Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Affinity panels dilute the target-specific signal that transfers to a held-out protein

**Adeliya Leleytner**

a.leleytner@gmail.com

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute)

**Maxim Fedorov**

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute)

**Research Article**

**Keywords:** affinity prediction, molecular representations, selectivity, transfer, censored data, reproducibility

**Posted Date:** November 20th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11529558/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Affinity panels dilute the target-specific signal that transfers to a held-out protein

Adeliya Leleytner[1*] and Maxim Fedorov[1]

[1]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Bolshoy Karetny per. 19, Moscow, 127051, Russia.

*Corresponding author(s). E-mail(s): a.leleytner@gmail.com;

**Abstract**

Virtual affinity panels are used as molecular descriptors, on the assumption that more columns describe a molecule more fully. We ask what a panel is worth when the protein in question was never calibrated. Eight scores from six programs — empirical scoring, learned rescoring and co-folding — were evaluated on 909 molecules and ten proteins against censored $IC_{50}$ measurements in their original assay contexts. The decision endpoint is a measured trade-off: a replacement that strengthens one interaction and weakens another.

For a protein excluded from calibration, that protein's own score recovered 17.8–34.2% of measured trade-offs, against 2.6–4.1% for a calibrated baseline of nonlinear chemistry and overall score level. All eight gained, by 13.7 to 31.0 points, every 95% interval from 400-draw chemical-group resampling excluding zero. The same scores given as a whole profile recovered 3.3–9.5%, three to five times less.

The ordering reverses on new molecules at calibrated proteins, where the scores themselves fell 8.6–14.9 points below that baseline and the whole profile matched it. Panel breadth did not supply the difference: the leading component explained 55–83% of score variance, a joint 102-column panel had an effective dimensionality of 1.63, 106 of 240 model selections chose zero profile weight, and acquisition curves ended 8.0 and 2.4 points below chemistry alone. How tightly a score concentrated tracked how much of it transferred, effective dimensionalities of 1.43–2.93 ordering the eight as their gains did (Spearman 0.80).

One combination escaped: for Nesso-1 on new molecules, the queried target's residual added 6.1 points [0.6, 12.3] over chemistry.

Trade-off information sits in the queried column; pooling columns into a profile destroys it.

**Scientific Contribution.** Profile descriptors are usually judged by pooled accuracy; we judge them by the measured two-direction trade-offs they are used to anticipate, separately for new molecules and for withheld proteins. The transferable signal sits in the queried target's own column and is lost when that column is pooled into a profile vector, by three to five fold across eight scores spanning three unrelated formulations. The released package assesses any panel by the trade-offs recovered per column purchased, not by its size.

**Keywords:** affinity prediction, molecular representations, selectivity, transfer, censored data, reproducibility

ORCID: Adeliya Leleytner, https://orcid.org/0009-0001-8214-2496; Maxim Fedorov, https://orcid.org/0000-0003-3901-3565.

# 1 Introduction

A virtual affinity panel promises more than a collection of binding scores. Its attraction is a biological description of a molecule: which proteins it prefers, which liabilities accompany its desired activity, and which chemical modification shifts that balance. Docking workflows and learned scoring models have made broad panels cheap enough that this promise is now routine practice, and panels are used directly as molecular descriptors [13, 32, 51, 72]. The same logic underlies target prediction from chemical similarity [29, 38], the large-scale prediction of side-effect targets [35], and the reuse of assay profiles as descriptors in other modalities [56]. DOCKSTRING supplies a broad reproducible docking matrix and demonstrates regression, screening and design applications, including a computational selectivity objective [20]. The second broad panel we use is the secondary pharmacology set assembled to flag adverse-event liabilities across pharmaceutical companies [7].

These studies establish that profiles can be useful. They do not establish *which* use survives which kind of generalization.

Two bodies of work bear on the question without answering it. The first concerns how well a score tracks affinity at all. Comparative assessments of scoring functions have run for two decades [11, 57, 70], learned scoring functions improve on empirical ones under some protocols [3, 46, 49], and co-folding models with a trained affinity head are the current frontier [2, 27, 71]. Whether structures predicted by such models help a learned affinity model is itself under test [23], and sequence-based predictors dispense with structure altogether [6]. Alongside this, careful studies find that deep networks often recover ligand-side regularities rather than the protein-ligand interaction itself [8, 18, 66, 73], and that generated poses can fail basic physical checks [9].

The second concerns how such claims are measured. Benchmark construction can reward memorisation instead of generalisation [55, 67], curated task collections try to standardise the comparison [24], and debiased datasets have been built specifically for compound-protein interaction prediction [22], decoy sets carry hidden bias [10, 41], the choice of split changes what is being tested [53], and evaluation guidance has become its own literature [5, 68]. Simple baselines repeatedly match complex models once the comparison is made carefully [21, 26]. A long vector of target-labelled scores can principally describe a property shared across targets, in which case adding proteins makes a profile look more detailed without adding the biological distinctions a decision requires.

The distinction matters most at the moment a panel is actually consulted. A medicinal chemist rarely asks whether a molecule binds more overall; that question is answered well by chemistry alone. The question that justifies the panel is directional and two-sided: does this replacement strengthen the intended interaction while weakening the off-target one? That is a trade-off, it is the object that work on interaction concordance formalizes [44], and it is the decision we evaluate here.

Recent work sharpens the setting. Representation quality depends on the task used to assess it [25]. Assay-neighbour profiles can recover much of the benefit of a larger

experimental panel for activity-cliff prediction [1], which suggests a small task-related subset carries the value of a large one. Evaluation design can reorder apparent model rankings [43]. A recent GPCR study documented compressed protein representations inside learned interaction models [14]. Together these make pooled accuracy, panel size and spectral dimensionality insufficient answers to the practical question: what does a profile let a researcher decide that chemistry and the molecule's overall scoring tendency do not?

We separate three things that are usually treated as one: the shape of a score matrix, what a fitted model can get out of it, and what the score says on its own. We hold the decision fixed and vary what the model is allowed to see — chemistry alone, chemistry plus the panel mean, chemistry plus the whole residual profile, chemistry plus the queried target's own residual column — and we evaluate each under two settings that are routinely conflated: a new molecule at a calibrated protein, and a protein withheld from calibration entirely.

The answer is a reversal. On new molecules the calibrated chemical model is the thing to beat, and the raw score alone is substantially worse than it. On a withheld protein the ordering inverts: the raw score of that protein recovers trade-offs that neither chemistry nor a model trained on the whole profile recovers, by a factor of three to five, for every one of eight scoring methods of unrelated formulation. The signal is carried by one column. Pooling columns into a profile dilutes it.

# 2 Results

## 2.1 A withheld protein's own score recovers trade-offs that a profile model loses

The decision endpoint is a measured trade-off. Given two molecules and two assay contexts, we ask whether replacing molecule A with molecule B strengthens one interaction and weakens the other, and we count a prediction correct only when it recovers both directions. Trade-offs are identified from the experimental records; measurements at a reporting limit retain their censoring relations and are never treated as inactivity.

We first examine proteins excluded from local calibration. The model may use every experimental label except those belonging to the queried protein, and the scores themselves are available throughout, since computing a score requires no measurement. The comparison is against a deliberately strong baseline: a nonlinear chemical kernel over Morgan fingerprints and six physicochemical descriptors, plus the molecule's mean score across the panel, which the model may use differently at different proteins. This baseline can express a change in relative preference; it is not a model forced to have zero interaction contrast.

In this setting the baseline is nearly helpless on trade-offs. It recovers 2.6–4.1% of them across the eight scores, because without any calibration at the queried protein there is nothing from which to produce a protein-specific differential.

That protein's own score, used as computed with no fitting, recovers 17.8–34.2%. The two Boltzina settings gave identical values at this precision and share a row, so the table has seven rows for eight scores; Table 2 is laid out the same way.

**Table 1** Trade-offs recovered on a protein withheld from calibration. Percentages are the fraction of identified trade-offs recovered; gains are against the chemistry-and-level baseline, with 2.5th–97.5th percentiles of 400 paired chemical-group resampling draws. The two Boltzina settings share a row.

| | Baseline | That protein's own score | | | Whole profile | |
|---|---|---|---|---|---|---|
| Score | % | % | gain | [95% interval] | % | gain |
| AutoDock4 | 3.3 | 17.8 | +14.5 | [+10.5, +18.4] | 3.3 | +0.0 |
| GNINA (own search) | 4.1 | 17.8 | +13.7 | [+8.6, +18.3] | 5.5 | +1.4 |
| Vina | 3.1 | 18.2 | +15.2 | [+10.4, +19.8] | 3.8 | +0.8 |
| GNINA (Vina pose) | 4.0 | 22.9 | +18.8 | [+15.6, +22.9] | 6.2 | +2.3 |
| Boltz-2 | 2.8 | 24.0 | +21.1 | [+15.1, +25.8] | 7.5 | +4.7 |
| Boltzina (far/zero) | 2.6 | 24.9 | +22.2 | [+15.7, +26.3] | 8.1 | +5.5 |
| Nesso-1 | 3.2 | 34.2 | +31.0 | [+24.1, +36.7] | 9.5 | +6.4 |

Every one of the eight gains is positive and every 95% interval from 400-draw paired chemical-group resampling excludes zero. The scores span three unrelated formulations — empirical scoring functions, a convolutional rescorer, and co-folding models with a learned affinity head — and the effect does not depend on which.

