0.10	**0.78**±0.08	0.40±0.12	0.63±0.11	0.44±0.11	0.50±0.12	0.20±0.10	0.12±0.14	0.45±0.15
pfkfb3	40	**0.81**±0.05	0.76±0.06	0.64±0.09	0.64±0.10	0.22±0.16	0.39±0.12	0.28±0.15	-0.04±0.13	0.53±0.09
shp2	26	0.88±0.05	0.71±0.08	**0.92**±0.03	0.06±0.24	0.35±0.15	0.44±0.14	0.29±0.14	0.03±0.20	0.41±0.12
syk	44	**0.71**±0.07	**0.71**±0.08	0.46±0.14	0.67±0.10	0.42±0.16	0.44±0.16	0.28±0.11	0.05±0.15	0.53±0.13
tnks2	27	0.56±0.10	0.54±0.18	0.77±0.08	0.20±0.20	0.64±0.13	**0.83**±0.07	0.44±0.18	0.51±0.18	0.39±0.16
Miscellaneous										
btk	6	0.82±0.25	0.88±0.29	0.56±0.45	**0.94**±0.24	0.69±0.36	0.32±0.52	0.74±0.29	0.54±0.31	0.78±0.26
cdk8	10	-0.23±0.32	0.66±0.15	0.51±0.33	0.64±0.27	0.10±0.37	**0.72**±0.14	0.35±0.44	0.50±0.25	-0.36±0.32
faah	24	0.79±0.10	**0.85**±0.08	0.40±0.15	0.77±0.11	0.53±0.12	0.52±0.18	0.61±0.15	0.50±0.10	0.27±0.21
galectin	26	**0.89**±0.04	0.87±0.04	0.72±0.12	0.73±0.11	0.72±0.09	0.30±0.18	0.42±0.21	0.73±0.13	0.65±0.13
hiv1_protease	13	**0.52**±0.18	-0.01±0.34	-0.14±0.25	-0.14±0.26	-0.44±0.24	0.38±0.28	-0.08±0.31	-0.47±0.20	0.15±0.28
Scaffold hopping										
bace1	3	0.91±0.84	**0.96**±0.02	-0.94±0.03	0.82±0.08	-0.44±0.81	0.93±0.84	-0.62±0.82	-0.42±0.81	-0.88±0.05
factor_xa	3	-0.94±0.03	-0.98±0.01	0.98±0.85	**1.00**±0.00	-0.95±0.84	-0.95±0.02	-0.23±0.82	-0.43±0.82	-0.96±0.84
Water										
brd4	8	**0.94**±0.04	0.05±0.39	0.83±0.24	0.66±0.23	0.77±0.25	0.92±0.22	-0.64±0.43	0.62±0.28	-0.03±0.52
chk1	13	0.71±0.10	0.78±0.11	0.76±0.10	0.86±0.09	0.84±0.08	**0.95**±0.03	0.71±0.25	0.56±0.18	0.46±0.20
hsp90_kung	11	0.81±0.09	**0.83**±0.14	0.74±0.17	0.68±0.22	0.78±0.12	0.81±0.13	0.69±0.20	0.62±0.20	-0.36±0.22
hsp90_woodhead	4	0.97±0.02	**1.00**±0.00	0.25±0.75	0.72±0.61	0.71±0.50	0.51±0.71	-0.74±0.75	0.95±0.49	-0.34±0.61
scyt_dehyd	7	0.66±0.27	**0.88**±0.22	-0.31±0.42	0.81±0.17	-0.27±0.37	0.58±0.30	0.79±0.14	-0.09±0.40	0.06±0.43
taf12	8	0.84±0.16	**0.91**±0.22	0.62±0.21	0.86±0.42	0.82±0.11	0.61±0.29	-0.40±0.40	0.45±0.31	0.21±0.37
thrombin	21	0.70±0.11	0.75±0.08	0.67±0.12	0.46±0.17	0.59±0.16	**0.77**±0.09	0.50±0.13	0.41±0.19	-0.02±0.23
urokinase	4	0.65±0.50	0.93±0.06	0.70±0.86	0.77±0.09	0.52±0.79	0.72±0.84	0.83±0.44	**0.94**±0.46	0.92±0.05
PLB / FEP+ 4										
compound-weighted	87	**0.86**±0.03	0.77±0.04	0.81±0.04	0.66±0.07	0.77±0.04	0.54±0.09	0.41±0.10	0.50±0.08	0.31±0.10
cdk2	16	**0.94**±0.02	0.65±0.14	0.82±0.09	0.77±0.08	0.83±0.06	0.82±0.08	0.48±0.19	0.59±0.13	0.73±0.09
jnk1	21	**0.92**±0.04	0.84±0.05	0.64±0.11	0.76±0.07	0.81±0.05	0.36±0.23	0.39±0.26	0.53±0.18	-0.16±0.26
p38	34	0.79±0.07	0.70±0.08	**0.87**±0.04	0.43±0.16	0.68±0.08	0.40±0.16	0.56±0.11	0.55±0.10	0.58±0.12
tyk2	16	0.84±0.08	**0.93**±0.03	0.87±0.08	0.89±0.04	0.83±0.09	0.81±0.09	0.05±0.25	0.23±0.26	-0.05±0.29
CASP16										
compound-weighted	140	**0.77**±0.03	–	0.65±0.05	–	0.59±0.05	0.56±0.06	0.56±0.05	0.49±0.06	–
autotaxin	123	**0.80**±0.03	–	0.69±0.05	–	0.60±0.06	0.58±0.06	0.60±0.06	0.56±0.05	–

