DARPINBENCH: A DATASET AND LIBRARY-SCREENING BENCHMARK FOR DESIGNED ANKYRIN REPEAT PROTEINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pulling a handful of experimentally confirmed binders to the front of a library whose members differ by one or two point mutations is a recurring machine-learning problem, from antibody affinity maturation to the triage step of a generative design pipeline. Designed ankyrin repeat proteins (DARPins) instantiate it: one shared solenoid scaffold, sparse positives, and negatives that are near neighbors. Public sources keep sequences, complex coordinates, and binding readouts apart, so neither an auditable table nor a frozen ranking protocol exists for this scaffold. We release both. DARPinDB annotates sequence, binding label, K_{D}, and evidence source on a single record: 188 records, 179 binders, and 37 targets, of which 121 carry a numeric K_{D}, with structures tiered and released with chain maps. DARPinBench freezes three screening modes on that table. Mode 1 scores ten mainstream methods from the sequence, structure, rescoring, and docking families on ten near-neighbor libraries of 500 sequences each; Modes 2 and 3 release a shared 20,000-member library and a leave-one-target-out split as held-out material, unscored here. No method reaches the top 1% of a library reliably: the best rate at the top 1% is 0.30 of ten units, AF3 attains the highest recall at the top 10% (0.80) while reaching only 0.10 at the top 1%, and the median enrichment factor at 5% is zero for eight of the ten methods. A high DockQ does not travel with a front-of-library rank. We introduce no new scorer; the contribution is a frozen reference line, together with a diagnosis of where off-the-shelf scores fail, for later evolution and training to exceed.

1 INTRODUCTION

Computational screening of a designed binder library is a ranking problem under an unusual constraint. The candidates differ from one another by one or two point mutations on a shared scaffold, and only a handful of them bind. A useful method has to put those few at the very front of the list, because the wet-lab budget reaches tens of candidates while the library holds thousands. The same shape recurs in antibody affinity maturation, in directed evolution campaigns, and in the triage step of any generative design pipeline.

Designed ankyrin repeat proteins (DARPins) instantiate this problem cleanly. An N-terminal cap, a stack of ankyrin repeats, variable positions on the concave face, and a C-terminal cap form a stable solenoid; binding diversity sits at randomized concave positions while framework residues stay fixed under consensus design (Binz et al., 2003; Kohl et al., 2003; Plückthun, 2015). Randomly drawn library members remain soluble and thermodynamically stable, and crystal structures show a regular consensus fold with small backbone deviations between repeats (Kohl et al., 2003). A DARPin is selected from such a library rather than predicted as an arbitrary globular heterodimer (Binz et al., 2004; Steiner et al., 2008). Positives are rare; negatives are near neighbors.

Existing resources answer adjacent questions. Antibodies and nanobodies have jointly annotated sequence–structure resources (Dunbar et al., 2014; Shrestha et al., 2025); generic complexes and mutational affinities are compiled in PDBbind and SKEMPI 2.0 (Liu et al., 2015; Jankauskaitė et al., 2019); CASP15, CAPRI-Q, and all-atom suites such as FoldBench ask whether a predicted pose returns to experimental coordinates (Kryshtafovych et al., 2023; Collins et al., 2024; Xu et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2025). None of them asks which member of a shared-scaffold library is the binder, and reading success at generic heterodimer prediction as evidence of library-screening ability skips exactly that difference.

The data gap is more specific still. The PDB contains DARPin–target co-crystals but also DARPin-only monomers that supply scaffold coordinates and no binding interface (Kohl et al., 2003); crystal entries typically carry neither library-screening labels nor a numeric K_{D} with a literature source, and treating a monomer as a complex puts later docking and pose evaluation on the wrong structural premise. Curated binder databases and paper supplements provide sequences and binding readouts, yet structures and chain correspondences have to be recovered from coordinate files. Learning methods are therefore left with sequence labels without structure, complex coordinates without screening semantics, or ad hoc joins across supplements. What is missing is one public table that carries sequence, structure where it exists, affinity, and evidence source on the same record, and that keeps complex, target-only, and structure-free entries apart.

Without that table a frozen library-ranking protocol cannot be defined, and later learning methods have no reproducible training and comparison benchmark. Existing binder or PPI benchmarks target antibodies or generic interactions, or classify against synthetic negatives rather than ranking sparse positives among near-neighbor decoys (Liu et al., 2026; Zhang et al., 2025). Conformational success on CASP/CAPRI does not imply that the same model retrieves an experimentally labeled positive from 500 near neighbors, and testing that implication needs a protocol placing both questions on the same $(\mathrm{DARPin}, \mathrm{target})$ pair. No such protocol has been available.

We therefore release the data product and the evaluation protocol separately (Figure 1). **DARPinDB** annotates sequence, binding label, K_{D} when available, and evidence source on the same record: 188 records, 179 binders, and 37 targets, of which 121 have a numeric K_{D}. Structures are tiered as complex, target-only, or none, and released with coordinates and frozen chain maps, so that later methods train on consistent label and structure premises. **DARPinBench** instantiates screening as three modes. Mode 1 is a controlled near-neighbor ranking on nine biological targets with experimental co-crystals. HER2 is split by the subdomain I / IV epitopes into two computational units (Steiner et al., 2008; Jost et al., 2013), which yields ten libraries of 500 sequences. On those we evaluate ten mainstream methods from the sequence, structure, rescoring, and docking families and report frozen scores. Mode 2 uses one shared library of 20,000 DARPins and Mode 3 performs leave-one-target-out on it (≈200,000 pairs); both release sequences and labels only.

Evaluating four families on the same near-neighbor decoys puts inexpensive sequence scores, fold-model confidence, and energy-like docking or rescoring scores on one ranking. DockQ on a crystal subset asks a separate mechanistic question: whether a method that recovers the pose also ranks that binder at the front of the 500-member library.

Contributions.

- **DARPinDB**, a record-level table that jointly annotates DARPin sequence, target, binding label, K_{D} where available (121 of 188 records), and a structure tier with frozen chain maps, contact-verified for the 13 crystal cases released with coordinates.
- **DARPinBench**, a frozen sparse-positive ranking protocol on near-neighbor libraries, in three screening modes: a scored near-neighbor screen (Mode 1), shared-library panning (Mode 2), and leave-one-target-out generalization (Mode 3). Mode 1 carries the frozen scores for the baseline evaluation in this work; Modes 2 and 3 are released as held-out material for later evolution and training.
- A frozen ten-method reference line across the sequence, structure, rescoring, and docking families. Off-the-shelf scores do not place experimental positives at the front of these libraries, and we diagnose where they fail. We do not propose a new scoring architecture.

2 Related Work

2.1 Binder databases and screening benchmarks

Existing binder resources are annotated for other scaffolds and scored on other tasks. Antibodies and nanobodies have jointly annotated sequence–structure resources such as SAbDab and SAbDab-

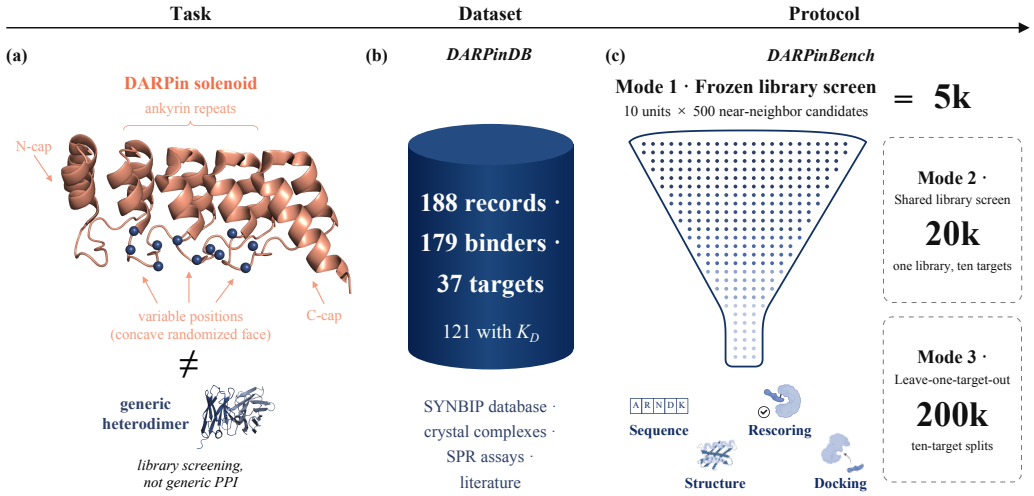

Figure 1: Overview of DARPinBench. (a) Ankyrin-repeat library screening versus generic heterodimer interaction prediction. (b) DARPinDB: 188 records over 37 targets, each carrying its evidence class and structure tier. (c) Protocol: Mode 1 is a frozen ten-method screen on 10×500 near-neighbor libraries; Mode 2 (20k shared-library panning) and Mode 3 (200k, leave-one-target-out) are released without model scores.

nano (Dunbar et al., 2014; Schneider et al., 2022), and generic complexes and mutational affinities are compiled in PDBbind, SKEMPI 2.0, and later PPI affinity sets such as PPB-Affinity (Liu et al., 2015; Jankauskaitė et al., 2019; Liu et al., 2024). SKEMPI 2.0 records mutational effects on closely related backbones, which is not a DARPin combinatorial-library screening table. DARPin sequences themselves stay scattered across public resources such as the SYNBIP database (Li et al., 2025) and paper supplements, which rarely carry a verifiable structure tier. Hand-joining those sources does not yield an auditable record-level table; **DARPinDB** is that table, released with a structure tree and chain maps.

Binder design and PPI classification already have benchmarks for nanobody binding, designed binders, and proteome-scale interaction (Shrestha et al., 2025; Liu et al., 2026; Zhang et al., 2025), and virtual-screening decoy construction, as in DUD-E, can inflate enrichment when decoys differ from actives in bulk properties rather than in binding (Mysinger et al., 2012; Chen et al., 2019). Those tasks classify, regress affinity, or score design success, with negatives from property matching or random cross-target draws rather than a one-residue neighbor on the same DARPin scaffold (Xu et al., 2026). Selecting candidates from a generated pool is also treated as ranking under a fixed budget (Lee et al., 2026), but there the pool comes from a design generator and the labels are proxy scores. Reading either kind of ranking as DARPin library-screening ability skips the scaffold constraint.

2.2 Structure prediction and docking evaluation

Structure assessments score proximity to coordinates, not rank in a library. CASP15, CAPRI-Q, and FoldBench measure whether a pose returns to experimental coordinates (Kryshtafovych et al., 2023; Collins et al., 2024; Xu et al., 2025). A high DockQ says the pose is close to the crystal, not that the same DARPin rises into the top five of 499 neighbors. Those confidence outputs are increasingly reused as objectives, including as an energy function for binder design and virtual screening (Nori et al., 2025), which makes what they rank the more pressing question.

2.3 Computational DARPin design and screening

Consensus ankyrin design fixes framework residues and places diversity at predetermined interaction sites (Binz et al., 2003; Kohl et al., 2003); full-consensus folding studies describe a highly stable consensus fold, not point mutations remodeling the backbone (Wetzel et al., 2008), and the scaffold

and its uses have been reviewed (Plückthun, 2015). High-affinity binders have been selected from such libraries (Binz et al., 2004; Steiner et al., 2008), and HER2 cannot be treated as one epitope: fragments lacking versus containing domain IV select distinct classes, later mapped to subdomain I (4HRL, 4HRM) and subdomain IV (4HRN) (Steiner et al., 2008; Jost et al., 2013). Scaffold similarity makes homology modeling straightforward on the DARPin side, yet the target can undergo large induced fit, as for BCL-W and APH, so docking is not uniformly rigid-body (Radom et al., 2019); public rigid docking and SPR tables exist for BCL-W (Gaudreault et al., 2023). The scaffold remains in active use: DARPins against TBXT dissect its regulatory network in chordoma (Umbaugh et al., 2025), and analyses of DARPin-based binding moieties explain what makes a bioPROTAC efficient (Winkelvoß et al., 2025). Those studies address single targets; to our knowledge no public DARPin combinatorial-library package jointly provides experimental binder/nonbinder labels and a frozen near-neighbor protocol.

2.4 Sequence PPI and affinity scoring

Sequence PPI and affinity models include D-SCRIPT, ESMDNN-PPI, ProtT5-embedding regression, and PLM-interact (Sledzieski et al., 2021; Li et al., 2024; Lou, 2025; Liu et al., 2025), with ProtT5 from ProtTrans (Elnaggar et al., 2022); the structure side contributes AF3, Boltz-2, ESMFold, and ESMFold2 (Abramson et al., 2024; Passaro et al., 2025; Lin et al., 2023; Candido et al., 2026). Neither a sequence PPI score nor a fold confidence can be read directly as a DARPin affinity prediction.

3 Dataset and Benchmark

3.1 DARPinDB

The release unit of DARPinDB is a $(\mathrm{DARPin}, \mathrm{target})$ record rather than a paper or a PDB entry. Each record carries the DARPin and target sequences, a binding label, a label source with a readable citation (literature DOI, SYNBIP accession, PDB code, or SPR table), a structure tier, and, where a numeric value can be obtained, K_{D} in nM. Chain maps are frozen separately so that unreliable RCSB entity names are not taken as DARPin or target chains.