The last two columns are the point of the table. A model given exactly the same scores, arranged as a whole residual profile, recovers 3.3–9.5%. Three of its eight intervals include zero. The arrangement of the information, not its availability, decides whether the decision is recoverable.

The same ordering holds on the complementary measurement subset, where at least one of the four measurements sits at a reporting limit: gains for that protein's own score run +17.1 to +29.6 points, and for the whole profile +0.0 to +4.5, with the same eight-of-eight consistency. The result is therefore not an artefact of excluding censored records. The full table for that subset is Additional file 1, Table S1.

Direction accuracy, which asks only whether the preference moves the right way, behaves differently, and that difference is informative. For Boltz-2 it is 57.9% for the baseline, 60.3% for the model trained on the whole profile, 70.7% when the queried target's own residual column is supplied explicitly, and 70.4% for that protein's own score; for Nesso-1 the corresponding values are 58.4, 58.3, 70.5 and 71.0. Giving the

model that one column by name matches what the score itself achieves. Burying the same column inside a long vector does not.

## 2.2 The ordering reverses for new molecules at calibrated proteins

We repeat the comparison with the protein's experimental labels available and the molecule withheld. Here the baseline is strong in absolute terms, recovering 28.5–32.9% of trade-offs, because calibration at that protein lets chemistry carry the differential.

That protein's own score now falls below it, by 8.6 to 14.9 points for seven of eight scores; five of those intervals exclude zero. The whole-profile model neither helps nor hurts: its gains run −3.2 to +7.5 points and every interval includes zero.

**Table 2** Trade-offs recovered on new molecules at calibrated proteins. Columns and row grouping as in Table 1.

| Score | Baseline % | Own score % | gain | [95% interval] |
|---|---|---|---|---|
| AutoDock4 | 31.8 | 16.9 | −14.9 | [−21.8, −5.4] |
| Vina | 32.6 | 18.1 | −14.5 | [−22.9, −4.5] |
| GNINA (own search) | 32.9 | 19.0 | −13.9 | [−22.4, −3.4] |
| Boltzina (far/zero) | 31.7 | 20.0 | −11.7 | [−15.9, −0.9] |
| Boltz-2 | 31.0 | 21.5 | −9.5 | [−15.1, +1.5] |
| GNINA (Vina pose) | 32.1 | 23.5 | −8.6 | [−18.8, +0.1] |
| Nesso-1 | 28.5 | 36.4 | +7.8 | [−0.9, +15.7] |

A single model ordering therefore cannot serve both settings. Which arrangement of the same scores is best depends entirely on which axis is new, and the two settings are frequently reported under one heading of "generalization".

## 2.3 One combination escapes

Across the whole grid of scores, model variants and settings, exactly one combination shows the profile contributing beyond chemistry with an interval excluding zero: for Nesso-1 on new molecules, the queried target's own residual column adds 6.1 points

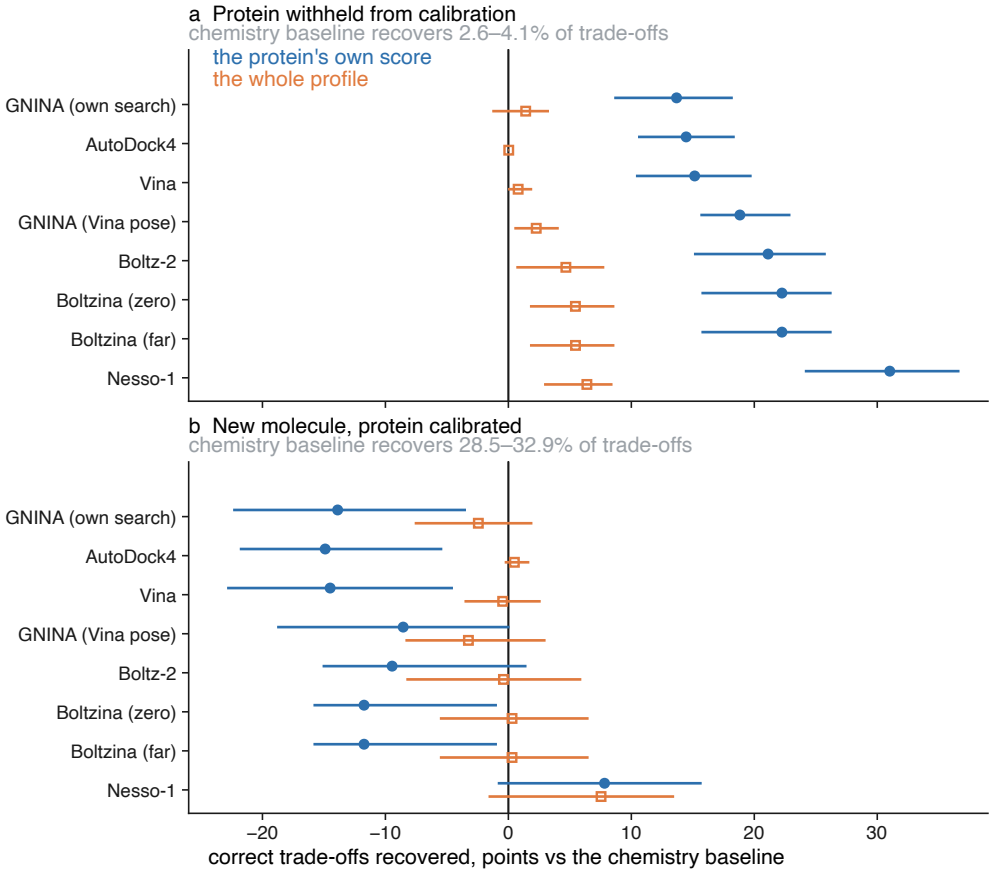


**Fig. 1 The reversal.** Correct trade-offs recovered, in percentage points against the chemistry-and-level baseline. **a** Protein withheld from calibration. **b** New molecule at a calibrated protein. Points are paired differences; bars are 2.5th–97.5th percentiles of 400 chemical-group resampling draws. Both panels share one horizontal scale, and each prints its own baseline value.

[0.6, 12.3]. Its direction accuracy rises in the same direction, from 75.4% for the baseline to 78.4% with the full profile and 79.2% with the matching-target residual.

This is the boundary condition of the paper's claim, and it is consistent with the mechanism rather than an exception to it. The variant that succeeds is again the one that names the query column; what differs is that Nesso-1's scores are informative enough at a calibrated protein for that column to add to chemistry rather than merely to substitute for calibration.

## 2.4 Why: the panel's variance is one shared molecular axis

Low effective dimensionality was a shared property of the score matrices. On the same 909 molecules with complete outputs, the eight scores had effective dimensionalities from 1.43 to 2.93, and their first principal components explained 55.1–83.3% of standardized score variance. Every score's leading component had loadings of the same sign across all ten targets: its dominant variation raises or lowers a molecule's scores across the whole panel at once.

The pattern is not confined to related proteins. Two complementary broad panels — DOCKSTRING, with 58 columns weighted towards kinases, enzymes, proteases and nuclear receptors, and Docking44, with 44 columns of the secondary-pharmacology panel [42] weighted towards GPCRs, ion channels and transporters — share nine gene labels and give 93 distinct genes across 102 columns. On the 384 molecular identities complete in both, the joint effective dimensionality is 1.63 and the leading component accounts for 78.3% of variance. Molecular row means from the two sources correlate at $r = 0.973$. Adding diverse target labels does not remove the shared molecular scoring tendency.

Removing each molecule's source-specific mean exposes a much broader residual spectrum: the joint effective dimensionality rises from 1.63 to 20.35 and the first-component fraction falls from 78.3% to 15.9%. But that richer geometry is uneven in agreement. On the 384 complete profiles, row-centered scores correlate at 0.678 for DRD2 and 0.613 for PTGS2, and at 0.114 for ADRB2 and 0.133 for LCK. A near-identical overall molecular ordering across panels coexists with much weaker agreement about particular targets.