continued

Series / target	*n*	*RUBY*	*FEP+*‡	*Nesso-1*	*OpenFE RBFE*‡	*Boltz-2*	*LigUnity*	*Mol. weight*	*AQAffinity*	*GNINA*
`chymase`	17	**0.55**±0.10	–	0.41±0.27	–	0.54±0.19	0.37±0.21	0.30±0.13	0.00±0.23	–

Table 8. Additional public-potency metrics: Spearman ρ for OpenBind EV-A71 2A, as shown in Figure 4, and Kendall τ for CASP16. The CASP16 mean weights targets by compound count, and its target rows show autotaxin and chymase separately. Values carry one bootstrap standard error where per-compound predictions are available. OpenBind Nesso-1 and ApherisAffinity are cited aggregates without matched errors. Its MolWt, GNINA, Boltz-2 and AQAffinity values use published per-compound predictions, and LigUnity was run on the same cohort. OpenBind uses three decimals and CASP16 two.

Series / target	*n*	*RUBY*	*Nesso-1*	*Boltz-2*	*LigUnity*	*Mol. weight*	*AQAffinity*	*GNINA*	*ApherisAffinity*
OpenBind EV-A71 2A									
EV-A71 2A	494	**0.574**±0.034	0.512	0.403±0.038	0.112±0.044	0.487±0.036	0.114±0.044	0.454±0.037	0.537
CASP16									
compound-weighted	140	**0.57** ±0.04	0.47 ±0.05	0.38 ±0.05	0.39 ±0.05	0.42 ±0.05	0.35 ±0.05	–	–
`autotaxin`	123	**0.60** ±0.04	0.49 ±0.04	0.40 ±0.05	0.42 ±0.05	0.42 ±0.05	0.39 ±0.05	–	–
`chymase`	17	**0.41** ±0.14	0.26 ±0.19	0.25 ±0.19	0.18 ±0.20	0.40 ±0.14	0.07 ±0.20	–	–

F. Partner-program analyses

F.1. Per-assay performance

Table 9. Per-assay performance on the eight Genesis–Incyte assays underlying Figure 5, with methods in columns and metrics in rows. The first row group gives the equally weighted assay average, with n summed across assays. Each setting reports Spearman ρ, Pearson r, and Kendall τ_b, with MAE in pX units additionally reported after recalibration. The n column distinguishes the full zero-shot cohort, the shared fitted cohort for recalibrated correlations, and eligible exact measurements for MAE. Values show point estimates and one standard error from 1,000 within-assay compound bootstrap resamples (seed 0), rounded to two decimals. Assay-average errors use the Figure 5 bootstrap estimates where available, with independent assay errors propagated for Kendall τ_b and molecular weight. Recalibrated correlations use pooled rolling predictions for every method.