The main table has 188 records covering 37 targets, 179 labeled binder and 9 nonbinder, with 121 carrying a numeric K_{D}; label sources fall into four public classes (Figure 2a). The SPR assays come from a public BCL-W screen (Gaudreault et al., 2023), where 11 of 20 show binding and 9 show no detectable binding. Those 9 are the only experimentally confirmed negatives in the table. They are neither synthetic decoys nor a default "affinity not reported" tag. Literature positives include sub-nanomolar DARPins from phage display (Steiner et al., 2008) and the HER2 subdomain co-crystals (Jost et al., 2013), and SYNBIP records keep the SYNBIP accession when no DOI can be matched. Crystal recovery admits only co-crystals whose chain maps pass verification, and known erroneous entries (for example historical 7TXD) are isolated from the complex path.

Structure is tiered rather than a binary "has a PDB entry" flag (Figure 2b): an entry may be a DARPin–target complex, target-only, or a DARPin-only monomer, and the last two are not complex poses. Monomers stay in the structure tree for scaffold reference without counting in the complex tier, and records with no experimental structure still carry sequence, label, and source. Affinity is released as a class as well as a value (Figure 2c); unknown marks measurements that cannot be stably converted to a numeric K_{D}.

Among the 37 targets, HER2 is the only receptor split into two entries by subdomain. Subdomains I and IV of its extracellular region are non-overlapping epitopes, so merging them into one binding column would record positives against different epitopes as a single binding event; the two classes have to be fused bispecifically before they become cytotoxic (Steiner et al., 2008; Jost et al., 2013). We therefore follow that literature split and label records HER2_DI (primary anchor 4HRL) and HER2_DIV (primary anchor 4HRN) rather than introducing a new domain cut. If the literature assigns the same sequence to both domains, each target keeps a record; treating the other side as a negative would turn epitope specificity into a false negative.

The structure tree is organized by target, and Mode 1 additionally releases primary-anchor coordinates for the ten computational units.

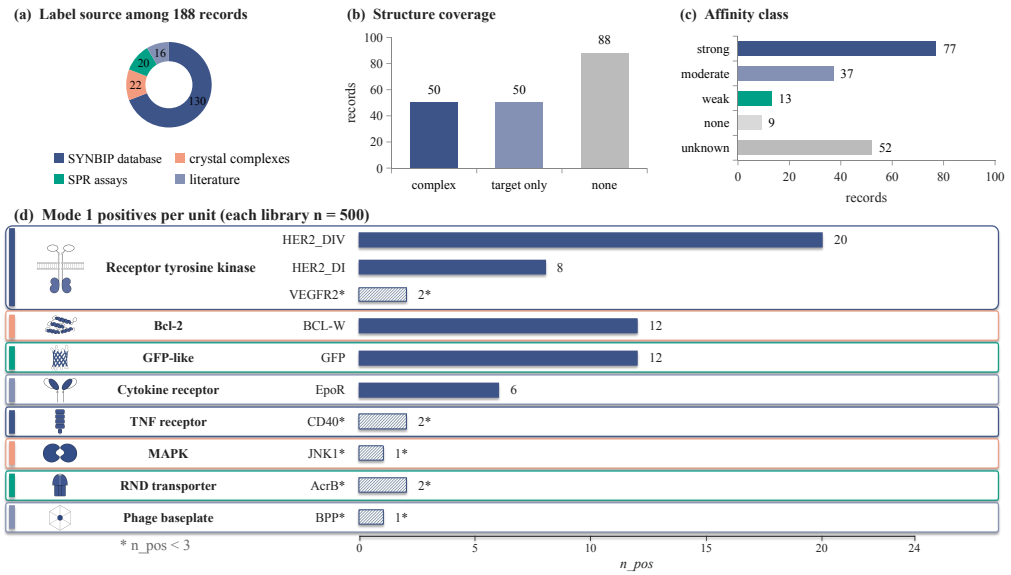

Figure 2: Anatomy of DARPinDB and the Mode 1 units. (a) Label source among 188 records. (b) Structure tiers: complex 50, target-only 50, none 88. (c) Affinity class over all 188 records; 121 carry a numeric K_{D}, and the weak class also holds six qualitative SPR calls. (d) n_{pos} per Mode 1 unit, grouped by target family; units with $n_{\mathrm{pos}} < 3$ are marked.

3.2 DARPinBench

DARPinBench writes screening as sparse-positive ranking: a method assigns each $(\mathrm{DARPin}, \mathrm{target})$ a higher-better scalar and ranks the library, where positives are the sparse experimentally labeled binders and the rest are decoys. Evaluation asks for early enrichment, since display and panning workflows assay the front of a ranked list first. Negatives are sequence neighbors of positives rather than random chains, and we describe those relations by sequence Hamming distance, not as backbone RMSD observations (Kohl et al., 2003; Radom et al., 2019). Screening also does not stop at one neighborhood of one target, so the three modes scale the same task: Mode 1 builds controlled near-neighbor libraries and reports scores, Mode 2 shares one library across targets, and Mode 3 holds targets out.

3.2.1 Mode 1: target admission and near-neighbor libraries

Libraries are deliberately small and hard. A unit is one target together with its own 500-member library, and each unit carries a primary anchor, the experimental DARPin–target co-crystal against which DockQ is measured, so that pose recovery and library ranking can be compared on the same $(\mathrm{DARPin}, \mathrm{target})$ pair. Coverage is nine biological targets (Table S1 in the appendix); a target without such a co-crystal, for example EGFR (ErbB1), does not enter Mode 1. Before freezing, a biology expert reviewed the published assays, the relevance of each target, and whether scaffold and complex pose could be verified. HER2 is split by the literature epitopes of Section 3.1 into HER2_DI and HER2_DIV (Steiner et al., 2008; Jost et al., 2013), so the evaluation protocol has ten units and asks about epitope rather than receptor name.

For Mode 1 positives that still lack their own co-crystal, ESMFold2, Boltz-2, and AF3 each generate 10 candidate complex poses and the same experts select a predicted reference to release. Predicted structures are stored separately from experimental co-crystals and do not replace the crystal primary anchors as DockQ natives.

Each computational unit contains exactly 500 unique sequences. Positives are the binders already labeled for that unit in DARPinDB. Decoys are $k \in \{1, 2, 3\}$ near-neighbor mutations only at consensus-design variable sites, with framework residues held fixed; when positives are few, in-scaffold random sequences fill the library. All ten methods see the same pairs table. Near-neighbor

decoys sit at a median normalized Hamming distance of 0.013 from their unit's nearest positive, about two substitutions on length ~ 155; per-unit means run from 0.006 (JNK1) to 0.18 (HER2_DIV, whose shared pool leaves many k-mutants far from its own positives). That hardness is by design, not leakage: random decoys are much farther (for example mean Hamming about 0.97 on JNK1). Table S1 in the appendix lists the primary anchor, n_{pos}, and decoy mix of every unit; Figure 2d shows n_{pos} grouped by target family. HER2_DIV is densest at $n_{\mathrm{pos}} = 20$, BCL-W and GFP follow at 12, HER2_DI at 8, EpoR at 6; VEGFR2, AcrB and CD40 hold 2 each, and JNK1 and BPP 1. Units with $n_{\mathrm{pos}} < 3$ are marked high-variance, and that mark travels with every later summary. Decoy mix tracks positive density: dense units are almost entirely k-mutants while very sparse ones are filled out with in-scaffold random sequences, and the two HER2 units additionally contain binders to the other domain, which sit at Hamming distance $\gtrsim 0.10$ and implement the epitope split of Section 3.1 rather than duplicate-sequence contamination.

3.2.2 Mode 2 and Mode 3: shared-library panning and cross-target generalization

Mode 2 matches experimental panning of one library against multiple targets: one shared library of 20,000 DARPins serves all ten units, and a sequence counts as a binder only on the unit where it is a Mode 1 positive. Mode 3 reuses those sequences and freezes a leave-one-target-out split over $20\mathrm{k} \times 10 = 200{,}000$ pairs, asking whether a trainable method can rank the positives of a held-out target into the front without treating other targets' binders as hits. Because HER2_DI and HER2_DIV share candidates, holding out one epitope leaves related-epitope transfer rather than a new target. Neither mode is scored here.

3.2.3 Frozen scoring matrix

The ten methods form four families (Table S7 in the appendix). The sequence family comprises the pair score of ESMC-600M (Candido et al., 2026), paired pooling of ProtT5 (Elnaggar et al., 2022), and the interaction probability of D-SCRIPT (Sledzieski et al., 2021); none of the three is a K_{D} regressor. The structure family uses the ipTM of ESMFold2 and Boltz-2 as a structural proxy rather than an affinity (Candido et al., 2026; Passaro et al., 2025), while AF3 contributes its model ranking score (Abramson et al., 2024). Rescoring applies BioScore to both sets of poses (Zhu et al., 2025), and docking uses MEGADOCK FFT scores from the same two fold poses (Matsuzaki et al., 2013).

The matrix was frozen before evaluation: the ten methods, their versions, and their scores were fixed in advance and never tuned against the results reported here. That leaves the scores as a public reference rather than a search for a new fused scorer.

4 Experimental Setup

Mode 1 evaluates the ranking of $n = 500$ sequences within each computational unit, with every score converted so that larger is better (Figure 3a).

The primary metric is Top-k success: a unit succeeds when at least one labeled positive appears in the top $k \in \{5, 25, 50\}$, that is 1%, 5%, and 10% of the library, and the rate is the fraction of the ten units that succeed. The definition still holds at $n_{\mathrm{pos}} = 1$, where it asks whether the unique positive entered the front rather than whether every positive was recalled.

The enrichment factor EF@5% measures positive density in the top 25 relative to random. Half of the units have $n_{\mathrm{pos}} < 3$, and on most units no method places a positive in the top 25, so the ten-unit median is 0 for 8 of 10 methods; mean and median are reported side by side, and per-unit values show where failures concentrate (Figure 3c).

Uncertainty is estimated over computational units rather than library members. Top-k success is a binomial count over the ten units, so every 95% interval reported here is a Wilson score interval for that count (whiskers in Figure 3b); unlike a normal interval it stays inside $[0, 1]$ and remains defined where a rate is 0, which happens twice at Top-1%. The intervals are wide because there are only ten units and five are high-variance, so small gaps between point estimates do not support a unique winner and family comparisons are read through interval overlap. The delete-one-unit

jackknife standard errors, which describe spread between units rather than between library members, are released beside them.

Timing is the wall-clock hours needed to score one 500-pair unit, plotted against target residues (Figure 3d); DARPin length varies only between 154 and 159 residues, so that axis reflects target size. Sequence methods and both BioScore runs are nearly flat in length and collapse into one baseline, the per-unit median of those five.

Pose recovery is evaluated only on the crystal binders that can be aligned to that unit's primary-anchor coordinates: 13 complexes, one Boltz-2 and one AF3 comparison each, 26 valid comparisons. That is an interface evaluation on the crystal subset, not a structural statistic over the 5,000 library members. AF3 ranks by its model ranking score, and its per-structure ipTM is released beside it, so pose quality can be read against ipTM on one scale for both structure models.

5 RESULTS

Across the ten units the 95% intervals span roughly 0.3–0.5, so this comparison resolves whether a family reaches the front of a library at all, not which method ranks first.

5.1 EARLY ENRICHMENT ON NEAR-NEIGHBOR LIBRARIES

No method reaches the top 1% of a library reliably, and the extra compute a structure model spends buys width, not a front-row hit. Figure 3 summarizes the frozen comparison over $10 \times 500 \times 10$. At Top-1%, the top 5 ranks, no method forms a stable advantage: ProtT5 succeeds on 0.30 of the units with interval $[0.11, 0.60]$ and ESMC on 0.20 $[0.06, 0.51]$, while AF3 succeeds on 0.10 $[0.02, 0.40]$ and both BioScore←Boltz-2 and MEGADOCK←Boltz-2 succeed on 0.00 $[0.00, 0.28]$, never placing a positive in the top 5 of these ten units (aggregate rates in Table S8).

Once the window widens to the top 5% and top 10%, the families take different shapes instead of forming one ranking. AF3 reaches Top-10% success 0.80 $[0.49, 0.94]$, the highest wide-window recall among the ten, with Top-1% only 0.10; ESMFold2 reaches Top-5% 0.60 $[0.31, 0.83]$, and D-SCRIPT Top-10% 0.70 with Top-1% also 0.20. Mean EF@5% leads with ESMFold2 (3.95) and AF3 (3.80), while the median is 0 for eight methods, 2.08 for ESMFold2 and 0.83 for AF3.

Per-unit EF@5% (Figure 3c) jumps between 0 and large values for the same method across targets, so the aggregate bars cannot settle which method is the more reliable one. What the comparison supports is weak and complementary early enrichment: sequence methods occasionally hit at the very front, structure methods recover positives only after a larger fraction has been screened, and rescoring and docking do not stably improve the poses they depend on.