This is the mechanism behind Section 2.1. The variance a profile contains is dominated by a molecule-level quantity that chemistry already predicts. Six ordinary descriptors recover 77% of a molecule's mean score across the joint panel under chemical-group cross-validation, and a median of 14% of what remains once that mean

is removed, column by column. The target-specific part that encodes a trade-off is a small residual; it differs between sources at the very targets where it matters; and a model asked to find it inside a long vector does not recover it.

## 2.5 How concentrated a score is tracks how much of it transfers

The eight scores differ in how tightly their variance concentrates, from an effective dimensionality of 1.43 for GNINA with its own search to 2.93 for Nesso-1, and they differ in how much each recovers on a withheld protein when used as that protein's own score, from +13.7 to +31.0 points. The two orderings agree. Spearman's rho between them is 0.80 on the exact four-cell subset and 0.88 on the censor-involving subset, and omitting any single score leaves it between 0.70 and 0.95 — so the association does not rest on Nesso-1, the one score whose effective dimensionality sits far from the rest.

We report this as an observation, not as a calibrated relationship. Six of the eight effective dimensionalities fall between 1.43 and 1.54, a spread we have not shown to exceed the resampling noise of the spectrum itself, and the eight scores include two GNINA configurations and two Boltzina variants that share method lineages, so they are not eight independent draws. The comparison that would settle it is a paired chemical-group bootstrap of the effective dimensionality under the scheme already used for the decisions, which we have not performed. Leave-one-out values and the per-score geometry are Additional file 1, Tables S2 and S3.

Should it hold, the consequence is practical. An effective dimensionality is computed from a score matrix alone and needs no experimental measurement of any kind, so it would serve as a pre-screen for which scorer to trust when a new target appears — the one setting in which, by Section 2.1, the choice of scorer decides whether the decision is recoverable at all.

## 2.6 The loss is a property of pooling, not of one profile model

Three controls separate the finding from the particular model used to read the profile.

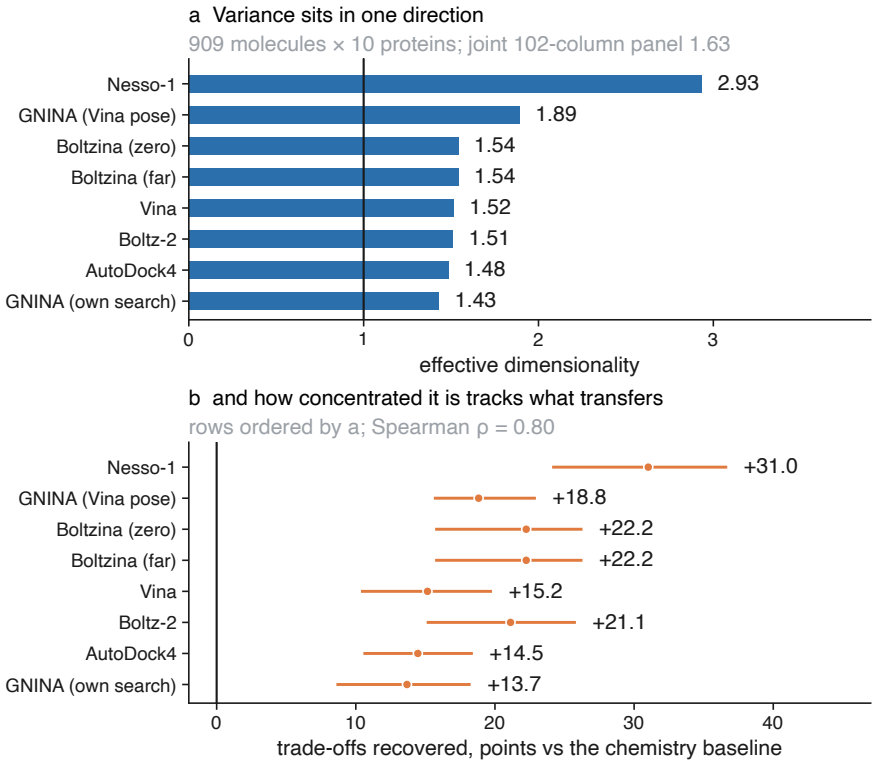


**Fig. 2 Concentration and transfer. a** Effective dimensionality of each score matrix on 909 molecules and ten proteins; the vertical line marks one direction. **b** What that protein's own score recovers on a withheld protein, rows in the order of **a**. Bars are 2.5th–97.5th percentiles of 400 chemical-group resampling draws.

First, the profile block's regularization strength was varied independently of the chemical block, with both linear and RBF representations and with candidates selected either by interval error or by preference-direction accuracy inside the training data. Across the 240 outer model selections, 106 selected exactly zero profile strength. When the full profile is allowed to be discarded, it is discarded in 44% of selections.

Second, allowing the profile to shrink separately removes most of the deterioration rather than converting it into a gain. In DOCKSTRING new-molecule evaluation the pointwise-selected linear variant reaches 82.9% direction accuracy and 40.4% correct trade-offs, against the baseline's 82.9% and 40.5%: the tuned profile model converges on the baseline, it does not beat it.

Third, the loss shows up in pooled error as well, not only in the decision endpoint. On Docking44 new molecules, adding the full profile moved exact-contrast RMSE from 1.257 to 1.290 $pIC_{50}$ units and direction accuracy from 64.3% to 61.8%. Both moved against the profile, and neither change separates from zero under the resampling: +0.033 [-0.024, +0.132] units and −2.5 points [-5.3, +3.1]. Choosing a decision endpoint over a regression metric is not what makes the profile look bad.

## 2.7 Panel breadth is not worth its cost

Acquisition curves were computed across the full panel-size axis, $k = 0 \ldots 58$ for DOCKSTRING and $k = 0 \ldots 44$ for Docking44, under three random and three family-stratified acquisition orders, with every mean, mask and residual computed from only the purchased columns and hyperparameters selected on chemistry alone inside each training fold. A model using chemistry alone recovers 43.4% of trade-offs on DOCKSTRING and 19.1% on Docking44.

Buying columns does not improve on chemistry alone; it does worse. Individual orders wander up to 6 points above chemistry alone at small panel sizes, but every order drifts below as columns accumulate, reaching 11.4 and 5.6 points under it at the worst point and settling 8.0 and 2.4 points under it at full panel size. The drift is common to random and family-stratified orders, so it is not an artefact of which columns are bought first. Biological diversity of the purchased columns is valuable for coverage; it does not buy decision value, and at these panel sizes it costs some.

By contrast, the query-specific comparators require one or two columns for a target pair and are shown separately from the shared-panel size axis. A single residual column selected per known assay raises correct trade-offs from 40.5% to 45.4% in DOCKSTRING and from 13.7% to 19.5% in Docking44, where the full profile gives 27.7% and 13.6%. The purchase that pays is the queried column, not the panel.

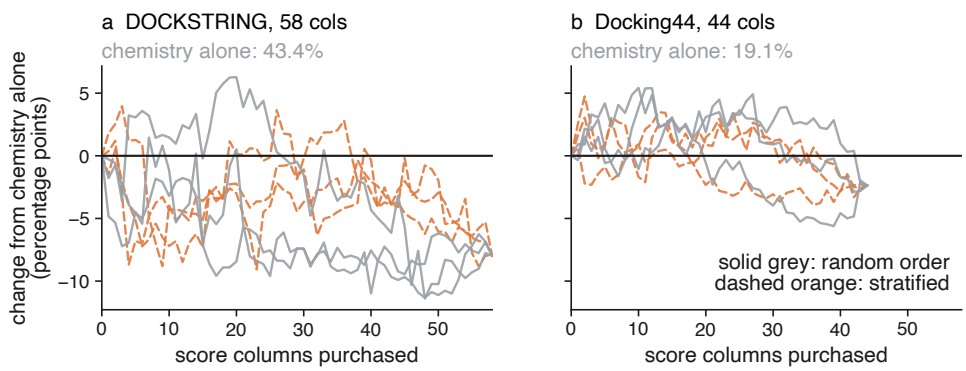


**Fig. 3 What buying columns costs.** Change in correct trade-offs against a model using chemistry alone as score columns are purchased, for new molecules at calibrated proteins. Each line is one acquisition order; the lines describe composition variation, not uncertainty. Both panels share one vertical scale, and each prints its own chemistry-only value.

## 2.8 The broad panels agree in direction with less precision

The decisive comparison was repeated on both broad panels. The direction is the same — that protein's own score gains over the baseline on withheld proteins, by 3.3 points on Docking44 and 6.7 points on DOCKSTRING — but the 95% intervals are [-2.1, +9.3] and [-3.3, +16.4] and therefore include zero.