Program task	Mode	Metric	n	*RUBY*	*Nesso-1*	*Boltz-2*	*GNINA*	*LigUnity*	*AQAffinity*	*Mol. weight*
Assay average	Zero-shot	ρ	22,036	**0.61**±0.01	0.39±0.01	0.46±0.01	0.19±0.01	0.11±0.01	0.30±0.01	0.43±0.01
		r	22,036	**0.59**±0.01	0.36±0.01	0.45±0.01	0.19±0.01	0.11±0.01	0.29±0.01	0.41±0.01
		τ	22,036	**0.44**±0.00	0.27±0.01	0.32±0.01	0.13±0.01	0.08±0.01	0.21±0.01	0.30±0.01
	Recalibrated	ρ	20,619	**0.71**±0.01	0.44±0.01	0.51±0.01	0.39±0.01	0.41±0.01	0.45±0.01	0.49±0.01
		r	20,619	**0.70**±0.01	0.42±0.01	0.49±0.01	0.36±0.01	0.40±0.01	0.42±0.01	0.47±0.01
		τ	20,619	**0.53**±0.01	0.31±0.01	0.36±0.01	0.27±0.01	0.29±0.01	0.31±0.01	0.34±0.01
		MAE	17,631	**0.53**±0.01	0.76±0.01	0.74±0.01	0.81±0.01	0.81±0.01	0.80±0.01	0.77±0.01
Incyte-A biochem	Zero-shot	ρ	3,560	**0.73**±0.01	0.61±0.01	0.63±0.01	-0.01±0.02	-0.23±0.02	0.34±0.01	0.41±0.02
		r	3,560	**0.72**±0.01	0.60±0.01	0.62±0.01	-0.01±0.02	-0.19±0.02	0.35±0.01	0.41±0.02
		τ	3,560	**0.54**±0.01	0.43±0.01	0.45±0.01	-0.01±0.01	-0.15±0.01	0.23±0.01	0.29±0.01
	Recalibrated	ρ	3,387	**0.83**±0.01	0.63±0.01	0.67±0.01	0.44±0.02	0.49±0.01	0.50±0.01	0.57±0.01
		r	3,387	**0.83**±0.01	0.61±0.01	0.67±0.01	0.40±0.02	0.47±0.01	0.48±0.01	0.55±0.02
		τ	3,387	**0.64**±0.01	0.45±0.01	0.49±0.01	0.30±0.01	0.34±0.01	0.34±0.01	0.41±0.01
		MAE	3,001	**0.68**±0.01	1.04±0.01	1.03±0.01	1.30±0.01	1.28±0.01	1.23±0.01	1.20±0.01
Incyte-A cellular	Zero-shot	ρ	3,360	**0.54**±0.01	0.36±0.02	0.40±0.01	0.03±0.02	-0.30±0.02	0.20±0.02	0.21±0.02
		r	3,360	**0.52**±0.01	0.38±0.02	0.36±0.01	0.02±0.02	-0.28±0.02	0.20±0.02	0.21±0.02
		τ	3,360	**0.37**±0.01	0.25±0.01	0.27±0.01	0.02±0.01	-0.20±0.01	0.14±0.01	0.14±0.01