5.2 POSE RECOVERY IS NOT LIBRARY RANKING

A high DockQ and a front-of-library rank do not move together. On the crystal-binder subset, Boltz-2 and AF3 each produce one DockQ for 13 complexes ($n = 26$; all values in Table S13 in the appendix), and they trade wins across targets. On JNK1 Boltz-2 reaches 0.810 (high) against AF3 0.431 (acceptable); on BPP the outcome reverses, AF3 0.825 (high) against Boltz-2 0.029 (incorrect); on CD40 Boltz-2 is medium on both chains while both AF3 poses are incorrect (≈ 0.007). GFP inverts the pattern once more, AF3 acceptable at 0.365 where Boltz-2 is incorrect at 0.013.

Two units are reproduced almost exactly by both models: EpoR (0.862 and 0.853) and HER2_DIV (0.919 and 0.911), the latter with ligand RMSD under 1 Å. On AcrB neither model recovers a single native contact ($f_{\mathrm{nat}} = 0$ in all four comparisons), placing its geometry outside either method's range; on VEGFR2 AF3 reaches medium (0.662 and 0.782) where Boltz-2 stays incorrect. On HER2_DI, Boltz-2 is acceptable (0.437) against an incorrect AF3 (0.011).

A high score together with a wrong interface still recurs, and it is sharpest when the score in question is the one that ranks the library. On BPP, Boltz-2 reports an ipTM of 0.894 and recovers almost none of the interface (DockQ 0.029). Read instead through each model's complex confidence, which is not what orders the library, AcrB and AF3 on CD40 join the same pattern. Either way a confident score certifies neither the interface nor the rank of that binder in its library.

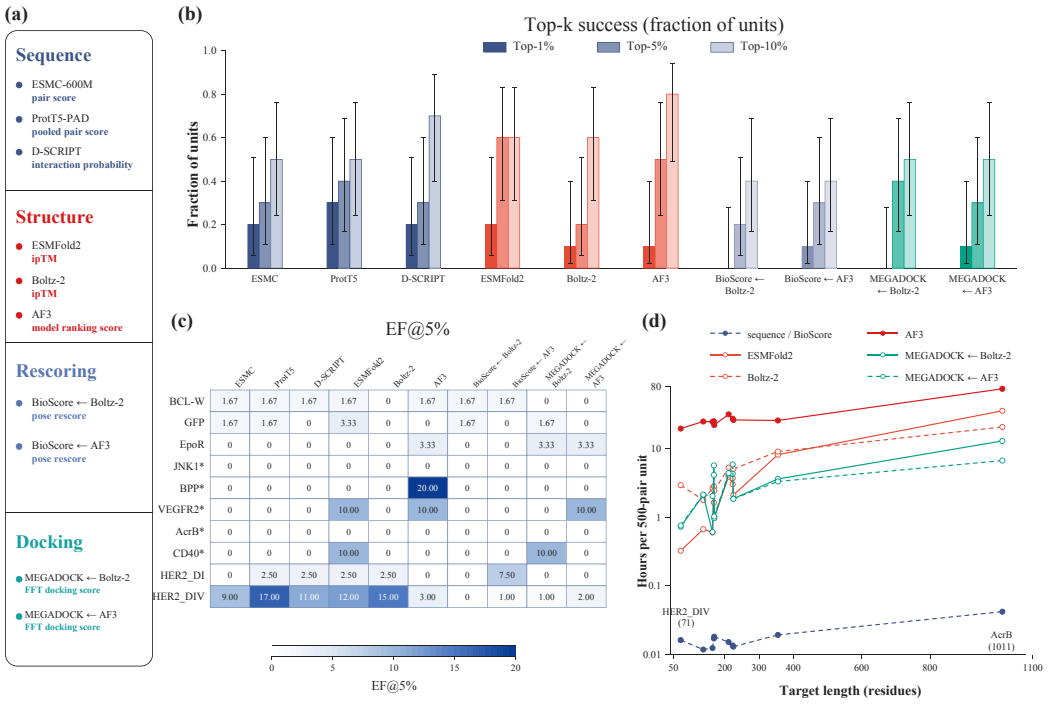

Figure 3: Mode 1 methods, ranking, and cost. (a) Frozen ten-method scoring matrix; the quantity in bold under each method name is the one that orders the library (higher is better throughout). ESMFold2 and Boltz-2 rank by ipTM, AF3 by its model ranking score. (b) Fraction of the ten units with at least one positive in the top 1/5/10%; whiskers are 95% Wilson score intervals. (c) Per-unit EF@5%; * marks units with fewer than three positives. (d) Wall-clock hours per 500-pair unit versus target length (log scale); the baseline line is the per-unit median of the three sequence methods and both BioScore runs.

If pose recovery were equivalent to library ranking, a method that reproduces the crystal interface would place that binder at the front of its 500-member library. EpoR settles it: Boltz-2 and AF3 reproduce its interface at almost the same quality (0.862 and 0.853), yet Boltz-2 leaves that binder at rank 90 while AF3 places it at rank 2, and the other eight methods spread from rank 9 to rank 147 (pose overlays in Figure S4). Interface quality that close cannot account for library ranks two orders of magnitude apart. HER2_DIV shows the converse, high DockQ with rank 1 for six methods, so the two are independent rather than opposed. Mode 1 measures rank in a near-neighbor library, DockQ measures the interface near crystal coordinates, and neither replaces the other.

6 Discussion

On high-variance units a single EF@5% cell swings between 0 and a large value because the denominator holds almost no positives; that is how AF3 reaches 20.00 on BPP, which cannot be read as a lead across targets. AcrB is an outer-membrane multidrug efflux pump with a large interface; its per-unit enrichment is almost uniformly 0 and its crystal pose recovery is incorrect, which points to target geometry outside the range where current fold–dock assumptions hold rather than a chance miss among 500 sequences. The other two units expose different mechanisms. Once HER2 is split into DI and DIV, binders to the other domain act as decoys inside the shared pool, so a model that scores by receptor name is penalized and epitope specificity becomes part of the task (Steiner et al., 2008; Jost et al., 2013). BCL-W carries 12 positives, second only to HER2_DIV, and its early enrichment remains weak: the number of positives does not by itself make a near-neighbor library separable.

Cost and early hits do not track each other once wall-clock time is read against target length. The sequence / BioScore baseline stays near 0.01–0.04 hours per 500 pairs at every length. Fold and docking methods rise with target size: AF3 takes about 19.3 hours on HER2_DIV (71 residues) and about 72.8 hours on AcrB (1011 residues), with a ten-unit median near 24.9 hours. ESMFold2, Boltz-2, and MEGADOCK follow the same trend below AF3. AF3 holds the highest wide-window recall, yet it reaches the top 1% on one unit of ten, a rate below both sequence baselines. Those hours therefore do not buy a reliably earlier first positive, and a more expensive structure model does not belong at the first stage of screening. Rescoring is cheap once poses exist and still does not lift fold models into stable early winners. The scores also fail as probabilities: rescaled to $[0, 1]$ they stay poorly calibrated across the ten units (Table S15), so a high score is not an experimental hit rate that would justify skipping the assay.

Difficulty in Mode 1 therefore does not reduce to a compute shortfall: near-neighbor decoys compress the separable signal into one or two residues, sparse labels make unit-level metrics unstable, and targets such as AcrB and HER2 add interface geometry and epitope assignment on top. A method that improves here will have to gain on residue-level discrimination rather than on global pose quality, which is the distinction the frozen scores make measurable.

7 Limitations

Mode 1 conclusions are bounded by ten computational units, five of which hold fewer than three positives. That sample size, rather than the methods under test, sets the width of every interval reported here. Coverage is bounded by admission: only a target with a positive DARPin–target co-crystal can anchor a unit, so the 37 targets reduce to nine that can anchor a unit, the rest being released as data records outside the Mode 1 aggregates. The HER2 DI/DIV split inherits its epitope assignment from the literature, and if annotation and literature diverge, both DARPinDB and Mode 1 labels are affected.

Structural evidence is tiered rather than uniform: most records carry no complex coordinates, so DARPinDB is a traced joint table rather than a structural set. The pose evidence is bounded in the same way and does not extrapolate to targets without crystals. Supplementary predicted reference complexes are expert selections from sampled candidates; they are predictions and do not serve as DockQ natives. Modes 2 and 3 are held-out protocols released as sequences and labels, so they cannot be read as completed cross-target generalization experiments, and experimental screening and fused ranking pipelines are out of scope. The training corpora of the evaluated methods are not fully public, so exposure to an individual anchor can be bounded only by the release dates in Table S5, not excluded. Near-neighbor k-mutant decoys are a deliberate hardness design: they ask whether nearly identical scaffolds can be separated, not whether a new binding mode can be found in chemical space unrelated to the positives. Library size is likewise a design choice rather than a reproduction of experimental scale, since 500 members keep all ten methods affordable across ten units; Mode 2 is where the library grows.

8 Conclusion

DARPinDB and DARPinBench give DARPin library screening what public sources leave missing: an auditable joint table, and a frozen ranking protocol on top of it.

Off-the-shelf methods do not yet place experimental positives stably at the front of these libraries. More expensive structure models raise recall in a wider window without winning the very front, and the additional wall-clock time does not buy an earlier first positive. On the crystal subset, pose recovery can reach high and can also be fully incorrect at high internal confidence; pose quality is not the same fact as that binder's rank among 500 sequences. The resource leaves data and protocol on reproducible splits. It does not replace experimental screening and does not introduce a new scorer; Mode 1 marks the frozen reference that later evolution and training must exceed, and Modes 2 and 3 reserve splits for that work. Later methods must answer, on the same $(\mathrm{DARPin}, \mathrm{target})$ pair, whether sequence, structure, and binding evidence align, and whether positives can be taken from the front of the library. Modes 2 and 3 are where that question turns from scoring to evolution and training, and we expect the first substantial gains on this benchmark to come from methods that use them.

AI USE STATEMENT

We used generative AI tools for language polishing and for reference sketches that informed the design of some figures. The manuscript, the figures as delivered, and all scientific content were written and produced by the authors, who verified every experimental result, numerical value, and bibliographic entry. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

DARPinDB aggregates sequences, structures, and binding annotations that were already public in the PDB, the literature, curated binder databases, and published SPR tables. No new human-subjects data were collected. Predicted reference complexes are labeled as predictions and are not presented as experimental co-crystals. The benchmark is intended for computational method comparison; it does not provide laboratory protocols for constructing or testing binders. Dual-use risk is that of existing public DARPin and PPI resources: sequences of known binders are already available. We do not release wet-lab guidance.

REPRODUCIBILITY STATEMENT

Mode 1 pairs, labels, and frozen method scores, together with DARPinDB records, structure tiers, and chain maps, are released as supplementary material. Appendix H states what that archive contains. Target admission, HER2 epitope splitting, k-mutant decoy generation, and the ten-method scoring matrix are specified in Section 3. Metrics, the Wilson intervals together with the delete-one-unit jackknife standard errors, and the DockQ crystal subset ($n = 26$) are specified in Section 4. Modes 2 and 3 release sequences and labels without scores, as held-out protocols for later work. Every scoring run was executed as a single job on one NVIDIA A40 GPU, so the wall-clock hours reported in the appendix are per-job times on one device rather than aggregate GPU hours. Implementation details that do not affect the claims of the main text are provided in the appendix.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL http://dx.doi.org/10.1038/s41586-024-07487-w.

H Kaspar Binz, Patrick Amstutz, Andreas Kohl, Michael T Stumpp, Christophe Briand, Patrik Forrer, Markus G Grütter, and Andreas Plückthun. High-affinity binders selected from designed ankyrin repeat protein libraries. *Nature Biotechnology*, 22(5):575–582, Apr 2004. ISSN 1546-1696. doi: 10.1038/nbt962. URL http://dx.doi.org/10.1038/nbt962.

H.Kaspar Binz, Michael T Stumpp, Patrik Forrer, Patrick Amstutz, and Andreas Plückthun. Designing repeat proteins: Well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *Journal of Molecular Biology*, 332(2):489–503, Sept 2003. ISSN 0022-2836. doi: 10.1016/s0022-2836(03)00896-9. URL http://dx.doi.org/10.1016/S0022-2836(03)00896-9.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth

Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. *bioRxiv*, June 2026. doi: 10.64898/2026.06.03.729735. URL http://dx.doi.org/10.64898/2026.06.03.729735.

Lieyang Chen, Anthony Cruz, Steven Ramsey, Callum J. Dickson, Jose S. Duca, Viktor Hornak, David R. Koes, and Tom Kurtzman. Hidden bias in the dud-e dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLOS ONE*, 14(8): e0220113, Aug 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0220113. URL http://dx.doi.org/10.1371/journal.pone.0220113.

Keeley W. Collins, Matthew M. Copeland, Guillaume Brysbaert, Shoshana J. Wodak, Alexandre M.J.J. Bonvin, Petras J. Kundrotas, Ilya A. Vakser, and Marc F. Lensink. Capri-q: The capri resource evaluating the quality of predicted structures of protein complexes. *Journal of Molecular Biology*, 436(17):168540, Sept 2024. ISSN 0022-2836. doi: 10.1016/j.jmb.2024.168540. URL http://dx.doi.org/10.1016/j.jmb.2024.168540.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. Sabdab: the structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, Nov 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1043. URL http://dx.doi.org/10.1093/nar/gkt1043.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, Oct 2022. ISSN 1939-3539. doi: 10.1109/tpami.2021.3095381. URL http://dx.doi.org/10.1109/tpami.2021.3095381.