The reason is support, not disagreement. The broad-panel trade-off analyses rest on 1,731 and 904 identified trade-offs across 61 and 79 chemical groups, against roughly 30,000 cases across 217 groups on the ten-target panel, because only 199 and 416 molecules in those panels have experimental measurements. We report the broad panels as a consistency check on target composition, and we do not treat them as independent confirmations of the effect size.

## 2.9 What to compute, by situation

The comparisons above resolve into a short routing rule. Section references give the evidence; percentages are changes against the baseline named in that section.

**Table 3** What to compute, by situation. Section references give the evidence.

| Situation | Compute | Do not | Section |
|---|---|---|---|
| New molecule, protein already calibrated | the calibrated chemical model | that protein's own score alone (8.6–14.9 points worse); the whole profile (no interval excludes zero) | 2.2 |
| Protein withheld from calibration | that protein's own score (+13.7 to +31.0 points, all eight intervals excluding zero) | a model trained on the whole profile, which recovers three to five times less | 2.1 |
| Choosing which scorer to run for a new target | prefer the score whose panel spectrum is least concentrated | ranking scorers by pooled accuracy on calibrated targets | 2.5 |
| Deciding how many columns to buy | the queried target's column | the broad panel, which ends 2.4–8.0 points below a model using chemistry alone | 2.7 |
| Nesso-1 scores, new molecules | the queried target's residual (+6.1 points [0.6, 12.3]) | — | 2.3 |

The rule that generalizes across the table: name the column you are asking about, and spend on that rather than on breadth.

---

# 3 Discussion

The practical consequence is a rule about what to compute when a new target appears. The information that answers a two-sided preference question about a protein is in that protein's own score, and the standard move of training a model on the whole profile loses most of it. Three to five fold is a large enough loss to change a workflow: it is the difference between recovering a sixth to a third of measured trade-offs and recovering under a tenth.

The reversal across settings explains why the literature has been able to report both that profiles are useful descriptors and that they add little. Both are true, under different held-out designs. A study that withholds molecules at calibrated proteins is measuring whether the profile adds to chemistry, and the honest answer is usually that it does not. A study that withholds proteins is measuring whether the scores

transfer, and there the scores carry something chemistry cannot produce — provided the queried column is not dissolved into a vector first. These two designs are routinely both called generalization, and a single model ranking derived from one does not carry to the other.

The spectral result should be read as the explanation, not the verdict. An effective dimensionality measures how score variance is distributed; it assigns no biological value to any direction. Row centering illustrates the gap precisely: it changes the spectrum from 1.63 to 20.35 while leaving the differences between targets, and hence the trade-off the score itself predicts, untouched, because subtracting the same molecular mean from every target cancels in a difference. Centering is a useful diagnostic decomposition whose consequences for a calibrated predictor must be measured rather than inferred from the resulting rank. The low dimensionality is why the trade-off signal is fragile; it is not itself evidence that the signal is absent. Related limits are familiar from activity-cliff prediction, where small structural changes carry large potency changes that models systematically miss [1, 60, 62], and from the ceiling that experimental error places on any model [54].

That said, the spectrum is not merely descriptive here. The order in which the eight scores concentrate their variance is the order in which they transfer (Section 2.5), which makes a quantity computable from scores alone a candidate pre-screen for a decision that otherwise needs held-out measurements to evaluate. We have deliberately stopped short of calling this a calibrated relationship, and Section 2.5 names the bootstrap that would decide it.

Our findings connect to several recent results. That a small task-related subset can recover much of a large panel's benefit [1] is the same phenomenon measured on a different endpoint: here the subset has size one and is named by the query. That representation quality is task-dependent [25] and that evaluation design reorders model rankings [43] both hold in our data, and Section 2.2 gives a concrete instance in which

the ordering inverts rather than merely shifts. Target-specific workflow choice is itself an active question, with consensus pipelines and algorithm-selection studies addressing which protocol to run for a given target [31, 69]; the present analysis instead asks what is gained by carrying additional targets into the molecular representation. The compression reported for learned interaction models [14] is consistent with the matching-target variant outperforming the model trained on the whole profile on withheld proteins: the limitation we observe is in how the score vector is presented to that model, and our result does not establish an information limit for other protein representations or training objectives.

The panel we use for the broad analysis was assembled to flag secondary-pharmacology liabilities [7, 42], and the association between antitarget docking scores and acute toxicity reported there is a different question from the one asked here. We note that our chemical baseline — Morgan fingerprints with molecular weight, logP, topological polar surface area, hydrogen-bond donors and acceptors and rotatable bonds — is precisely the comparator such an association requires, since a score threshold selects for size and lipophilicity, and that the present work supplies that comparator for the trade-off endpoint but not for the toxicity endpoint.

Limitations constrain the scope of the claim in four specific ways. The headline comparison rests on ten proteins, of which five are aminergic GPCRs and three are monoamine transporters; a withheld protein in this panel often has a close relative in calibration, and the broad-panel replication agrees in direction without the precision to settle the effect size. Local protein holdout excludes that protein's experimental labels from our calibration but does not establish its absence from the original scorers' training data, so pretrained-model exposure is a separate audit that we have not performed. The experimental records mix assay contexts and reporting limits. Heterogeneous public affinity data carry substantial measurement uncertainty, and mixed $IC_{50}$ values from different assays are only partly comparable [28, 30]; we retain actual

contexts and censoring rather than imposing a new activity boundary, and we report both measurement subsets, but assay heterogeneity remains. Finally, as Section 2.5 notes, the eight scores are not eight independent draws, so eight-of-eight consistency is weaker evidence than eight independent methods would be.

The remedy these results suggest is cheap. Before buying a panel, compute the chemistry-only model and the queried column, and measure what the remaining columns add on the decision and setting actually intended. This is the same discipline that validation practice has long asked of QSAR models [63] and that bioactivity benchmarks apply to model families [34]. Our release performs this assessment for any panel, reporting trade-offs recovered per column purchased.

---

# 4 Conclusions

Virtual affinity panels carry a dominant shared molecular component and a small target-specific residual. When a protein is withheld from calibration, that residual — reached through the protein's own column — recovers 17.8–34.2% of measured trade-offs against 2.6–4.1% for a strong chemical baseline, consistently across eight scores spanning three unrelated formulations. A model given the same scores pooled into a profile recovers 3.3–9.5%. On new molecules at calibrated proteins the ordering reverses and the chemical baseline is the thing to beat. Researchers evaluating a panel should state which axis is new, compare against a score-free chemical model, and keep the queried target's score as an explicit input rather than one coordinate of a long vector.

---

# 5 Additional files

**Additional file 1** — PDF. *Supplementary tables and notes.* Tables S1–S3: the main comparison on the censor-involving measurement subset; leave-one-out values of the rank correlation in Section 2.5; per-score effective dimensionality recomputed from the frozen score arrays. Two notes record which growth cohort the acquisition curves use and what the resampling does not cover.

**Additional file 2** — CSV. *Support accounting.* Possible, observed, identified, exact, tied and undefined comparisons for every assay in scope, with molecule, gene and assay counts, including strata with zero support.

**Additional file 3** — CSV. *Decision summaries.* The compact per-model summaries behind the main tables, across the ten-target panel and both broad panels, for each setting and weighting.

**Additional file 4** — CSV. *Acquisition curves and column costs.* Every curve point for k = 0…58 and k = 0…44, under three random and three family-stratified orders, with the acquired columns and the query-column costs.

**Additional file 5** — CSV. *Paired estimates and uncertainty by measurement subset.* Point estimates, paired differences and 2.5th/97.5th percentiles from 400 chemical-group resampling draws, separately for the exact four-cell and censor-involving subsets.

---

# 6 Methods

## 6.1 Study design and the panels used

We separate two things: the ordering a score produces on its own, and what that score's panel adds once a model has been fitted to experimental measurements. Every

comparison reuses score outputs that already existed. No new docking, model inference or pretraining was run for this work.

The experimental measurements come from the public secondary pharmacology resource of Sutherland et al. [59], released through Zenodo [58]. The molecular set keeps its frozen axis of 930 molecules and adds 469 molecules that have scores but no selected experimental labels. Those 469 are used only to describe the shape of score matrices. They are excluded from model fitting and from the normalisation that fitting uses.

Two broad panels are kept separate. DOCKSTRING has 58 target columns and 890 source rows, which map to 885 distinct molecules. Docking44 has 44 columns and 668 molecules. DOCKSTRING rows were kept when they overlapped the union of the frozen 930-molecule axis and the Docking44 axis; Docking44 rows were kept when they overlapped either the experimental axis or the selected DOCKSTRING rows. These are therefore linked cohorts, not an analysis of the whole DOCKSTRING library. DOCKSTRING comes from the released docking benchmark [19, 20]; Docking44 is a frozen export of the data behind Nikitin et al. [42], cited here for provenance. Of their molecules, 416 and 199 carry experimental labels, and 113 of those are shared. Two panels sharing a gene name does not mean they used the same structure, construct or docking settings, so we never average their columns together.