	Recalibrated	ρ	3,260	**0.81**±0.01	0.63±0.01	0.65±0.01	0.60±0.01	0.61±0.01	0.61±0.01	0.63±0.01
		r	3,260	**0.81**±0.01	0.62±0.01	0.64±0.01	0.59±0.01	0.60±0.01	0.60±0.01	0.62±0.01
		τ	3,260	**0.61**±0.01	0.45±0.01	0.46±0.01	0.43±0.01	0.44±0.01	0.44±0.01	0.44±0.01
		MAE	3,109	**0.48**±0.01	0.80±0.01	0.80±0.01	0.86±0.01	0.84±0.01	0.85±0.01	0.85±0.01
Incyte-B assay 1	Zero-shot	ρ	1,418	**0.69**±0.01	0.45±0.02	0.42±0.02	0.22±0.03	0.19±0.02	0.46±0.02	0.43±0.02
		r	1,418	**0.63**±0.01	0.42±0.02	0.41±0.02	0.27±0.02	0.13±0.02	0.45±0.02	0.41±0.02
		τ	1,418	**0.50**±0.01	0.32±0.02	0.31±0.02	0.16±0.02	0.13±0.02	0.33±0.02	0.31±0.02
	Recalibrated	ρ	1,166	**0.54**±0.02	0.40±0.03	0.38±0.03	0.27±0.03	0.23±0.03	0.40±0.03	0.35±0.03
		r	1,166	**0.53**±0.02	0.34±0.02	0.31±0.02	0.19±0.03	0.20±0.03	0.33±0.03	0.29±0.03
		τ	1,166	**0.39**±0.02	0.28±0.02	0.27±0.02	0.18±0.02	0.16±0.02	0.28±0.02	0.25±0.02
		MAE	846	**0.57**±0.02	0.71±0.02	0.71±0.02	0.73±0.02	0.76±0.02	0.70±0.02	0.72±0.02
Incyte-B assay 2	Zero-shot	ρ	1,418	**0.72**±0.01	0.49±0.02	0.47±0.02	0.29±0.03	0.24±0.02	0.49±0.02	0.51±0.02
		r	1,418	**0.63**±0.01	0.43±0.02	0.42±0.02	0.29±0.02	0.19±0.02	0.44±0.02	0.44±0.02
		τ	1,418	**0.53**±0.01	0.35±0.01	0.34±0.02	0.21±0.02	0.17±0.02	0.34±0.01	0.36±0.02
	Recalibrated	ρ	1,112	**0.60**±0.02	0.43±0.03	0.40±0.03	0.29±0.03	0.31±0.03	0.39±0.03	0.39±0.03
		r	1,112	**0.56**±0.02	0.33±0.03	0.31±0.03	0.21±0.03	0.26±0.03	0.31±0.03	0.31±0.03
		τ	1,112	**0.44**±0.02	0.30±0.02	0.28±0.02	0.19±0.02	0.21±0.02	0.27±0.02	0.27±0.02
		MAE	837	**0.45**±0.01	0.56±0.01	0.57±0.01	0.58±0.02	0.58±0.01	0.57±0.01	0.57±0.01