Francis Gaudreault, Jason Baardsnes, Yuliya Martynova, Aurore Dachon, Hervé Hogues, Christopher R. Corbeil, Enrico O. Purisima, Mélanie Arbour, and Traian Sulea. Exploring rigid-backbone protein docking in biologics discovery: a test using the darpin scaffold. *Frontiers in Molecular Biosciences*, 10, Aug 2023. ISSN 2296-889X. doi: 10.3389/fmolb.2023.1253689. URL http://dx.doi.org/10.3389/fmolb.2023.1253689.

Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, July 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty635. URL http://dx.doi.org/10.1093/bioinformatics/bty635.

Christian Jost, Johannes Schilling, Rastislav Tamaskovic, Martin Schwill, Annemarie Honegger, and Andreas Plückthun. Structural basis for eliciting a cytotoxic effect in her2-overexpressing cancer cells via binding to the extracellular domain of her2. *Structure*, 21(11):1979–1991, Nov 2013. ISSN 0969-2126. doi: 10.1016/j.str.2013.08.020. URL http://dx.doi.org/10.1016/j.str.2013.08.020.

Andreas Kohl, H. Kaspar Binz, Patrik Forrer, Michael T. Stumpp, Andreas Plückthun, and Markus G. Grütter. Designed to be stable: Crystal structure of a consensus ankyrin repeat protein. *Proceedings of the National Academy of Sciences*, 100(4):1700–1705, Feb 2003. ISSN 1091-6490. doi: 10.1073/pnas.0337680100. URL http://dx.doi.org/10.1073/pnas.0337680100.

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xv. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1539–1549, Nov 2023. ISSN 1097-0134. doi: 10.1002/prot.26617. URL http://dx.doi.org/10.1002/prot.26617.

Gyubok Lee, Kiwoong Yoo, Jimin Seo, Jiyoun Kim, Kyunghoon Hur, and Edward Choi. Natural-language-guided generator-agnostic shortlisting for protein binder design. *arXiv*, 2026. doi: 10.48550/arXiv.2608.20755. URL https://arxiv.org/abs/2608.20755.

Yane Li, Chengfeng Wang, Haibo Gu, Hailin Feng, and Yaoping Ruan. Esmdnn-ppi: a new protein–protein interaction prediction model developed with protein language model of esm2 and deep neural network. *Measurement Science and Technology*, 35(12):125701, Sept 2024. ISSN 1361-6501. doi: 10.1088/1361-6501/ad761c. URL http://dx.doi.org/10.1088/1361-6501/ad761c.

Yanlin Li, Fengcheng Li, Zixin Duan, Ruihan Liu, Wenzhe Jiao, Haibo Wu, Feng Zhu, and Weiwei Xue. SYNBIP 2.0: epitopes mapping, sequence expansion and scaffolds discovery for synthetic binding protein innovation. *Nucleic Acids Research*, 53(D1):D595–D603, 2025. doi: 10.1093/nar/gkae893. URL https://doi.org/10.1093/nar/gkae893.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, Mar 2023. ISSN 1095-9203. doi: 10.1126/science.ade2574. URL http://dx.doi.org/10.1126/science.ade2574.

Cong Liu, Milong Ren, Jiaqi Guan, Chengyue Gong, Jinyuan Sun, Xinshi Chen, and Wenzhi Xiao. Protdbench: A unified benchmark of protein binder design and evaluation. *arXiv*, 2026. doi: 10.48550/arXiv.2605.04118. URL https://arxiv.org/abs/2605.04118.

Dan Liu, Francesca Young, Kieran D. Lamb, Adalberto Claudio Quiros, Alexandrina Pancheva, Crispin J. Miller, Craig Macdonald, David L. Robertson, and Ke Yuan. Plm-interact: extending protein language models to predict protein-protein interactions. *Nature Communications*, 16(1), Oct 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-64512-w. URL http://dx.doi.org/10.1038/s41467-025-64512-w.

Huaqing Liu, Peiyi Chen, Xiaochen Zhai, Ku-Geng Huo, Shuxian Zhou, Lanqing Han, and Guoxin Fan. Ppb-affinity: Protein-protein binding affinity dataset for ai-based protein drug discovery. *Scientific Data*, 11(1), Dec 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03997-4. URL http://dx.doi.org/10.1038/s41597-024-03997-4.

Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31 (3):405–412, Oct 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu626. URL http://dx.doi.org/10.1093/bioinformatics/btu626.

Hongfu Lou. Prott-affinity: Sequence-based protein-protein binding affinity prediction using prott5 embeddings. *arXiv*, 2025. doi: 10.48550/arXiv.2511.16113. URL https://arxiv.org/abs/2511.16113.

Yuri Matsuzaki, Nobuyuki Uchikoga, Masahito Ohue, Takehiro Shimoda, Toshiyuki Sato, Takashi Ishida, and Yutaka Akiyama. Megadock 3.0: a high-performance protein-protein interaction prediction software using hybrid parallel computing for petascale supercomputing environments. *Source Code for Biology and Medicine*, 8(1), Sept 2013. ISSN 1751-0473. doi: 10.1186/1751-0473-8-18. URL http://dx.doi.org/10.1186/1751-0473-8-18.

Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, July 2012. ISSN 1520-4804. doi: 10.1021/jm300687e. URL http://dx.doi.org/10.1021/jm300687e.

Divya Nori, Anisha Parsan, Caroline Uhler, and Wengong Jin. Bindenergycraft: Casting protein structure predictors as energy-based models for binder design. *arXiv*, 2025. doi: 10.48550/arXiv.2505.21241. URL https://arxiv.org/abs/2505.21241.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, June 2025. doi: 10.1101/2025.06.14.659707. URL http://dx.doi.org/10.1101/2025.06.14.659707.

Andreas Plückthun. Designed ankyrin repeat proteins (darpins): Binding proteins for research, diagnostics, and therapy. *Annual Review of Pharmacology and Toxicology*, 55(1):489–511, Jan 2015. ISSN 1545-4304. doi: 10.1146/annurev-pharmtox-010611-134654. URL http://dx.doi.org/10.1146/annurev-pharmtox-010611-134654.

Filip Radom, Emanuele Paci, and Andreas Plückthun. Computational modeling of designed ankyrin repeat protein complexes with their targets. *Journal of Molecular Biology*, 431(15):2852–2868, July 2019. ISSN 0022-2836. doi: 10.1016/j.jmb.2019.05.005. URL http://dx.doi.org/10.1016/j.jmb.2019.05.005.

Constantin Schneider, Matthew I J Raybould, and Charlotte M Deane. Sabdab in the age of biotherapeutics: updates including sabdab-nano, the nanobody structure tracker. *Nucleic Acids Research*, 50(D1):D1368–D1372, Nov 2022. ISSN 1362-4962. doi: 10.1093/nar/gkab1050. URL http://dx.doi.org/10.1093/nar/gkab1050.

Palistha Shrestha, Chandana S. Talwar, Jeevan Kandel, Kwang-Hyun Park, Kil To Chong, Eui-Jeon Woo, and Hilal Tayara. Nanobinder: a machine learning assisted nanobody binding prediction tool using rosetta energy scores. *Journal of Cheminformatics*, 17(1), June 2025. ISSN 1758-2946. doi: 10.1186/s13321-025-01040-1. URL http://dx.doi.org/10.1186/s13321-025-01040-1.

Samuel Sledzieski, Rohit Singh, Lenore Cowen, and Bonnie Berger. D-script translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems*, 12(10):969–982.e6, Oct 2021. ISSN 2405-4712. doi: 10.1016/j.cels.2021.08.010. URL http://dx.doi.org/10.1016/j.cels.2021.08.010.

Daniel Steiner, Patrik Forrer, and Andreas Plückthun. Efficient selection of darpins with sub-nanomolar affinities using srp phage display. *Journal of Molecular Biology*, 382(5):1211–1227, Oct 2008. ISSN 0022-2836. doi: 10.1016/j.jmb.2008.07.085. URL http://dx.doi.org/10.1016/j.jmb.2008.07.085.

Charles S. Umbaugh, Marie Groth, Cihan Erkut, Kwang Seok Lee, Joana Marinho, Simon Linder, Florian Iser, Jonas N. Kapp, Petra Schroeter, Simay Dolaner, Asli Kayserili, Dominic Helm, Martin Schneider, Julia Hartmann, Philipp Walch, Thomas F. E. Barth, Kevin Mellert, Birgit Dreier, Jonas V. Schaefer, Andreas Plückthun, Stefan Fröhling, and Claudia Scholl. Selective targeting of TBXT with DARPins identifies regulatory networks and therapeutic vulnerabilities in chordoma. *Science Advances*, 11(36):eadu2796, 2025. doi: 10.1126/sciadv.adu2796. URL https://doi.org/10.1126/sciadv.adu2796.

Svava K. Wetzel, Giovanni Settanni, Manca Kenig, H. Kaspar Binz, and Andreas Plückthun. Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *Journal of Molecular Biology*, 376(1):241–257, Feb 2008. ISSN 0022-2836. doi: 10.1016/j.jmb.2007.11.046. URL http://dx.doi.org/10.1016/j.jmb.2007.11.046.

Dorothea Winkelvoß, David Vukovic, Anna-Carina Hänny, Luca Riermeier, Anto Udovcic, Jonas Kolibius, Annemarie Honegger, Peer R. E. Mittl, Erich Michel, Simon Hansen, and Andreas Plückthun. Molecular features defining the efficiency of bioPROTACs. *Communications Biology*, 8(1):946, 2025. doi: 10.1038/s42003-025-08352-w. URL https://doi.org/10.1038/s42003-025-08352-w.

Ao Xu, Yongchan Hong, Jordy Homing Lam, and Vsevolod Katritch. PLATE-VS: a web server for protein–ligand assay curation and cross-target virtual screening datasets. *Nucleic Acids Research*, 54(W1):W49–W56, 2026. doi: 10.1093/nar/gkag509. URL https://doi.org/10.1093/nar/gkag509.

Sheng Xu, Qiantai Feng, Lifeng Qiao, Hao Wu, Tao Shen, Yu Cheng, Shuangjia Zheng, and Siqi Sun. Benchmarking all-atom biomolecular structure prediction with foldbench. *Nature Communications*, 17(1), Dec 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-67127-3. URL http://dx.doi.org/10.1038/s41467-025-67127-3.

Jing Zhang, Ian R. Humphreys, Jimin Pei, Jinuk Kim, Chulwon Choi, Rongqing Yuan, Jesse Durham, Siqi Liu, Hee-Jung Choi, Minkyung Baek, David Baker, and Qian Cong. Predicting

protein-protein interactions in the human proteome. *Science*, 390(6771), Oct 2025. ISSN 1095-9203. doi: 10.1126/science.adt1630. URL http://dx.doi.org/10.1126/science.adt1630.

Yuchen Zhu, Jihong Chen, Yitong Li, Xiaomin Fang, Xianbin Ye, Jingzhou He, Xujun Zhang, Jingxuan Ge, Chao Shen, Xiaonan Zhang, Tingjun Hou, and Chang-Yu Hsieh. Bioscore: A foundational scoring function for diverse biomolecular complexes. *arXiv*, 2025. doi: 10.48550/arXiv.2507.10877. URL https://arxiv.org/abs/2507.10877.

A How to read this appendix

This appendix carries the material that the main text uses but does not have room to display: the full-precision tables behind every aggregate, the per-unit views that an aggregate hides, and the release contracts that make the numbers checkable. It is organized by question rather than by artifact.

How the resource is built, and what a label means. Appendix B gives the ten Mode 1 units, the record-level field dictionary, the negative-type taxonomy over all $5{,}000$ candidates together with the distance at which those negatives sit, the primary anchors with their release dates, and the Mode 2 / Mode 3 split contracts. Figure 2 in the main text is the picture of that construction; the tables here are its full form.

What the frozen comparison found. Appendix C states what each of the ten methods contributes as a score, reports the aggregate rates with their Wilson intervals, and then resolves them per unit: Figure S1 shows the ranking landscape, and the accompanying tables give the first-positive rank of every unit $\times$ method cell and the between-unit spread behind every interval.

What sits at the front of a library when a method fails. Appendix D resolves the top window by candidate class (Figure S2), which turns the near-neighbor construction from a design claim into an observation and separates three different ways of being wrong.

What the structural control does and does not show. Appendix E holds the crystal subset: pose quality against model confidence (Figure S3), the 26 DockQ comparisons, four case-level pose overlays (Figure S4), and the interface components that DockQ combines.

What the comparison costs, and what a score is worth as a probability. Appendix F reads hit rate against screening depth beside wall-clock time per unit (Figure S5), with the calibration error and the timing grids in full.

What later work inherits. Appendix G states the measurement standard a later method has to meet on this benchmark, and Appendix H states what the supplementary archive contains and the contracts on those files.

B Data records and library construction

B.1 Mode 1 units

A unit is one target together with its own 500-member library and its primary anchor, the experimental co-crystal against which DockQ is measured. Positive density varies by an order of magnitude across the ten units, and the decoy mix follows it: dense units are almost entirely k-mutants, while units with one or two positives are filled out with in-scaffold random sequences because there are too few positives to mutate around. Both properties travel with every later summary, which is why the high-variance mark is repeated in each table rather than stated once.