The smaller, experimentally labelled panel holds cached scores for 930 molecules and ten targets. Eight scores are compared: Boltz-2 affinity [45]; Boltzina affinity under its far and zero Vina settings [17]; GNINA CNNaffinity on a Vina pose [46]; Nesso-1 affinity [52]; AutoDock4 [40, 50]; GNINA CNNaffinity after its own search [36, 37]; and Vina [15, 61, 64]. These are eight scores from six programs, not eight independent method families: the two Boltzina scores are built on Boltz-2, the two GNINA scores share one scoring network [36], and both broad panels also use Vina. Energy-like

outputs are oriented so that a larger value means stronger predicted binding. We use the released or cached oriented scores without further clipping.

## 6.2 Experimental measurements and censoring

Measured affinities of this kind are curated at scale in ChEMBL [74]; the resource used here is the secondary pharmacology set of Sutherland et al. [59]. The binding data contain 33,050 records from 105 assays, grouped into 82 assay groups covering 74 human genes. Of these, 4,247 report an exact concentration and 28,803 report a strict "greater than" relation. Assay identifiers, source rows, relations, units and metadata stay attached to each observation. Grouping is used for annotation only; we never pool alternative measurements of the same assay. Functional measurements are held in a separate table and are not mixed with binding measurements.

For a source concentration $c$ in micromolar, we use $y = 6 - \log_{10} c$. An exact result gives the point interval $[y, y]$. A reported $\mathrm{IC}_{50} > c$ gives the open interval $(-\infty, 6 - \log_{10} c)$. This is an operational $\mathrm{pIC}_{50}$; it is not a conversion to $K_d$ or to a thermodynamic affinity. Missing outcomes stay missing. No active/inactive boundary is introduced anywhere in the analysis.

When one molecule has several source variants, we keep them as separate inputs whose weights sum to one for that molecule. Labels attach to each variant with those weights. We never average scores or decisions before evaluating them. A molecular pair always contains two different molecules, and the weights of the variant combinations of a pair sum to one. Column metadata records the nominal gene label, the current gene mapping, the organism and the protein context.

## 6.3 Molecular and protein representations

Each of the 930 molecules is represented by a 2,048-bit chirality-aware Morgan fingerprint of radius three [48], computed with RDKit [47], together with six descriptors in a fixed order: molecular weight, logP, topological polar surface area, hydrogen-bond

donors, hydrogen-bond acceptors and rotatable bonds. The existing Tanimoto matrix, chemical groups and fold assignments are reused unchanged. Groups follow Bemis-Murcko scaffolds [4]. There are 676 chemical groups; the 623 non-empty scaffolds do not straddle groups, and the 62 molecules with an empty or acyclic scaffold keep their earlier group.

Proteins are annotated with canonical human sequences from UniProt [65], covering 132 genes: 74 with experimental measurements and 58 more from the broad panels. All 93 gene labels in the panels map to a canonical sequence. A protein is represented by its amino-acid and dipeptide frequencies, of length 20 and 400; each block is scaled to unit length and then by $1/\sqrt{2}$. The first ten protein representations are inherited unchanged. A canonical sequence is a gene-level representation that transfers between settings. It does not certify the sequence, species, mutation, complex or construct actually used in an assay or a docking calculation.

## 6.4 Measuring how concentrated a score matrix is

This part of the analysis is descriptive and is kept separate from everything fitted to measurements. Each molecule carries equal total weight across its source variants. Analyses that need a complete profile keep only rows with every column present; analyses that tolerate gaps keep an explicit mask. We never fill a missing score to complete a profile. Four views are reported: raw scores, scores with each molecule's row mean removed, per-column ranks, and ranks followed by row-mean removal. Column moments use molecule weights, computed after whichever transformation is in force.

For a standardised score matrix $Z$ with normalised row weights $W$, the singular value decomposition of $W^{1/2}Z$ gives eigenvalues $\lambda_j$ proportional to the squared singular values. Writing $f_j = \lambda_j \,/\, \sum \lambda$, we report the full spectrum, the fraction in the first component, and $1/\sum_j f_j^2$, the effective number of directions the variance occupies. This quantity is also known as the participation ratio, and it is one of several

ways to summarise how many directions a dataset occupies [16]. Numerical rank uses a floating-point tolerance and is not a scientific cutoff. Joint summaries over two panels report equal weighting of columns and equal weighting of the two sources separately. Projections between panels are in-sample reconstructions across all retained directions; they are not held-out predictions, and they do not count independent biological dimensions.

To ask how much of the leading direction ordinary chemistry already carries, each molecule's mean score across the joint panel is regressed on the six source descriptors — molecular weight, cLogP, topological polar surface area, heavy atom count, ring count and rotatable bonds — by ordinary least squares, and we report the out-of-fold coefficient of determination over five folds that hold out whole chemical groups. The same model is fitted once per column to that column's score with the molecule's own panel mean removed, and the 102 out-of-fold values are summarised by their median. Like the rest of this subsection these fits are descriptive; nothing in the decision analysis uses them.

### 6.5 The baseline, and what each model is given in addition

For a purchased set of $k$ score columns, let $s_{ij}$ be an observed score, $g_i$ the mean of that molecule's observed columns, and $r_{ij} = s_{ij}$ - $g_i$ the remainder. The baseline is given nonlinear chemistry, $g_i$, coverage, and an availability indicator per column. All moments are estimated from training inputs only, using the source-variant weights. A missing standardised feature is encoded as zero alongside its availability indicator, which does not change the underlying missing score or outcome. Availability and residual blocks are scaled by $1/\sqrt{k}$.

The chemical kernel mixes Morgan–Tanimoto similarity with an RBF kernel over the six training-standardised descriptors:

$$K_{\text{chem}} = (1-\omega)K_{\text{Tanimoto}} + \omega K_{\text{RBF}}, \qquad K_{\text{RBF}}(i,j) = \exp\{-\|d_i - d_j\|^2/(2\times 6)\}.$$

Kernel centering and the positive-eigenvalue feature map use training molecules only. The protein kernel is the dot product of the sequence vectors. For a protein the model has seen, it is averaged with an assay-identity kernel so that each assay can have its own response and slope. For a protein held out of fitting, only the sequence kernel carries over, and its feature map is fitted on training assays.

With molecular features $x_i$, assay features $z_a$ and pair features $q_{ia\ell}$, the fitted model is

$$\hat{y}_{ia} = b + x_i^{\mathsf{T}}\beta + z_a^{\mathsf{T}}\gamma + x_i^{\mathsf{T}} W z_a + \sum_{\ell} q_{ia\ell}\left(\eta_\ell + z_a^{\mathsf{T}}\theta_\ell\right).$$

This is a proteochemometric form, in which molecule and protein descriptors enter together with their interaction [12]. So even the baseline can let the same mean score act differently at different proteins. It is not restricted to one global slope. This matters: it is why beating the baseline means something.

Four fitted models share this specification: the baseline alone; the baseline plus the queried target's own remainder; the baseline plus one selected remainder column; and the baseline plus the whole remainder profile. Every one of them is told whether the matching score is available. A pair-specific remainder uses one score scale fitted across the observed training molecule–assay entries. For a protein the model has seen, the single column is the one most strongly associated, in absolute value, with the training orders identified inside that assay; the regression then estimates its sign. Where labels cannot identify a training order, the fallback is the sequence-nearest column. For a held-out protein, the column is chosen by sequence similarity alone, with no held-out labels. Selection rules and fallback uses are recorded. A further stage compares

baseline, selected-column and whole-profile models across all 74 genes, including genes absent from any panel, which tests the profile as a molecular descriptor separately from recognising the named target.

Parameters minimise a weighted squared distance to each observed outcome interval, with a ridge penalty on everything except the intercept:

$$\frac{\sum_{ia} w_{ia}\ \mathrm{dist}(\hat{y}_{ia}, I_{ia})^2 + \alpha \|\vartheta_{\text{non-intercept}}\|_2^2}{\sum_{ia} w_{ia}}.$$

For an open interval the distance is its infimum, so a prediction sitting exactly at a censoring boundary can have zero loss without being a correct identified outcome. Unobserved cells carry zero weight. L-BFGS-B runs with at most 5,000 iterations, function tolerance $10^{-10}$, gradient tolerance $10^{-6}$ and at most 40 line-search steps. Failure to converge is a numerical error, not a scientific result. The score itself is kept as a separate, named comparison and receives no calibration of any kind.

## 6.6 Held-out settings and tuning

For new molecules at a protein the model has seen, we reuse five outer molecular folds that keep chemical groups intact. A test assay's gene must appear in training.