continued

Program task	Mode	Metric	n	*RUBY*	*Nesso-1*	*Boltz-2*	*GNINA*	*LigUnity*	*AQAffinity*	*Mol. weight*
Incyte-C	Zero-shot	ρ	2,302	**0.41**±0.02	0.24±0.02	0.34±0.02	0.07±0.02	0.26±0.02	0.31±0.02	0.34±0.02
		r	2,302	**0.43**±0.02	0.23±0.02	0.33±0.02	0.10±0.02	0.26±0.02	0.30±0.02	0.31±0.02
		τ	2,302	**0.28**±0.01	0.16±0.01	0.23±0.01	0.05±0.02	0.19±0.01	0.22±0.01	0.24±0.01
	Recalibrated	ρ	2,153	**0.70**±0.01	0.34±0.02	0.43±0.02	0.33±0.02	0.37±0.02	0.38±0.02	0.43±0.02
		r	2,153	**0.71**±0.01	0.33±0.02	0.45±0.02	0.34±0.02	0.39±0.02	0.39±0.02	0.43±0.02
		τ	2,153	**0.52**±0.01	0.23±0.01	0.30±0.01	0.22±0.01	0.26±0.01	0.26±0.01	0.30±0.01
		MAE	1,714	**0.60**±0.01	0.79±0.01	0.77±0.01	0.79±0.01	0.78±0.01	0.76±0.01	0.78±0.01
Incyte-D	Zero-shot	ρ	451	0.42±0.04	0.16±0.05	0.35±0.04	0.24±0.05	0.28±0.04	0.35±0.04	**0.45**±0.04
		r	451	**0.45**±0.03	0.08±0.05	0.37±0.04	0.23±0.04	0.29±0.04	0.34±0.04	**0.45**±0.04
		τ	451	0.29±0.03	0.11±0.03	0.25±0.03	0.17±0.03	0.19±0.03	0.24±0.03	**0.32**±0.03
	Recalibrated	ρ	318	**0.65**±0.04	0.09±0.05	0.35±0.05	0.25±0.05	0.28±0.05	0.31±0.05	0.44±0.05
		r	318	**0.66**±0.03	0.08±0.05	0.33±0.04	0.21±0.05	0.26±0.05	0.28±0.05	0.44±0.04
		τ	318	**0.48**±0.03	0.07±0.03	0.24±0.03	0.17±0.03	0.19±0.03	0.21±0.03	0.30±0.03
		MAE	232	**0.55**±0.03	0.85±0.04	0.77±0.03	0.78±0.04	0.84±0.04	0.80±0.04	0.79±0.03
Incyte-E biochem	Zero-shot	ρ	5,181	**0.69**±0.01	0.43±0.01	0.58±0.01	0.31±0.01	0.27±0.01	0.13±0.01	0.57±0.01
		r	5,181	**0.69**±0.01	0.41±0.01	0.58±0.01	0.30±0.01	0.29±0.01	0.16±0.01	0.56±0.01
		τ	5,181	**0.50**±0.01	0.30±0.01	0.41±0.01	0.21±0.01	0.18±0.01	0.09±0.01	0.40±0.01
	Recalibrated	ρ	4,994	**0.76**±0.01	0.55±0.01	0.64±0.01	0.55±0.01	0.53±0.01	0.52±0.01	0.58±0.01
		r	4,994	**0.76**±0.01	0.56±0.01	0.63±0.01	0.52±0.01	0.56±0.01	0.54±0.01	0.59±0.01
		τ	4,994	**0.57**±0.01	0.38±0.01	0.45±0.01	0.38±0.01	0.37±0.01	0.36±0.01	0.41±0.01
		MAE	4,444	**0.56**±0.01	0.85±0.01	0.77±0.01	0.89±0.01	0.91±0.01	0.92±0.01	0.78±0.01
Incyte-E cellular	Zero-shot	ρ	4,346	**0.67**±0.01	0.36±0.01	0.49±0.01	0.36±0.01	0.18±0.01	0.11±0.02	0.50±0.01
		r	4,346	**0.64**±0.01	0.32±0.01	0.48±0.01	0.35±0.01	0.19±0.01	0.10±0.02	0.48±0.01
		τ	4,346	**0.48**±0.01	0.24±0.01	0.34±0.01	0.25±0.01	0.12±0.01	0.07±0.01	0.35±0.01
	Recalibrated	ρ	4,229	**0.75**±0.01	0.49±0.01	0.59±0.01	0.44±0.01	0.48±0.01	0.46±0.01	0.54±0.01
		r	4,229	**0.74**±0.01	0.47±0.01	0.56±0.01	0.41±0.01	0.45±0.01	0.43±0.01	0.52±0.01
		τ	4,229	**0.56**±0.01	0.33±0.01	0.41±0.01	0.30±0.01	0.32±0.01	0.31±0.01	0.38±0.01
		MAE	3,448	**0.36**±0.01	0.50±0.01	0.48±0.01	0.51±0.01	0.53±0.01	0.53±0.01	0.48±0.01

F.2. Paired comparisons

Table 10. Paired bootstrap comparisons on eight partner assays. Each assay receives equal weight, and compounds are resampled within assays on shared draws. Zero-shot correlations use 22,036 scored registrations. Recalibrated correlations use 20,619 shared fitted predictions, and MAE uses 17,631 eligible exact measurements. The comparison method is selected separately for each metric by its aggregate point estimate. Advantage is RUBY minus the comparison method for correlations and the comparison method minus RUBY for MAE. Intervals are bootstrap percentiles, and tests condition on the selected RUBY readout and recalibration fits. With 1,000 resamples, zero unfavorable draws are reported as $p < 0.001$.