Table S1: Mode 1 units. Each library has $n = 500$. HV marks $n_{\text{pos}} < 3$. HER2_DI / HER2_DIV have n_{pos} of 8 and 20. The nine BCL-W entries assayed without detectable binding are library members labeled negative, not synthetic decoys.

Unit	PDB	n_{pos}	HV	Decoy mix
BCL-W	4K5A	12		9 assayed nonbinder / 479 k-mut
GFP	9F24	12		488 k-mut
EpoR	6MOE	6		360 k-mut / 134 rand
JNK1	5LW1	1	*	60 k-mut / 439 rand
BPP	3HG0	1	*	60 k-mut / 439 rand
VEGFR2	5OYJ	2	*	120 k-mut / 378 rand
AcrB	3NOG	2	*	120 k-mut / 378 rand
CD40	7P3I	2	*	120 k-mut / 378 rand
HER2_DI	4HRL	8		473 k-mut (+ other-domain binders)
HER2_DIV	4HRN	20		473 k-mut (+ other-domain binders)

B.2 Record fields and annotation rules

One row per released field. Conditional requirements are stated with the field they apply to. In the structure manifest, `model`, `dockq_row_key` and `prediction_rank` are required only for the predicted roles `boltz2_top1` and `af3_top1`. In a per-case chain map, the pre-clean `pdb_auth_*` chains are required only on the native row. Finally, `prerequisite_pose_source` is required only for the four methods that consume an external pose: both BioScore and both MEGADOCK runs.

Table S2: DARPinDB data dictionary and annotation rules. One row per released field. *Rule* states the value domain where one is controlled and the missing-value convention otherwise; an empty cell in the release means absent, and the strings `NA`, `None` and – are never used as sentinels. Conditional requirements are stated with the field they apply to.

Field	Type	Rule and description
`record_id`	string	required for dataset rows. Stable identity key for a DARPin–target record. Applies to `identity;dataset`.
`darpin_seq`	string	Values: amino-acid alphabet without X. required; X letters deleted not replaced. DARPin amino-acid sequence after X-stripping. Applies to `identity;dataset;mode1`.
`scaffold_id`	string	empty if unknown. Internal scaffold identifier. Applies to `identity;dataset`.
`sbp_id`	string	Values: SBP###### or empty. empty if not from the SYNBIP database. SYNBIP accession when applicable. Applies to `identity;dataset`.
`shortuid`	string	required in Mode1 pairs. Mode1 library member identifier used in scores and DockQ. Applies to `identity;mode1`.
`target_folder`	string	Values: registry folder names; HER2_DI or HER2_DIV not bare HER2. required for binders. Canonical target folder in DARPinDB / Mode1. Applies to `identity;dataset;mode1`.
`target_name`	string	empty if unknown. Human-readable target name. Applies to `identity;dataset`.
`target_uniprot`	string	empty if not applicable. UniProt accession for the biological target. Applies to `identity;dataset`.
`target_seq`	string	Values: amino-acid alphabet without X. required in Mode1 pairs. Target amino-acid sequence used for scoring. Applies to `identity;mode1`.
`label`	string	Values: binder—nonbinder. required. Categorical binding label. Applies to `binding;dataset`.
`label_is_binder`	integer	Values: 0—1. required. Binary binding label used in Mode1 metrics. Applies to `binding;dataset;mode1`.
`label_source`	string	Values: literature_or_sbp_positive—synthetic_kmut—synthetic_random—.... required in Mode1. Provenance of the Mode1 library member label. Applies to `binding;curation;mode1`.

Field	Type	Rule and description
confidence_tier	string	empty if unknown. Curation confidence for the dataset label. Applies to binding;curation;dataset.
evidence_source	string	empty if unknown. Underlying evidence stream (SBP, PDB, literature, SPR). Applies to binding;curation;dataset.
epitope_role	string	Values: includes cross_epitope_hard_neg. empty outside HER2 split. Mode1 epitope-split role; hard negatives are other-domain binders. Applies to binding;mode1.
binder_class	string	empty if not annotated. Fine binder class after HER2 epitope split. Applies to binding;mode1.
in_bench9	string	Values: yes—no. required on dataset positives. Biological nine-target set; HER2 remains one biology seat. Applies to binding;dataset.
affinity_raw	string	empty if unreported. Unparsed affinity string from source. Applies to affinity;dataset.
kd_nM	float	Values: nM. empty if unknown. Parsed equilibrium dissociation constant. Applies to affinity;dataset.
kd_nM_low	float	Values: nM. empty if no interval. Lower parsed Kd bound. Applies to affinity;dataset.
kd_nM_high	float	Values: nM. empty if no interval. Upper parsed Kd bound. Applies to affinity;dataset.
affinity_class	string	Values: strong—moderate—weak—none—unknown. unknown if unparsed. Binned affinity class; not a model score. Applies to affinity;dataset.
affinity_assay	string	empty if unknown. Experimental assay used for affinity. Applies to affinity;dataset.
affinity_parse_note	string	empty if clean parse. Curation note on affinity parsing. Applies to affinity;curation;dataset.
has_gt_complex	string	Values: yes—no. required. Whether an experimental DARPin–target complex exists. Applies to structure;dataset.
gt_pdb_id	string	Values: 4-char PDB or empty. empty if no complex. Experimental complex PDB code. Applies to structure;dataset.
structure_tier	string	Values: complex—target_only—darpin_only—none. required. Highest available experimental structure tier. Applies to structure;dataset.
structure_qc_status	string	Values: ok—quarantine—.... required when a path is present. QC status of linked experimental coordinates. Applies to structure;curation;dataset.
path_complex	string	Values: repo-relative path. empty if none; never an absolute server path in SI tables. Dataset path to experimental complex. Applies to structure;dataset.
gt_darpin_chain	string	Values: PDB auth chain. empty if unknown. Experimental DARPin chain ID before pdb_clean. Applies to structure;chain mapping;dataset.
gt_target_chain	string	Values: PDB auth chain. empty if unknown. Experimental target chain ID before pdb_clean. Applies to structure;chain mapping;dataset.
darpin_chains	string	Values: PDB auth chain(s). required for dataset_ok complexes. Auth DARPin chain(s) in chain map. Applies to chain mapping;structures.
target_chains	string	Values: PDB auth chain(s). required for dataset_ok complexes. Auth target chain(s) in chain map. Applies to chain mapping;structures.
pdb_clean_darpin_chain	string	Values: A. A after clean. After pdb_clean, DARPin is chain A. Applies to chain mapping;mode1;structures.
pdb_clean_target_chain	string	Values: B. B after clean. After pdb_clean, target is chain B. Applies to chain mapping;mode1;structures.
his_tag_policy	string	documented per construct. Whether His-tag residues were kept; HER2_DI 9_29 DockQ allows mismatches. Applies to curation;structure;mode1.
source_type	string	Values: literature—sbp—pdb—spr. required on dataset positives. External source class. sbp is not an external reference. Applies to curation;dataset.

Field	Type	Rule and description
`reference`	string	empty if none. Human-readable citation. Do not invent DOIs. Applies to `curation;dataset`.
`doi`	string	Values: bare DOI or empty. empty if none; never fabricated. Bare DOI when known. Applies to `curation;dataset`.
`paper`	string	empty if none. Internal paper/source key. Applies to `curation;dataset`.

B.3 What a negative is, and how near it sits

Mode 1 negatives are not one kind of object, and the distinction decides what a failure means. A method that ranks a k-mutant decoy first has failed to separate near neighbors; a method that ranks a cross-epitope binder first has failed at epitope specificity, which is a different deficiency; and a method that ranks an assayed non-binder first has been shown a genuine experimental negative and put it at the front. Only the last case is available in this release, and only in BCL-W: its nine SPR entries without detectable binding sit in the library as labeled negatives, and 176 of its k-mutants were generated around three of those nine, so that fraction of the decoy set is a near neighbor of an assayed non-binder rather than of a positive.

How near those neighbors sit is what makes the task hard, and it is measured rather than asserted. A k-mutant sits at a median normalized Hamming distance of 0.013 from its unit's nearest positive, about two substitutions on length ~ 155, while random fill sequences sit two orders of magnitude away. The per-unit means separate two regimes that the median hides: units whose decoys were generated around their own positives stay near 0.006, whereas HER2_DIV reaches 0.18 because its shared pool leaves many k-mutants far from its own positives. BCL-W's mean of 0.042 is the largest outside the two HER2 units for the reason just given, its decoys were generated around assayed non-binders.

Table S3: Negative-type taxonomy over the $10 \times 500 = 5{,}000$ Mode 1 candidates. A label of 0 does not by itself mean that a sequence was assayed and found not to bind: only the `experimental_nonbinder` rows carry that evidence, and they occur in BCL-W alone. Cross-epitope rows are genuine binders to the other HER2 subdomain and are negatives only relative to the unit they appear in.

Type	n	Meaning	Evidence basis
`positive`	66	Experimentally labeled binder for that unit in DARPinDB.	Literature DOI, SYNBIP accession, PDB co-crystal, or SPR table.
`experimental_nonbinder`	9	Assayed and reported without detectable binding. Present in BCL-W only.	Public BCL-W SPR screen (Gaudreault et al., 2023).
`cross_epitope_binder`	26	A binder to the other HER2 subdomain, acting as a task negative in this unit.	Literature epitope assignment (Steiner et al., 2008; Jost et al., 2013).
`synthetic_kmut_decoy`	2,753	Generated $k \in \{1, 2, 3\}$ mutant at consensus-design variable sites.	Generated; framework residues held fixed.
`synthetic_random_decoy`	2,146	In-scaffold random sequence used to fill sparse units.	Generated.
`unverified_task_negative`	0	Carries $\mathrm{label} = 0$ with none of the above applying.	None; the category is empty in this release.
Total	5,000	Ten units of 500 unique sequences each.	

Table S4: Normalized Hamming distance from each k-mutant decoy to the nearest positive of its own unit, recomputed after the BCL-W label correction. Distance is mismatches plus length difference over the longer sequence. The two HER2 units are far above the rest because they share one candidate pool, so many k-mutants are neighbors of the other epitope's positives rather than of their own. BCL-W is elevated for a different reason: 176 of its k-mutants were generated around three assayed non-binders, which are no longer positives.

Unit	n	mean	median	min	max
BCL-W	479	0.0417	0.0194	0.0065	0.1032
GFP	488	0.0129	0.0126	0.0062	0.0238
EpoR	360	0.0102	0.0083	0.0022	0.0191
JNK1	60	0.0063	0.0063	0.0032	0.0095
BPP	60	0.0163	0.0163	0.0081	0.0244
VEGFR2	120	0.0127	0.0127	0.0063	0.0191
AcrB	120	0.0131	0.0131	0.0065	0.0199
CD40	120	0.0127	0.0127	0.0063	0.0192
HER2_DI	473	0.0426	0.0180	0.0060	0.1299
HER2_DIV	473	0.1813	0.1198	0.0060	0.6013

B.4 Primary anchors and exposure bounds

Every unit is anchored to a deposited co-crystal, so the deposition and release dates bound what an evaluated method could have seen. The bound is one-sided: the training corpora of the evaluated methods are not fully public, so an early release date does not establish exposure and a late one does not exclude it.

Table S5: Mode 1 primary anchors with PDB deposition and release dates, retrieved from RCSB. The training corpora of the evaluated methods are not fully public, so these dates are given for the reader to judge exposure per method rather than as a claim that any anchor was unseen.

Unit	PDB	Deposited	Released
BPP	3HG0	2009-05-13	2009-09-08
AcrB	3NOG	2010-06-25	2011-05-18
HER2_DI	4HRL	2012-10-28	2013-10-16
HER2_DIV	4HRN	2012-10-28	2013-10-16
BCL-W	4K5A	2013-04-14	2014-04-16
JNK1	5LW1	2016-09-15	2017-12-13
VEGFR2	5OYJ	2017-09-10	2018-03-14
EpoR	6MOE	2018-10-04	2019-06-05
CD40	7P3I	2021-07-07	2022-04-06
GFP	9F24	2024-04-22	2024-10-23

B.5 Mode 2 and Mode 3 split contracts

Modes 2 and 3 are released as material rather than as results, so their contracts have to be explicit about what counts as a hit. One shared library of $20{,}000$ DARPins serves all ten units in Mode 2, and a sequence counts as a binder only on the unit where it is a Mode 1 positive; the same sequence is a negative everywhere else, which is what panning against multiple targets means experimentally. Mode 3 freezes a leave-one-target-out split over the resulting $200{,}000$ pairs. Because HER2_DI and HER2_DIV share candidates, holding out one epitope leaves related-epitope transfer rather than an unseen target, and a submission has to report it as such. Neither mode carries model scores in this work.

Table S6: Mode 2 and Mode 3 split contracts. Both modes are released as held-out material: sequences and labels only, with no model scores produced in this work. Counts are the frozen release sizes.