For a protein held out of fitting, every assay of that gene is removed before fitting, and each evaluated molecule must have an observed label at some training protein. Molecules failing that condition are counted, not quietly moved into the other setting. The ten-gene DOCKSTRING and small-panel comparisons use all 45 leave-two-genes-out blocks. The 31-gene Docking44 subset uses five deterministic gene blocks, and the all-genes stage uses five blocks over 74 genes. A four-cell contrast is evaluated only when both of its proteins were held out by the same fit. These schemes differ between datasets, and we report them as such rather than presenting them as one test.

A separate both-cold evaluation crosses five held-out molecular folds with five held-out gene blocks, giving 25 fits per source. Its results are reported on their own and are

never pooled with either main setting, because its pair coverage differs from exhaustive leave-two-genes-out evaluation.

Hyperparameters are chosen inside each outer training set using one predetermined inner molecular fold, or one deterministic inner gene block. This is nested training/validation selection, not full inner cross-validation, and it follows standard guidance on separating selection from evaluation [53, 63]. The baseline chooses $\omega \in \{0, 0.5\}$; each added block then chooses its own $\alpha \in \{0.1, 1, 10\}$ at that chemical weight, minimising inner interval loss, with ties broken deterministically. These grids and tolerances are algorithm settings, not thresholds of scientific importance. Every statement about something being new refers to local fitting only. Because the experimental data have been explored before, and because exposure inside the pretrained scoring models is unknown, these are not prospective tests and not pretraining-clean ones.

## 6.7 Sensitivity to the model and to how it was selected

After the first broad-panel comparison we ran a bounded, exploratory sensitivity check on the same 60 outer splits. The original baseline, selected-column and whole-profile predictions were kept exactly as they were, and the baseline's chemical weight and ridge penalty were held fixed. The added block was either the original linear remainders or a full RBF feature map over training-standardised remainders and availability indicators, with bandwidth denominator $2d$ for $d$ remainder and availability coordinates. Centering and the eigenbasis use training source rows only, weighted so each molecule has equal mass.

The added block was multiplied by $s \in \{0, 0.1, 1, 10\}$, which is equivalent to a block penalty $\alpha/s^2$ for $s > 0$. A strength of zero reproduces the baseline exactly. The same trained candidates were selected either by inner interval MSE or by direction accuracy on identified preference changes with equal weight per assay pair; ties choose

the smaller strength. Selected non-zero candidates were refitted on the outer training set, and identical selections share one fit. All four combinations of block type and selection objective are reported. This varies the representation and the selection rule, not the training loss, and we label it exploratory rather than describing it after the fact as preregistered.

## 6.8 Identifying a trade-off, and common support

Within one assay, two molecules have an identified order only when their experimental intervals force a strict ordering, open endpoints included. Exact ties and unidentified orders are counted separately. A predicted tie receives half credit for an identified binary order, and how often that happens is reported. Comparisons where all four measurements are exact, and comparisons where at least one is censored, are kept as separate subsets throughout. Fitted models are compared on identical source-variant support; a comparison that includes the score itself additionally requires that score to be finite on the same support. Variant outcomes are combined after orders or errors are computed, which preserves differences between variants that would be lost by averaging first.

For molecules $A$, $B$ and assays $t$, $u$, the four-cell change is

$$D = (y_{Bt} - y_{At}) - (y_{Bu} - y_{Au}).$$

For exact measurements the sign credit is that of the interaction concordance index [44]; our summaries add equal weighting per assay pair and the integration over source variants. Interval arithmetic identifies the sign of $D$ wherever that is possible. Products of exact source concentrations are compared in 60-digit decimal arithmetic before a logarithmic ratio is taken, which prevents false ties created by subtracting separately rounded logarithms. A constant offset between assays cancels in $D$; differences in assay scale, mechanism and context do not. We therefore describe these as operational

$IC_{50}$ contrasts, tag format, readout, technology, endpoint interpretation and species separately, and do not equate them with thermodynamic selectivity.

For a score used as computed, subtracting any molecular row mean cancels exactly from this contrast. Raw and row-centred versions of it are therefore not two independent experiments. Exact errors in $pIC_{50}$ units are reported for the fitted models. Summaries for scores used as computed rely on directions and associations, so that energy-like and log-affinity-like outputs can be compared at all. For the Nesso-1 illustration, its documented negative log-micromolar output differs from a nominal predicted $pIC_{50}$ by the constant six, which cancels from a change. Interval loss is a lower bound on error that is compatible with censoring. Association on exact contrasts uses the orientation-symmetric quantity $E(\hat{D}D)/\sqrt{E(\hat{D}^2)E(D^2)}$, retaining variant second moments.

We also record cases where the two measured potencies move in opposite directions. Predictions are classified as the correct joint direction, the reversed joint direction, a same-direction change at both assays, or a tie. No utility weights and no activity boundary are assigned.

## 6.9 Resampling and uncertainty

Every assay in the declared scope stays represented, including assays that contribute no identified pair. Support tables record possible, observed, identified, exact, tied and undefined comparisons, with molecule, gene and assay counts, and are exported in full as Additional file 2. Within-assay summaries pool held-out contributions inside each assay before averaging. Four-cell summaries are reported with equal weight per assay pair and, as a sensitivity, with equal weight per contrast. Molecules, assays, gene pairs and contrasts overlap each other, so none of these counts is an independent sample size.

Uncertainty is computed from the saved held-out predictions; the aim is to propagate the uncertainty already present in the measurements rather than to report a point estimate alone [39]. We draw chemical groups with replacement 400 times. Each contrast receives the product of the multiplicities of its two endpoint groups, or a single multiplicity when both molecules fall in the same group. A model and its baseline always use the same draws and the same finite support. Separate resampling of gene endpoints, and joint resampling of chemical groups and genes, measure sensitivity to panel composition; all assay alternatives for one gene move together. We report the original paired difference and the 2.5th and 97.5th percentiles of the draws. These distributions condition on fixed predictions and therefore exclude the variability of refitting. Where two models define different contexts, the contexts are intersected separately within each draw. Every paired estimate and its percentiles, by measurement subset, are exported as Additional file 5; the compact summaries behind the main tables are Additional file 3.

## 6.10 Acquisition curves

The acquisition stage compares the baseline and the whole-profile model at every $k = 0, \ldots, 58$ for DOCKSTRING and $k = 0, \ldots, 44$ for Docking44. Each source uses three random column orders and three family-stratified orders, with seeds 20261022, 20261023 and 20261024. Stratified acquisition cycles through randomised family classes and randomised columns inside each class. A chemistry-only inner fit chooses the kernel weight and the regularisation before any column is bought, and those choices are then reused across all $k$ and both models. The mean, remainders, coverage and masks at each step use only the columns bought so far. At $k = 0$ the model reads no score values and no availability masks, so it is a model using chemistry alone. At $k = 1$ the remainder is identically zero. Cost is counted in acquired columns. It is not a measured cost in money, GPU time or inference time.

Every curve point, acquisition order and column cost is exported as Additional file 4. Curves are reported on the cohort that matches the support of the main analysis. The data also contain a smaller metadata-matched cohort; the two are never averaged together, and the figures and text state which one is shown.

## 6.11 Comparing concentration with transfer

Two quantities are compared, one per score. The first is the effective number of directions from Section 5.4. The second is the gain of that protein's own score over the baseline from Section 5.8. We compute Spearman's rank correlation between them across the eight scores, with tied values given their mid-rank. It is reported for both measurement subsets, and we repeat it eight times, omitting one score each time, to show how much it depends on any single one. We do not resample the effective number of directions itself, and we say so where the result is discussed, because without that the ordering within a narrow band of values is not established.

## 6.12 Reproducibility

Prepared inputs use the schema `profile_information_inputs_v1`. Preparation ran on Python 3.12.11 with NumPy 1.26.4, pandas 2.3.2 and RDKit 2025.03.6. Fitting ran on Python 3.9.23 with NumPy 1.26.4, SciPy 1.13.1, pandas 2.3.1, scikit-learn 1.6.1 and RDKit 2023.09.6. Predictions, split identities, hyperparameter choices, solver diagnostics and support counts are saved together with hashes of the inputs, the code and the configuration, and a resumed run is accepted only when those signatures match. Where source scores, masks, rows, weights, target order and configuration are bit-identical, the run may share a recorded alias; such an alias is a record of identity, not an independent replication. Contrasts can be regenerated from the saved predictions without refitting, and the analysis code and prepared inputs are released and archived

with a DOI [33]. Independent computational checks cover interval arithmetic on synthetic cases, isolation of the splits, and the mapping back to source data. They are checks on the code, not a claim that a human has verified every number.