Mode	Metric	Comparison method	RUBY	Value	Advantage [95% CI]	One-sided p
Zero-shot	Spearman ρ	Boltz-2	0.61	0.46	0.15 [0.13, 0.16]	< 0.001
	Pearson r	Boltz-2	0.59	0.45	0.14 [0.13, 0.16]	< 0.001
Recalibrated	Spearman ρ	Boltz-2	0.71	0.51	0.19 [0.18, 0.21]	< 0.001
	Pearson r	Boltz-2	0.70	0.49	0.21 [0.20, 0.23]	< 0.001
	MAE (pX)	Boltz-2	0.53	0.74	0.20 [0.19, 0.21]	< 0.001

F.3. Genesis–Incyte chemical similarity

For each scored rolling test compound, we take the maximum Tanimoto similarity of its nonchiral Morgan fingerprint (radius 3, 2,048 bits) to an earlier exactly measured compound from the same assay. This temporal prefix describes novelty relative to earlier same-assay measurements, not to a comparator's pretraining corpus. We use fixed 0.1-wide bins and display a bin only when all eight Genesis–Incyte assays have at least six exact compounds. Both MAE and Spearman ρ use each method's rolling-recalibrated predictions. We retain only records with finite fitted predictions from every method, matching the recalibrated cohort described in Appendix C.4. MAE uses exact measurements, and Spearman additionally includes censored measurements at their reported bounds. We pool predictions across all rolling test windows within each assay and similarity bin before computing metrics, then average assays equally. The five bins spanning 0.4–0.9 meet the coverage criterion for all eight assays.

The bins are descriptive: the compounds and potency ranges differ between them. The equal-assay curves are smoother than individual assays, which need not be monotone. Within-bin ρ also omits comparisons between bins, so it cannot be compared directly with the whole-cohort correlation in Figure 5.

F.4. Data scaling for fine-tuning

We first tested whether updating the final potency-prediction layers of RUBY and the comparison models improved performance under the rolling-temporal evaluation used in the main paper. For most model–assay combinations, fine-tuning did not consistently improve on the zero-shot model over the duration of each program. Potency distributions shifted between the earlier validation data and later test compounds, and training settings selected on validation data did not reliably transfer to the subsequent test window. We therefore use the more robust assay-specific recalibration for the main evaluation and examine fine-tuning separately under a split based on chemical clustering as a way to probe each model's internal generalization capabilities.

For the data-scaling analysis, we retain up to 1,000 compounds per assay, sampled evenly across its measurement history.

Train–test split. For each assay, we group compounds by chemical similarity using Butina clustering with a Tanimoto threshold of 0.6 and assign the groups to four folds. We use a 4-fold evaluation setting, testing on each fold and using the other three as the training pool. Within each training pool, we repeat the clustering to create four inner folds for selecting training settings without using the test measurements. We evaluate training budgets of 12.5%, 25%, 50%, and 100% of the available training pool, selecting settings separately at each budget.

Fine-tuning and model selection. For RUBY, Boltz-2, Nesso-1, and AQAffinity, we update the last one, two, or three layers of the potency prediction head while keeping the rest of the model fixed. We precompute the inputs to these layers so that training does not require repeatedly running the full model. For each model and data budget, we select the number of layers to update and the other training settings using the two-stage validation search in Table 11. We rank configurations using a score that combines Spearman correlation, MAE, and average precision for identifying the highest-potency 20% of compounds. The three best configurations are refitted on the training pool at that data budget, and their predictions are averaged for the held-out test fold.

We combine held-out predictions across the four outer folds, compute metrics separately for each assay, and report the equally weighted mean across the eight assays.