Field	Mode 2	Mode 3
Task	shared_library_binder_identification	leave_one_target_out
Sample unit	sequence	pair
Sequences	20,000	20,000
Pairs	200,000	200,000
Split definition	One shared 20k DARPin library reused across 10 units; per-unit pairs_v1.tsv is the same 20k sequences paired to that unit target. label_is_binder=1 only for that unit Mode1 positives.	Cartesian 20k x 10 units = 200000 pairs. 10 LOTO folds: train = pairs with unit != holdout (180000), test = holdout unit (20000). HER2_DI/HER2_DIV are related-epitope transfer: the other epitope remains in train.
Labels released	true	true
Scored in this paper	false	false
Intended use	External scoring and retrieval on a held-out 20k library. This paper does not train or score Mode 2.	External LOTO training/evaluation. This paper does not train Mode 3 or report LOTO scores.
Exclusions	No model scores; no structure/docking outputs; old 10k shared_library_v1 is not the freeze.	No model scores; no LOTO metrics in this paper.

C THE FROZEN COMPARISON IN FULL

C.1 WHAT IS BEING COMPARED

The ten methods are not ten estimates of one quantity. Each contributes the score its own family produces, under its own definition and direction, and the matrix was fixed before evaluation so that none of the scores was tuned against the results reported here. Reading the table below beside any later row of numbers is what keeps a fold confidence from being read as an affinity.

Table S7: Frozen scoring matrix. Every method maps $(\mathrm{DARPin}, \mathrm{target})$ to a higher-better score, and the quantity named here is the one that actually orders the library: it was verified column by column against the released per-model outputs. ESMFold2 and Boltz-2 both contribute ipTM, a structural proxy rather than an affinity or K_{D} predictor; AF3 contributes its model ranking score, which is neither ipTM nor a calibrated probability.

Method	Family	Score that determines the rank
ESMC-600M	Sequence	pair score; not K_{D}
ProtT5-PAD	Sequence	paired pooling; not K_{D}
D-SCRIPT	Sequence	interaction probability; not K_{D}
ESMFold2	Structure	ipTM; not an affinity
Boltz-2	Structure	ipTM; not an affinity
AF3	Structure	model ranking score; not ipTM
BioScore←Boltz-2	Rescoring	pose rescore
BioScore←AF3	Rescoring	pose rescore
MEGADOCK←Boltz-2	Docking	FFT docking score
MEGADOCK←AF3	Docking	FFT docking score

C.2 AGGREGATE RATES

The aggregate is a count over ten units, and it should be read as one: a Top-1% rate of 0.20 is two units, not a fifth of a library. Mean and median EF@5% are reported side by side because they disagree by construction, the median being zero for eight of the ten methods.

Table S8: Aggregate Mode 1 metrics. Top-k is the fraction of the ten units with at least one positive in the top k. Brackets are 95% Wilson score intervals for that binomial count: unlike a normal or jackknife interval they stay inside $[0, 1]$ and remain defined when a rate is 0, which happens twice at Top-1%. Median EF@5% is 0 for eight of the ten methods, so the mean is reported beside wall-clock hours; hours are the median per 500-pair unit. Per-unit values are in Table S9 and Figure 3c.

Method	Family	Top-1%	Top-5%	Top-10%	mean EF@5% / h
ESMC	Sequence	0.20 [0.06, 0.51]	0.30 [0.11, 0.60]	0.50 [0.24, 0.76]	1.23 / 0.015
ProtT5	Sequence	0.30 [0.11, 0.60]	0.40 [0.17, 0.69]	0.50 [0.24, 0.76]	2.28 / 0.017
D-SCRIPT	Sequence	0.20 [0.06, 0.51]	0.30 [0.11, 0.60]	0.70 [0.40, 0.89]	1.52 / 0.030
ESMFold2	Structure	0.20 [0.06, 0.51]	0.60 [0.31, 0.83]	0.60 [0.31, 0.83]	3.95 / 1.86
Boltz-2	Structure	0.10 [0.02, 0.40]	0.20 [0.06, 0.51]	0.60 [0.31, 0.83]	1.75 / 3.29
AF3	Structure	0.10 [0.02, 0.40]	0.50 [0.24, 0.76]	0.80 [0.49, 0.94]	3.80 / 24.9
BioScore←Boltz-2	Rescoring	0.00 [0.00, 0.28]	0.20 [0.06, 0.51]	0.40 [0.17, 0.69]	0.33 / 0.015
BioScore←AF3	Rescoring	0.10 [0.02, 0.40]	0.30 [0.11, 0.60]	0.40 [0.17, 0.69]	1.02 / 0.015
MEGADOCK←Boltz-2	Docking	0.00 [0.00, 0.28]	0.40 [0.17, 0.69]	0.50 [0.24, 0.76]	1.60 / 2.83
MEGADOCK←AF3	Docking	0.10 [0.02, 0.40]	0.30 [0.11, 0.60]	0.50 [0.24, 0.76]	1.53 / 2.70

C.3 THE RANKING LANDSCAPE BEHIND THE AGGREGATE

Two views resolve what the aggregate rates leave out. Recall against screening depth shows the shape of a method's enrichment across the whole library rather than at three cut-offs, and the first-positive rank of every unit × method cell shows where the successes and failures actually sit. Read together they show that no family owns a column and no unit is easy for every method: the cells that succeed at the very front are scattered, and units such as AcrB stay dark for all ten methods.

Table S9: Rank of the first labeled positive in each Mode 1 unit, for every method. Smaller is better; 5 is the top 1% and 25 the top 5% of a 500-member library. * marks units with fewer than three positives. EpoR is the case that separates pose recovery from ranking: both structure models reproduce its crystal interface at high DockQ, yet Boltz-2 leaves that binder at rank 90 while AF3 places it at rank 2, and the remaining eight methods fall between rank 9 and rank 147.

Unit	n_{pos}	ESMC	ProtT5	D-SCRIPT	ESMFold2	Boltz-2	AF3	BioScore ←Boltz-2	BioScore ←AF3	MEGADOCK ←Boltz-2	MEGADOCK ←AF3
BCL-W	12	7	1	2	11	53	14	10	18	43	34
GFP	12	1	1	62	5	30	26	9	34	22	57
EpoR	6	86	74	39	147	90	2	95	142	9	22
JNK1*	1	368	42	42	153	280	46	52	239	152	228
BPP*	1	37	476	450	400	28	19	400	455	322	410
VEGFR2*	2	97	78	28	24	62	14	90	234	59	17
AcrB*	2	378	217	126	370	50	125	290	194	362	194
CD40*	2	453	459	49	22	31	307	100	357	13	91
HER2_DI	8	33	14	7	14	17	32	37	5	104	40
HER2_DIV	20	1	1	1	1	1	16	35	10	12	1

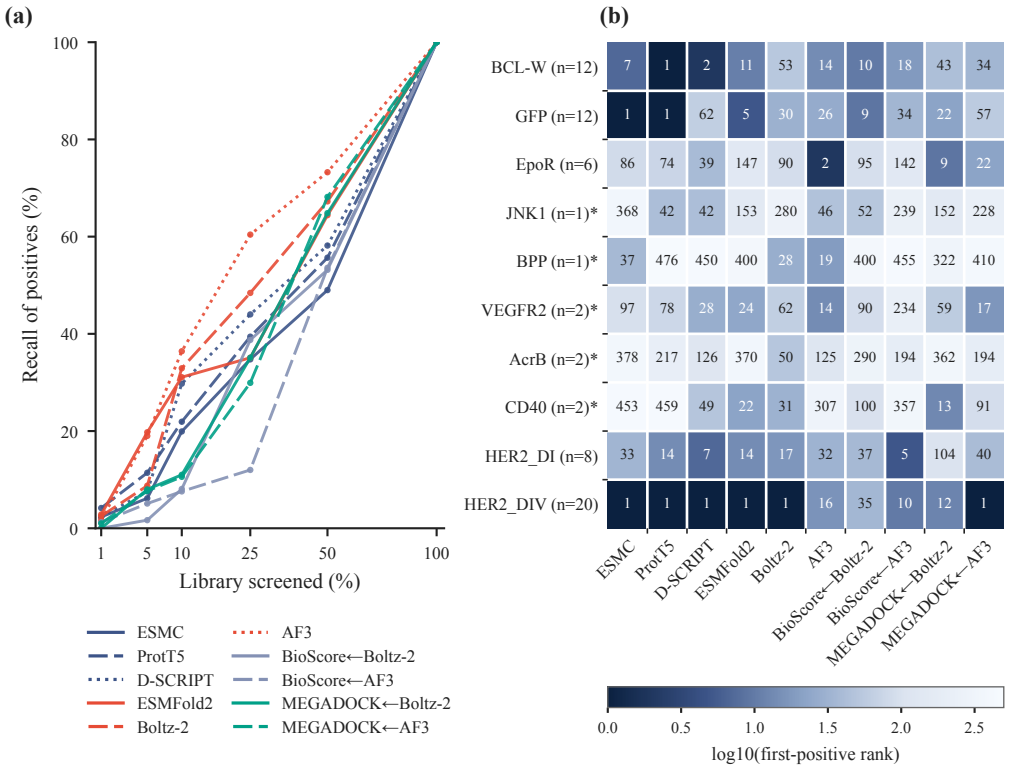

Figure S1: Ranking landscape over the ten Mode 1 units. (a) Recall of the full positive set as a function of the fraction of the 500-member library screened, aggregated over the ten units (macro mean) for all ten methods. Lines are colored by method family and distinguished within a family by line style; the legend lists every method by name. The x-axis uses a compressed, front-expanded scale. (b) First-positive rank for each unit $\times$ method cell: the rank (1 is best) of the first true positive within the 500 candidates. Colour encodes $\log_{10}$ of that rank, so darker is a smaller and better rank. Row labels give the number of positives per unit, and units with fewer than three positives ($n_{\mathrm{pos}} < 3$) are marked with an asterisk. BCL-W carries twelve positives here, after nine SPR entries assayed without detectable binding were moved out of its positive set and into the library as labeled negatives.

Table S10: Spread across the ten computational units, reported as delete-one-unit jackknife standard errors. These describe variation between units rather than between library members, and they are not independent-sample standard errors. The intervals in the main text are Wilson score intervals rather than 1.96 of these standard errors: for a binomial count over ten units the normal form collapses to zero width wherever a rate is 0 and can exceed $[0, 1]$ elsewhere. The last two columns give the normal form at Top-1% and Top-5% for comparison; the zero-width entries are exactly the cases that motivated the change.

Method	jackknife SE				normal-form interval	
	Top-1%	Top-5%	Top-10%	mean EF	Top-1%	Top-5%
ESMC	0.1333	0.1528	0.1667	0.8903	[0.00, 0.46]	[0.00, 0.60]
ProtT5	0.1528	0.1633	0.1667	1.6622	[0.00, 0.60]	[0.08, 0.72]
D-SCRIPT	0.1333	0.1528	0.1528	1.0905	[0.00, 0.46]	[0.00, 0.60]
ESMFold2	0.1333	0.1633	0.1633	1.5207	[0.00, 0.46]	[0.28, 0.92]
Boltz-2	0.1000	0.1333	0.1633	1.4930	[0.00, 0.30]	[0.00, 0.46]
AF3	0.1000	0.1667	0.1333	2.0507	[0.00, 0.30]	[0.17, 0.83]
BioScore←Boltz-2	0.0000	0.1333	0.1633	0.2222	[0.00, 0.00]	[0.00, 0.46]
BioScore←AF3	0.1000	0.1528	0.1633	0.7431	[0.00, 0.30]	[0.00, 0.60]
MEGADOCK←Boltz-2	0.0000	0.1633	0.1667	0.9960	[0.00, 0.00]	[0.08, 0.72]
MEGADOCK←AF3	0.1000	0.1528	0.1667	1.0089	[0.00, 0.30]	[0.00, 0.60]

D WHAT OCCUPIES THE FRONT OF A LIBRARY

Aggregate rates say how often a method fails; they do not say what it ranks instead. Resolving the top of each library by candidate class turns the near-neighbor construction from a design claim into an observation: positives occupy between 0.8 and 8 per cent of the top 5%, and k-mutants take roughly half to three quarters of it for every one of the ten methods. The same resolution separates three failures that an aggregate merges. Ranking a k-mutant first means near neighbors were not separated; ranking a cross-epitope binder first means epitope specificity was missed; ranking an assayed non-binder first means a genuine experimental negative was placed at the front. Only the last is available here, and only in BCL-W.

Which sequences reach the front is also not method-specific. A small number of task negatives are placed in the top window by many methods at once, which is what a shared blind spot looks like at candidate level and what a later method can be tested against directly.