### 6.13 Figure construction

Numerical figures are generated directly from the archived tables and keep vector axes and text. No generative image model supplied any experimental observation or plotted number.

### 6.14 Use of artificial intelligence tools

OpenAI Codex (GPT-6 Astra) and Anthropic Claude were used to develop code for data preparation, computational experiments and figure generation, and to assist with drafting and editing the manuscript. All results were checked by the authors, who take full responsibility for the content.

## Declarations

### Availability of data and materials

The analysis code, prepared inputs and saved predictions are released [33]. The experimental source is the secondary pharmacology resource of Sutherland et al. [59], released through Zenodo [58]. The DOCKSTRING source is version 1 of the Figshare dataset [19, 20]. The Docking44 input is the frozen source export associated with Nikitin et al. [42].

### Competing interests

The authors declare no competing financial interests. M.F. is a coauthor of the publication associated with the Docking44 source data [42].

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.


## Funding

This work was supported by the Ministry of Economic Development of the Russian Federation, IGK 000000C313925P4C0002, agreement No. 139-15-2025-010.


## Authors' contributions

A.R.L.: conceptualization, methodology, software, investigation, formal analysis, writing. M.V.F.: supervision, conceptualization. All authors read and approved the final manuscript.


## Acknowledgements

Not applicable.


# References


[1] Abe R, Miyao T, Bajorath J (2026) Accurate prediction of activity cliff compounds based on bioactivity profiles depends on assay nearest neighbor relationships. Journal of Cheminformatics 18:72. https://doi.org/10.1186/s13321-026-01210-9

[2] Abramson J, Adler J, Dunger J, et al (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630(8016):493–500. https://doi.org/10.1038/s41586-024-07487-w

[3] Ballester PJ, Mitchell JBO (2010) A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 26(9):1169–1175. https://doi.org/10.1093/bioinformatics/btq112

[4] Bemis GW, Murcko MA (1996) The Properties of Known Drugs. 1. Molecular Frameworks. Journal of Medicinal Chemistry 39(15):2887–2893. https://doi.org/10.1021/jm9602928

[5] Bender A, Schneider N, Segler M, et al (2022) Evaluation guidelines for machine learning tools in the chemical sciences. Nature Reviews Chemistry 6(6):428–442. https://doi.org/10.1038/s41570-022-00391-9

[6] Bidgoli AH, Mahdavi M, Malek H (2026) Structure-free drug–target affinity prediction using protein and molecule language models. Journal of Cheminformatics 18:21. https://doi.org/10.1186/s13321-025-01146-6

[7] Bowes J, Brown AJ, Hamon J, et al (2012) Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. Nature Reviews Drug Discovery 11(12):909–922. https://doi.org/10.1038/nrd3845

[8] Boyles F, Deane CM, Morris GM (2019) Learning from the ligand: using ligand-based features to improve binding affinity prediction. Bioinformatics 36(3):758–764. https://doi.org/10.1093/bioinformatics/btz665

[9] Buttenschoen M, Morris GM, Deane CM (2024) PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chemical Science 15(9):3130–3139. https://doi.org/10.1039/d3sc04185a

[10] Chen L, Cruz A, Ramsey S, et al (2019) Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PLOS ONE 14(8):e0220113. https://doi.org/10.1371/journal.pone.0220113

[11] Cheng T, Li X, Li Y, et al (2009) Comparative Assessment of Scoring Functions on a Diverse Test Set. Journal of Chemical Information and Modeling 49(4):1079–1093. https://doi.org/10.1021/ci9000053

[12] Cortes-Ciriano I, van Westen GJ, Lenselink EB, et al (2014) Proteochemometric modeling in a Bayesian framework. Journal of Cheminformatics 6(1). https://doi.org/10.1186/1758-2946-6-35

[13] Cortés-Ciriano I, Škuta C, Bender A, et al (2020) QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction. Journal of Cheminformatics 12:41. https://doi.org/10.1186/s13321-020-00444-5

[14] Dey S, Lu P, Wallqvist A, et al (2026) Evaluation of deep learning architectures for predicting ligand interactions with neurologically relevant GPCRs. Journal of Cheminformatics 18:139. https://doi.org/10.1186/s13321-026-01258-7

[15] Eberhardt J, Santos-Martins D, Tillack AF, et al (2021) Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. Journal of chemical information and modeling 61(8):3891–3898. https://doi.org/10.1021/acs.jcim.1c00203

[16] Facco E, d'Errico M, Rodriguez A, et al (2017) Estimating the intrinsic dimension of datasets by a minimal neighborhood information. Scientific Reports 7(1). https://doi.org/10.1038/s41598-017-11873-y

[17] Furui K, Ohue M (2025) Boltzina: Efficient and Accurate Virtual Screening via Docking-Guided Binding Prediction with Boltz-2. In: AI4Mat Workshop at NeurIPS, https://doi.org/10.48550/arXiv.2508.17555, URL https://openreview.net/forum?id=OwtEQsd2hN

[18] Gabel J, Desaphy J, Rognan D (2014) Beware of Machine Learning-Based Scoring Functions—On the Danger of Developing Black Boxes. Journal of Chemical Information and Modeling 54(10):2807–2815. https://doi.org/10.1021/ci500406k

[19] Garcia Ortegon M, Tripp A (2022) dockstring dataset. https://doi.org/10.6084/m9.figshare.16511577.v1, version 1. Accessed 22 October 2026

[20] García-Ortegón M, Simm GNC, Tripp AJ, et al (2022) DOCKSTRING: Easy Molecular Docking Yields Better Benchmarks for Ligand Design. Journal of Chemical Information and Modeling 62(15):3486. https://doi.org/10.1021/acs.jcim.1c01334

[21] Gil-Sorribes M, Molina A (2026) Addressing model overcomplexity in drug-drug interaction prediction with molecular fingerprints. Journal of Cheminformatics 18(1). https://doi.org/10.1186/s13321-025-01128-8

[22] Hao Y, Li B, Huang D, et al (2025) Benchmarking Sequence-Based Compound–Protein Interaction Prediction through Constructing a Debiased Data Set CDPN. Journal of Chemical Information and Modeling 65(23):12737. https://doi.org/10.1021/acs.jcim.5c02040

[23] Hsu WT, Grevtsev S, Herz AM, et al (2025) Can AI-Predicted Complexes Teach Machine Learning to Compute Drug Binding Affinity? Journal of Chemical Information and Modeling 65(24):13051–13056. https://doi.org/10.1021/acs.jcim.5c01848

[24] Huang K, Fu T, Gao W, et al (2022) Artificial intelligence foundation for therapeutic science. Nature Chemical Biology 18(10):1033–1036. https://doi.org/10.1038/s41589-022-01131-2

[25] Huber F, Pollmann J (2026) Count your bits: fingerprint benchmarking to assess broad chemical space representation. Journal of Cheminformatics 18:83. https://doi.org/10.1186/s13321-026-01245-y

[26] Janela T, Bajorath J (2022) Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models. Nature Machine Intelligence 4(12):1246–1255. https://doi.org/10.1038/s42256-022-00581-6

[27] Jumper J, Evans R, Pritzel A, et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2

[28] Kalliokoski T, Kramer C, Vulpetti A, et al (2013) Comparability of Mixed IC50 Data – A Statistical Analysis. PLoS ONE 8(4):e61007. https://doi.org/10.1371/journal.pone.0061007

[29] Keiser MJ, Setola V, Irwin JJ, et al (2009) Predicting new molecular targets for known drugs. Nature 462(7270):175–181. https://doi.org/10.1038/nature08506

[30] Kramer C, Kalliokoski T, Gedeck P, et al (2012) The Experimental Uncertainty of Heterogeneous Public ¡i¿K¡/i¿¡sub¿i¡/sub¿ Data. Journal of Medicinal Chemistry 55(11):5165–5173. https://doi.org/10.1021/jm300131x

[31] Lacour A, Agha H, Hirsch AKH, et al (2026) DockM8: an all-in-one open-source platform for consensus virtual screening in drug design. Journal of Cheminformatics 18:135. https://doi.org/10.1186/s13321-026-01287-2

[32] Laufkötter O, Sturm N, Bajorath J, et al (2019) Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold hopping capability. Journal of Cheminformatics 11:54. https://doi.org/10.1186/

s13321-019-0376-1

[33] Leleytner A, Fedorov M (2026) Affinity panels dilute the target-specific signal that transfers to a held-out protein: reproducibility code, data and figures. https://doi.org/10.5281/zenodo.23692523, URL https://github.com/AdeliyaLeleytner/virtual-affinity-profiles, Zenodo, version 1.3.0

[34] Lenselink EB, ten Dijke N, Bongers B, et al (2017) Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. Journal of Cheminformatics 9(1). https://doi.org/10.1186/s13321-017-0232-0