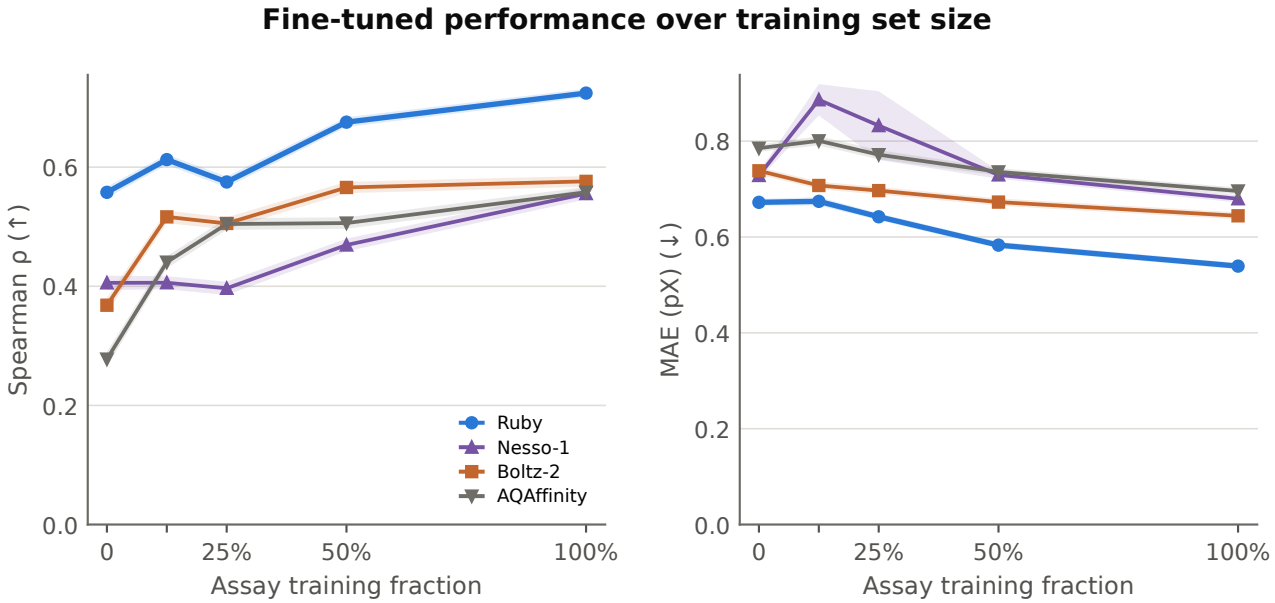


Figure 9. Data-scaling of assay fine-tuning, averaged with equal assay weight over the eight Genesis–Incyte assays. Each model is fine-tuned on nested fractions of up to 1000 measurements from each assay training pool, so metrics are not directly comparable to those in Figure 6. Shading represents ± one bootstrap standard error.

RUBY outperforms every comparison method on both metrics at every training fraction, widening its margin as the fraction approaches 100%. These results suggest that RUBY may have a richer internal representation than the other models.

Table 11. Two-stage hyperparameter search for the affinity models. Stage 1 is an exhaustive grid search. Stage 2 varies one parameter at a time for the two best configurations per loss function rather than forming a second Cartesian product. Both stages are scored only on inner-fold validation data, and the outer test fold is excluded from model selection.

Stage 1 — Screening grid

Axis	Values
Last linear layers updated	1, 2, 3
Learning rate	3×10^{-4}, 10^{-3}, 3×10^{-3}
Loss	MSE, Huber, pairwise
Relative penalty λ	0, 0.1
Dropout	0 (fixed)
LR schedule	constant (fixed)
Mean calibration	off (fixed)
Configurations	$3 \times 3 \times 3 \times 2 = 54$
Update checkpoints	10, 25, 50, 100, 200, 400
Scored points	$54 \times 6 = 324$

Stage 2 — Refinements

Axis	Values
Dropout	0.1
LR schedule	linear, cosine
Schedule horizon H	100, 400
Mean calibration	on (MSE/Huber only)
Configurations	34
Scored points	84
Total per fold	**88 cfg**