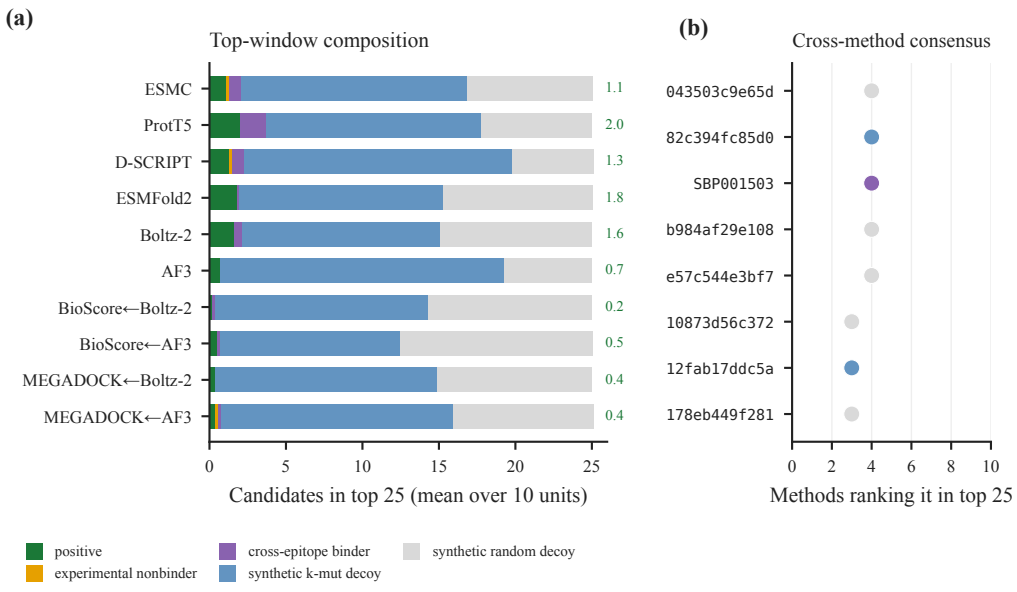

Figure S2: What occupies the front of the library. (a) Composition of the top window (top 25, that is 5% of 500), one row per method, as the mean over the ten units of the number of candidates of each member class. Segment order is fixed: positive, experimental nonbinder, cross-epitope binder, synthetic k-mutant decoy, synthetic random decoy. The number at the right of each bar is the mean positive count per unit, a decimal because it is averaged over units; per unit it is an integer, typically 0 or 1. The `experimental_nonbinder` class occurs only in BCL-W, in three top-window cells across all methods. Cross-epitope binders are genuine binders to the other HER2 subdomain that act as task negatives within their unit; they are neither non-binders, nor false positives, nor decoys. (b) Cross-method consensus for the highest-ranked task negatives: for each of the eight sequences most frequently placed in the top window, the number of methods ranking it within the top 25. Point color matches the member class in (a).

Table S11: What occupies the front of the library. Mean fraction of the top 5% (top 25 of 500) by candidate class, averaged over the ten units. Positives are a small minority for every method. A label of 0 does not by itself mean a sequence was assayed and found not to bind: only the *exp. nonbinder* column carries that evidence, and it occurs in BCL-W alone.

Method	positive	exp. nonbinder	cross-epitope	k-mutant	random
ESMC	0.044	0.004	0.032	0.592	0.328
ProtT5	0.080	0.000	0.068	0.564	0.288
D-SCRIPT	0.052	0.004	0.032	0.700	0.212
ESMFold2	0.072	0.000	0.004	0.532	0.392
Boltz-2	0.064	0.000	0.020	0.520	0.396
AF3	0.028	0.000	0.000	0.744	0.228
BioScore←Boltz-2	0.008	0.000	0.008	0.556	0.428
BioScore←AF3	0.020	0.000	0.004	0.472	0.504
MEGADOCK←Boltz-2	0.016	0.000	0.000	0.580	0.404
MEGADOCK←AF3	0.016	0.004	0.004	0.608	0.368

Rows sum to 1 up to rounding.

Table S12: Task negatives that reach the front of their library, restricted to the two classes that carry biological meaning. An assayed non-binder at the front means the method was shown a genuine experimental negative and ranked it first; a cross-epitope binder at the front means the method scored by receptor rather than by epitope.

Unit	Method	Rank	Negative type	Hamming
BCL-W	ESMC	1	experimental nonbinder	0.103226
BCL-W	D-SCRIPT	15	experimental nonbinder	0.083871
BCL-W	MEGADOCK←AF3	23	experimental nonbinder	0.083871
HER2_DI	ESMC	1	cross epitope binder	0.448052
HER2_DI	ESMC	2	cross epitope binder	0.441558
HER2_DI	ProtT5	1	cross epitope binder	0.441558
HER2_DI	ProtT5	2	cross epitope binder	0.448052
HER2_DI	ProtT5	3	cross epitope binder	0.415584
HER2_DI	D-SCRIPT	1	cross epitope binder	0.428571

E POSE RECOVERY AGAINST LIBRARY RANKING

E.1 DOCKQ ON THE CRYSTAL-BINDER SUBSET

Pose recovery is evaluated only where a crystal binder can be aligned to its unit's primary anchor: 13 complexes, one Boltz-2 and one AF3 comparison each, 26 valid comparisons. That is an interface evaluation on a subset, not a structural statistic over the $5{,}000$ library members, and it is not extrapolated to units without crystals.

Plotting DockQ against per-structure ipTM shows why a confidence-like score cannot stand in for interface quality. Both panels use ipTM on one shared scale, so the two structure models are directly comparable; within each panel, a high ipTM occurs together with an incorrect interface often enough that the ordering by ipTM is not the ordering by DockQ. Neither ipTM is a calibrated probability, and AF3 ranks the library by its model ranking score rather than by ipTM.

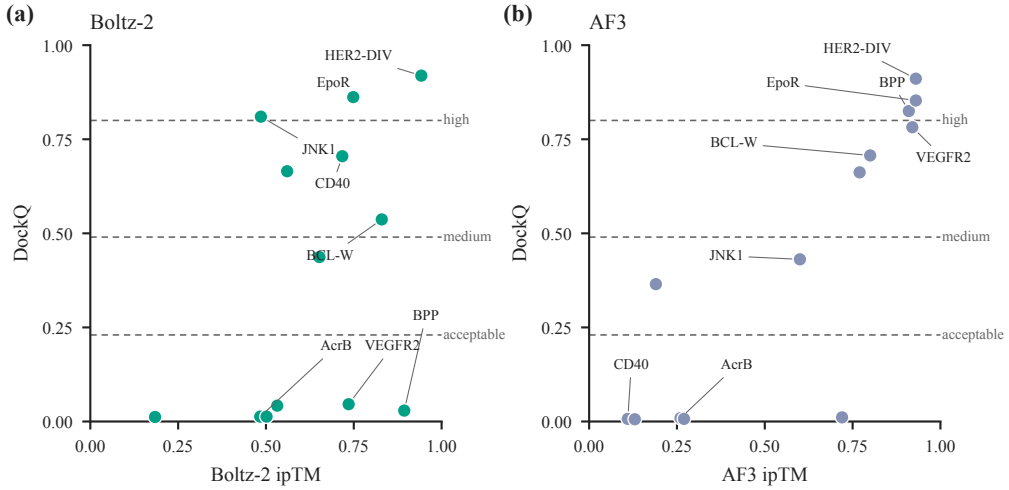

Figure S3: Pose quality against per-structure ipTM. DockQ of the top-1 predicted pose against that structure's ipTM on the crystal subset ($n = 26$: 13 native complexes $\times$ 2 models). (a) Boltz-2 and (b) AF3, both with x the ipTM of the scored structure, so the two panels share one axis and the models are directly comparable. Neither ipTM is a calibrated probability, and AF3 ranks the library by its model ranking score rather than by ipTM. Dashed lines mark the CAPRI DockQ thresholds 0.23 / 0.49 / 0.80 (acceptable / medium / high). The 13 complexes come from ten units; VEGFR2, AcrB and CD40 each contribute two crystal structures. EpoR, at 0.862 and 0.853, and HER2_DIV, at 0.919 and 0.911, are labeled. This subset is not extrapolated to the full library.

Table S13: DockQ on the crystal-binder subset: 13 complexes with one Boltz-2 and one AF3 comparison each ($n = 26$). CAPRI grades in parentheses. The HER2_DI row compares a literature positive against primary anchor 4HRL, whose construct differs in length from the Mode 1 sequence. Native chain pairings were verified computationally and by a structural-biology expert; the AcrB, EpoR, HER2_DIV and VEGFR2 natives were assembled directly from the deposited entries. This table reads the two models side by side; the interface components behind each DockQ value, together with each model's own confidence, are in Table S14.

Unit	Structure	Boltz-2	AF3
BCL-W	4K5A chain B	0.537 (medium)	0.707 (medium)
GFP	9F24 chain C	0.013 (incorrect)	0.365 (acceptable)
EpoR	6MOE chain A	0.862 (high)	0.853 (high)
JNK1	5LW1 chain A	0.810 (high)	0.431 (acceptable)
BPP	3HG0 chain D	0.029 (incorrect)	0.825 (high)
VEGFR2	5OYJ chain A	0.046 (incorrect)	0.662 (medium)
VEGFR2	5OYJ domain 4	0.042 (incorrect)	0.782 (medium)
AcrB	3NOG chain D	0.013 (incorrect)	0.009 (incorrect)
AcrB	3NOG chain E	0.012 (incorrect)	0.007 (incorrect)
CD40	7P3I chain B	0.665 (medium)	0.007 (incorrect)
CD40	7P3I chain D	0.705 (medium)	0.006 (incorrect)
HER2_DI	vs anchor 4HRL	0.437 (acceptable)	0.011 (incorrect)
HER2_DIV	4HRN chain A	0.919 (high)	0.911 (high)

E.2 CASE-LEVEL POSE CONTROLS

Four cases make the numeric table visible. Superposing both predictions on the same target shows what a DockQ of 0.86 and a DockQ of 0.01 look like as geometry, and it separates a displaced but contacting pose from one that recovers no native contact at all. EpoR is the case that carries the argument of Section 5.2: both models reproduce its crystal interface at nearly the same high quality, yet the two leave that binder 88 ranks apart in the same library. AcrB is the one failure that survives recomputation on rebuilt native coordinates.

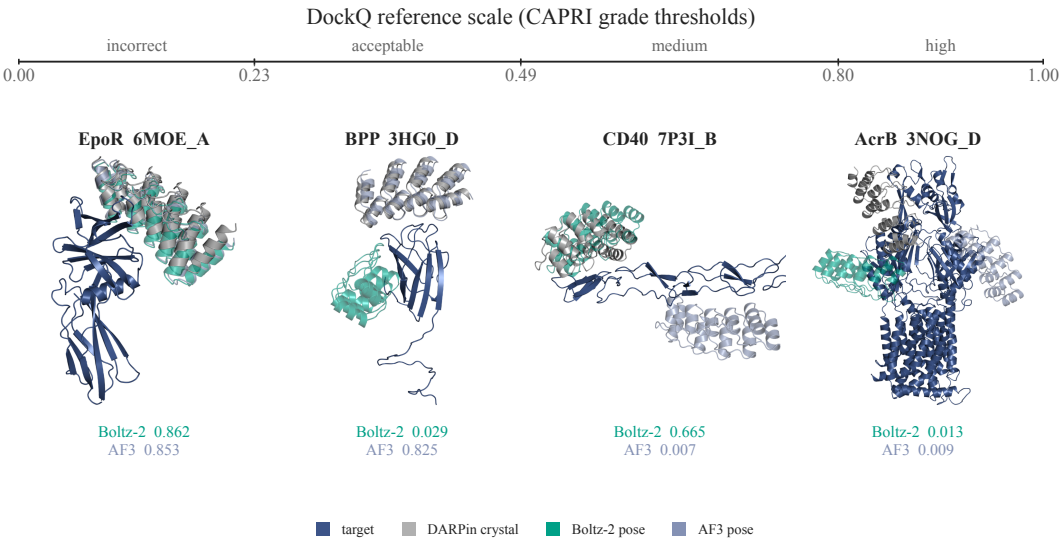

Figure S4: Crystal-positive pose controls. For four crystal-positive cases (EpoR 6MOE_A, BPP 3HG0_D, CD40 7P3I_B, AcrB 3NOG_D), the two model predictions are superposed on a single target so that pose differences are directly visible. Each overlay shows exactly four objects: the target (dark blue, from the native complex, drawn once), the crystal DARPin (light grey, ground-truth reference), the Boltz-2 DARPin pose (teal, semi-transparent) and the AF3 DARPin pose (slate, semi-transparent); the predicted target chains are not drawn, and superposition is anchored on the target chain. The top ruler marks the CAPRI DockQ grade thresholds (0.23 / 0.49 / 0.80). DockQ values below each panel are the values of Table S13. EpoR is the hero case: both models recover the crystal interface at nearly the same high quality (0.862 / 0.853), yet that binder sits at rank 90 under Boltz-2 and at rank 2 under AF3. AcrB is the one genuine failure that survives recomputation (0.013 / 0.009, with $f_{\mathrm{nat}} = 0.000$, so not a single native contact is recovered). BPP and CD40 show the two models trading wins. The crystal subset of 26 comparisons is not extrapolated to the full library.

E.3 INTERFACE COMPONENTS AND CONFIDENCE SEMANTICS

The same 26 comparisons, with the interface quantities that DockQ combines and with each model's confidence recorded under its own definition. Reading f_{nat} beside LRMSD separates two ways of being wrong: on AcrB no native contact is recovered in any of the four comparisons, whereas elsewhere a large LRMSD can coexist with recovered contacts under a displaced global pose. Neither confidence column is a calibrated probability, so a high value certifies neither the interface nor that binder's rank in its library.