[35] Lounkine E, Keiser MJ, Whitebread S, et al (2012) Large-scale prediction and testing of drug activity on side-effect targets. Nature 486(7403):361–367. https://doi.org/10.1038/nature11159

[36] McNutt AT, Francoeur P, Aggarwal R, et al (2021) GNINA 1.0: molecular docking with deep learning. Journal of Cheminformatics 13:43. https://doi.org/10.1186/s13321-021-00522-2

[37] McNutt AT, Li Y, Meli R, et al (2025) GNINA 1.3: the next increment in molecular docking with deep learning. Journal of Cheminformatics 17(1). https://doi.org/10.1186/s13321-025-00973-x

[38] Mervin LH, Afzal AM, Drakakis G, et al (2015) Target prediction utilising negative bioactivity data covering large chemical space. Journal of Cheminformatics 7(1). https://doi.org/10.1186/s13321-015-0098-y

[39] Mervin LH, Trapotsi MA, Afzal AM, et al (2021) Probabilistic Random Forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty. Journal of Cheminformatics 13(1). https:

//doi.org/10.1186/s13321-021-00539-7

[40] Morris GM, Huey R, Lindstrom W, et al (2009) Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. Journal of computational chemistry 30(16):2785–2791. https://doi.org/10.1002/jcc.21256

[41] Mysinger MM, Carchia M, Irwin JJ, et al (2012) Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. Journal of Medicinal Chemistry 55(14):6582–6594. https://doi.org/10.1021/jm300687e

[42] Nikitin I, Morgunov I, Safronov V, et al (2025) Towards Explainable Computational Toxicology: Linking Antitargets to Rodent Acute Toxicity. Pharmaceutics 17(12):1573. https://doi.org/10.3390/pharmaceutics17121573

[43] Özçelik R, Grisoni F (2025) How evaluation choices distort the outcome of generative drug discovery. Journal of Cheminformatics 17:169. https://doi.org/10.1186/s13321-025-01108-y

[44] Pahikkala T, Numminen R, Movahedi P, et al (2025) Interaction Concordance Index: Performance Evaluation for Interaction Prediction Methods. arXiv:2510.14419, https://doi.org/10.48550/arXiv.2510.14419, preprint, version 1

[45] Passaro S, Corso G, Wohlwend J, et al (2025) Boltz-2: Towards accurate and efficient binding affinity prediction. bioRxiv https://doi.org/10.1101/2025.06.14.659707

[46] Ragoza M, Hochuli J, Idrobo E, et al (2017) Protein–Ligand Scoring with Convolutional Neural Networks. Journal of Chemical Information and Modeling 57(4):942–957. https://doi.org/10.1021/acs.jcim.6b00740

[47] RDKit (2026) RDKit: Open-Source Cheminformatics Software. URL https://www.rdkit.org/, versions used are recorded in the reproducibility package. Accessed 22 October 2026

[48] Rogers D, Hahn M (2010) Extended-connectivity fingerprints. Journal of chemical information and modeling 50(5):742–754. https://doi.org/10.1021/ci100050t

[49] Rose D, Seidel T, Langer T (2026) GRIPHIN: grids of pharmacophore interaction fields for affinity prediction. Journal of Cheminformatics 18:97. https://doi.org/10.1186/s13321-026-01203-8

[50] Santos-Martins D, Solis-Vasquez L, Tillack AF, et al (2021) Accelerating autodock4 with gpus and gradient-based local search. Journal of chemical theory and computation 17(2):1060–1073. https://doi.org/10.1021/acs.jctc.0c01006

[51] Schifferstein J, Bernatavicius A, Janssen AP (2024) Docking-Informed Machine Learning for Kinome-wide Affinity Prediction. Journal of Chemical Information and Modeling 64(24):9196–9204. https://doi.org/10.1021/acs.jcim.4c01260

[52] Shenoy N, Errington D, Bengio E, et al (2026) Nesso-1: Accelerating open-source binding affinity predictions. bioRxiv https://doi.org/10.64898/2026.08.01.742196, URL https://www.biorxiv.org/content/10.64898/2026.08.01.742196v1

[53] Sheridan RP (2013) Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. Journal of Chemical Information and Modeling 53(4):783–790. https://doi.org/10.1021/ci400084k

[54] Sheridan RP, Karnachi P, Tudor M, et al (2020) Experimental Error, Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure–Activity Relationship Models? Journal of Chemical Information and Modeling 60(4):1969–1982. https://doi.org/10.1021/acs.jcim.9b01067

[55] Sieg J, Flachsenberg F, Rarey M (2019) In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. Journal of Chemical Information and Modeling 59(3):947–961. https://doi.org/10.1021/acs.jcim.8b00712

[56] Simm J, Klambauer G, Arany A, et al (2018) Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. Cell Chemical Biology 25(5):611–618.e3. https://doi.org/10.1016/j.chembiol.2018.01.015

[57] Su M, Yang Q, Du Y, et al (2018) Comparative Assessment of Scoring Functions: The CASF-2016 Update. Journal of Chemical Information and Modeling 59(2):895–913. https://doi.org/10.1021/acs.jcim.8b00545

[58] Sutherland J, Yonchev D, Fekete A, et al (2022) A novel preclinical secondary pharmacology resource illuminates target-adverse drug reaction associations of marketed drugs: Supplementary material. https://doi.org/10.5281/zenodo.8103950, accessed 22 October 2026

[59] Sutherland JJ, Yonchev D, Fekete A, et al (2023) A preclinical secondary pharmacology resource illuminates target-adverse drug reaction associations of marketed drugs. Nature Communications 14:4323. https://doi.org/10.1038/s41467-023-40064-9

[60] Tamura S, Miyao T, Bajorath J (2023) Large-scale prediction of activity cliffs using machine and deep learning methods of increasing complexity. Journal of Cheminformatics 15:4. https://doi.org/10.1186/s13321-022-00676-7

[61] Tang S, Ding J, Zhu X, et al (2024) Vina-gpu 2.1: Towards further optimizing docking speed and precision of autodock vina and its derivatives. IEEE/ACM

transactions on computational biology and bioinformatics 21(6):2382–2393. https://doi.org/10.1109/tcbb.2024.3467127

[62] van Tilborg D, Alenicheva A, Grisoni F (2022) Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. Journal of Chemical Information and Modeling 62(23):5938–5951. https://doi.org/10.1021/acs.jcim.2c01073

[63] Tropsha A (2010) Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics 29(6-7):476–488. https://doi.org/10.1002/minf.201000061

[64] Trott O, Olson AJ (2010) Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry 31(2):455–461. https://doi.org/10.1002/jcc.21334

[65] UniProt Consortium (2025) Uniprot: the universal protein knowledgebase in 2025. Nucleic acids research 53(D1):D609–D617. https://doi.org/10.1093/nar/gkae1010

[66] Volkov M, Turk JA, Drizard N, et al (2022) On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. Journal of Medicinal Chemistry 65(11):7946–7958. https://doi.org/10.1021/acs.jmedchem.2c00487

[67] Wallach I, Heifets A (2018) Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. Journal of Chemical Information and Modeling 58(5):916–932. https://doi.org/10.1021/acs.jcim.7b00403

[68] Walters WP, Barzilay R (2021) Critical assessment of AI in drug discovery. Expert Opinion on Drug Discovery 16(9):937–947. https://doi.org/10.1080/17460441.2021.1915982

[69] Wang JB, Cao S, Wu H, et al (2026) Molecular embedding-based algorithm selection in protein-ligand docking. Journal of Cheminformatics 18:47. https://doi.org/10.1186/s13321-026-01168-8

[70] Warren GL, Andrews CW, Capelli AM, et al (2005) A Critical Assessment of Docking Programs and Scoring Functions. Journal of Medicinal Chemistry 49(20):5912–5931. https://doi.org/10.1021/jm050362n

[71] Wohlwend J, Corso G, Passaro S, et al (2024) Boltz-1 Democratizing Biomolecular Interaction Modeling. bioRxiv https://doi.org/10.1101/2024.11.19.624167

[72] Wu J, Chen Y, Wu J, et al (2024) Large-scale comparison of machine learning methods for profiling prediction of kinase inhibitors. Journal of Cheminformatics 16:13. https://doi.org/10.1186/s13321-023-00799-5

[73] Yang J, Shen C, Huang N (2020) Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets. Frontiers in Pharmacology 11. https://doi.org/10.3389/fphar.2020.00069

[74] Zdrazil B, Felix E, Hunter F, et al (2023) The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research 52(D1):D1180–D1192. https://doi.org/10.1093/nar/gkad1004

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfile1.pdf
- additionalfile2.csv
- additionalfile3.csv
- additionalfile4.csv
- additionalfile5.csv