Table S14: DockQ on the crystal-binder subset with interface components and confidence semantics: 13 complexes, one Boltz-2 and one AF3 comparison each ($n = 26$). f_{nat} is the recovered fraction of native contacts, iRMSD the interface RMSD and LRMSD the ligand RMSD. The confidence column is each model's own output and the two are not on a common scale: AF3 reports a model ranking score, not ipTM and not a calibrated probability, and the Boltz-2 complex confidence is likewise not a calibrated binding probability. Reading f_{nat} beside LRMSD separates two ways of being wrong: on AcrB no native contact is recovered at all, whereas a large LRMSD with a non-zero f_{nat} indicates recovered contacts under a displaced global pose. The AcrB, EpoR, HER2_DIV and VEGFR2 natives were assembled directly from the deposited entries, and every pairing was verified computationally and by a structural-biology expert. Computed with DockQ 2.1.3 (bjornwallner/DockQ).

Unit	Structure	Model	DockQ	CAPRI	f_{nat}	iRMSD (Å)	LRMSD (Å)	ipTM	Model score
BCL-W	4K5A_B	Boltz-2	0.537	medium	0.655	1.39	10.03	0.83	0.905
	4K5A_B	AF3	0.707	medium	0.927	0.95	8.86	0.80	0.832
GFP	9F24_C	Boltz-2	0.013	incorrect	0.000	13.80	51.42	0.50	0.833
	9F24_C	AF3	0.365	acceptable	0.373	2.89	8.35	0.19	0.279
EpoR	6MOE_A	Boltz-2	0.862	high	0.870	0.54	3.83	0.75	0.909
	6MOE_A	AF3	0.853	high	0.926	0.74	3.88	0.93	0.920
JNK1	5LW1_A	Boltz-2	0.810	high	0.947	0.97	4.53	0.49	0.836
	5LW1_A	AF3	0.431	acceptable	0.614	3.28	8.40	0.60	0.622
BPP	3HG0_D	Boltz-2	0.029	incorrect	0.028	10.47	41.47	0.89	0.908
	3HG0_D	AF3	0.825	high	0.917	0.73	4.88	0.91	0.949
VEGFR2	5OYJ_A	Boltz-2	0.046	incorrect	0.065	16.14	32.31	0.74	0.885
	5OYJ_A	AF3	0.662	medium	0.839	1.69	5.48	0.77	0.778
	5OYJ_D4b	Boltz-2	0.042	incorrect	0.048	15.85	31.64	0.53	0.849
	5OYJ_D4b	AF3	0.782	medium	0.887	0.91	5.20	0.92	0.912
AcrB	3NOG_D	Boltz-2	0.013	incorrect	0.000	22.56	45.07	0.48	0.789
	3NOG_D	AF3	0.009	incorrect	0.000	20.03	58.97	0.26	0.341
	3NOG_E	Boltz-2	0.012	incorrect	0.000	22.72	46.04	0.18	0.727
	3NOG_E	AF3	0.007	incorrect	0.000	22.71	66.72	0.27	0.351
CD40	7P3I_B	Boltz-2	0.665	medium	0.690	1.43	4.49	0.56	0.873
	7P3I_B	AF3	0.007	incorrect	0.000	21.21	66.54	0.11	0.197
	7P3I_D	Boltz-2	0.705	medium	0.759	0.78	7.42	0.72	0.903
	7P3I_D	AF3	0.006	incorrect	0.000	22.64	69.46	0.13	0.214
HER2_DI	9_29	Boltz-2	0.437	acceptable	0.451	2.45	7.15	0.65	0.889
	9_29	AF3	0.011	incorrect	0.000	13.88	55.96	0.72	0.738
HER2_DIV	4HRN_A	Boltz-2	0.919	high	0.814	0.36	0.54	0.94	0.971
	4HRN_A	AF3	0.911	high	0.831	0.46	0.90	0.93	0.932

F ENRICHMENT, CALIBRATION, AND COST

What a screening method is worth in practice is a hit rate at a depth someone can afford to assay, and what it costs is wall-clock time at that depth. Both are read against the library base rate of 66 positives among 5,000 pairs, or 1.3 per cent: every method's hit-rate curve decays toward that rate well before the library is exhausted, so a high score enriches the very top and does not buy a rate that would let one skip the assay. Cost moves the other way, rising with target size for the fold and docking families while the sequence methods stay flat, and the extra hours do not purchase an earlier first positive (panel (b) of Figure S5).

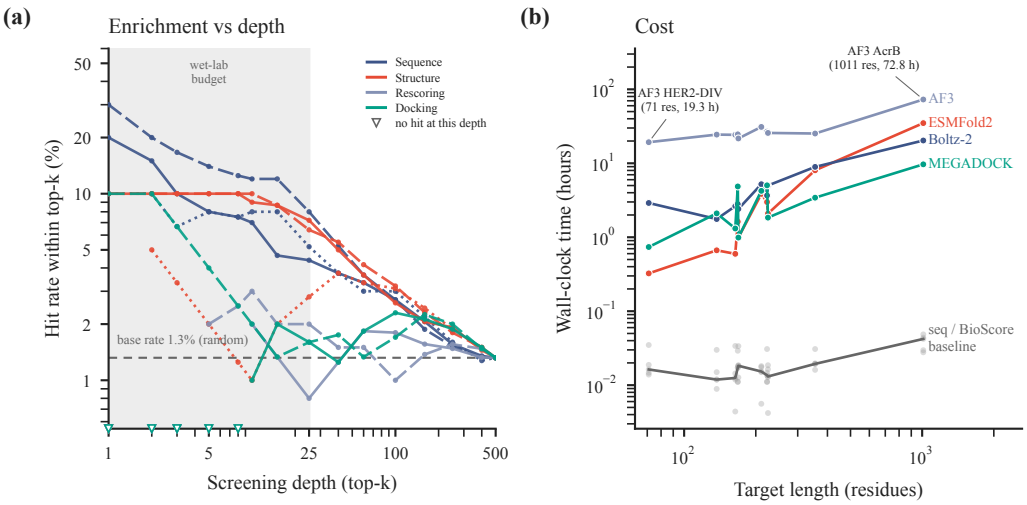

Figure S5: Enrichment and cost. (a) Hit rate, that is pooled precision, within the top-k against screening depth k on log–log axes, for all ten methods colored by family. The dashed horizontal line is the library base rate ($66/5000 = 1.3\%$), labeled as the random expectation; the shaded band marks the wet-lab budget, the first $\sim$25 candidates that form the top-5% window. Every curve decays toward the base rate by $k = 500$. (b) Cost: wall-clock time per one successful 500-pair run against target length, both axes logarithmic. The sequence-family methods and the two BioScore variants are summarized as a single sequence / BioScore baseline, drawn as the per-unit median with the individual methods shown as light semi-transparent scatter. ESMFold2, Boltz-2, MEGADOCK and AF3 are drawn as separate lines, MEGADOCK as the per-unit median of its two variants. AF3 is annotated at HER2_DIV (71 residues, $\sim$19.3 h) and AcrB (1011 residues, $\sim$72.8 h).

F.1 SCORE CALIBRATION

Panel (a) above replaces a reliability diagram, and the reason is the base rate. At 1.3 per cent a reliability curve mostly restates that base rate, and the mapped score is not claimed to be a probability at any point in this work. The calibration error is therefore reported as a scalar per method, with the empty-bin convention stated, and it bounds how far a high score is from an experimental hit rate rather than measuring a probability.

Table S15: Expected calibration error of the min–max rescaled scores, computed per unit on ten equal-width bins and averaged over units. An empty bin contributes zero to the weighted sum and the denominator stays the full pair count of the unit; no statistic is fabricated for a bin that holds no pairs. The rescaled score is a score-derived diagnostic, not an affinity and not a calibrated probability, so these values bound how far a high score is from an experimental hit rate rather than measuring a probability.

Method	mean ECE	worst unit	empty bins
ESMC	0.4945	0.7339	3
ProtT5	0.4311	0.6310	2
D-SCRIPT	0.3700	0.5121	12
ESMFold2	0.2208	0.4716	4
Boltz-2	0.5270	0.6875	0
AF3	0.3557	0.6493	0
BioScore←Boltz-2	0.5152	0.5926	0
BioScore←AF3	0.4856	0.5709	0
MEGADOCK←Boltz-2	0.2401	0.3339	2
MEGADOCK←AF3	0.2729	0.3637	1

F.2 WALL-CLOCK TIMING

Timing is reported per method and per unit, since the two carry different information: the per-method grid gives what a family costs, and the per-unit grid shows that the cost of the fold family is driven by target length rather than by library size, which is fixed at 500 pairs everywhere. Panel (b) of Figure S5 is the same quantity read against target length.

Table S16: Wall-clock hours to score one 500-pair unit, by method. Each unit was scored as a single job on one NVIDIA A40 GPU, so an entry is that job's wall-clock time on one device and not a sum of GPU hours over several of them. Hours are independent of the labels, so they are unchanged by any relabeling. The sequence family and both BioScore runs stay flat in target length; the fold and docking methods rise with it.

Method	Family	Units	Median	Mean	Min	Max
ESMC	seq	10	0.01	0.02	0.01	0.05
ProtT5	seq	10	0.02	0.02	0.01	0.03
D-SCRIPT	seq	10	0.03	0.02	0.00	0.03
ESMFold2	fold	10	1.86	5.61	0.32	34.93
Boltz-2	fold	10	3.29	5.57	1.76	20.32
AF3	fold	10	24.94	29.59	19.27	72.85
BioScore←Boltz-2	rescore	10	0.01	0.02	0.01	0.04
BioScore←AF3	rescore	10	0.01	0.02	0.01	0.04
MEGADOCK←Boltz-2	dock	10	2.83	3.65	0.60	12.74
MEGADOCK←AF3	dock	10	2.70	3.19	0.75	6.61

Table S17: Wall-clock hours per unit and method, with target length L in residues. Reading cost against target length rather than against accuracy is what shows that the additional hours of a larger structure model do not buy an earlier first positive: AF3 spends about 19.3 hours on the 71-residue HER2_DIV target and about 72.8 hours on the 1011-residue AcrB target, while the sequence family stays under a tenth of an hour at every length.

Unit	L	ESMC	ProtT5	D-SCRIPT	ESMFold2	Boltz-2	AF3	BioScore←Boltz-2	BioScore←AF3	MEGADOCK←Boltz-2	MEGADOCK←AF3
BCL-W	164	0.01	0.03	<0.01	0.60	2.63	24.2	0.01	0.01	0.60	2.02
GFP	225	0.02	0.02	<0.01	2.09	5.01	25.7	0.01	0.01	1.85	1.84
EpoR	211	0.02	0.02	<0.01	3.85	5.21	31.0	0.02	0.01	4.14	4.34
JNK1	355	0.02	0.02	0.03	8.04	8.93	25.3	0.02	0.02	3.57	3.28
BPP	137	<0.01	0.01	0.03	0.67	1.76	24.4	0.01	0.01	2.09	2.12
VEGFR2	223	0.01	0.01	0.03	2.99	3.67	26.8	0.01	0.01	4.24	5.83
AcrB	1011	0.05	0.03	0.03	34.9	20.3	72.8	0.04	0.04	12.7	6.61
CD40	168	0.01	0.01	0.03	1.62	2.80	24.6	0.02	0.02	5.62	4.10
HER2_DI	169	0.01	0.02	0.03	0.99	2.41	21.7	0.02	0.02	0.96	1.02
HER2_DIV	71	0.02	0.02	0.03	0.32	2.91	19.3	0.01	0.01	0.72	0.75

G WHAT THESE RESULTS ASK OF LATER METHODS

The frozen scores set a measurement standard as well as a reference line, and the standard is stricter than a single aggregate number. Ten units, five of them with fewer than three positives, put most 95% Wilson intervals between 0.3 and 0.5 in width, and leave a rate of 0 with an upper bound near 0.28. A later method that reports a Top-k rate higher by one or two units has moved inside that width, so an improvement claim on this benchmark needs the per-unit view of Appendix C.3 beside the aggregate: which units changed, whether the gain is confined to the units with many positives, and whether the first-positive rank moved on the units that all ten methods currently miss.

The metric a later method should be judged on follows from the same evidence. Median EF@5% is zero for eight of the ten methods, which makes it a poor discriminator near the current operating point, and a rate defined at three cut-offs discards the distance between rank 26 and rank 300. The first-positive rank distribution over the ten units, together with the hit rate at the depth a wet-lab campaign can actually assay, carries what the aggregate cannot: both are released here, in Table S9 and in panel (a) of Figure S5.

Where the separable signal lives also constrains what kind of gain is plausible. A k-mutant decoy sits about two substitutions from its nearest positive at variable positions on a fixed framework, so the discrimination a method needs is residue-level within one fold, not global pose quality. The crystal subset shows the two come apart in both directions: on EpoR the two structure models reach almost the same DockQ yet leave its binder at rank 90 and rank 2 respectively, while HER2_DIV combines high DockQ with rank 1 for six methods. A method that improves the pose of a complex it already gets roughly right should not be expected to move a library ranking, and a method that improves ranking need not improve pose.

Modes 2 and 3 are where the question changes from scoring to training, and they carry their own reporting requirements. In Mode 2 a sequence is a binder only on the unit where it is a Mode 1 positive, so cross-target hits are errors and a submission that pools targets hides them. In Mode 3 the HER2 epitope pair shares candidates, so a fold that holds out one epitope tests related-epitope transfer and has to be reported separately from the folds that hold out an unseen target. Neither mode carries scores here, which leaves the split rather than a ranking as the object under comparison.

Two boundaries stay in force for any claim built on this release. The crystal evidence covers 26 comparisons on 13 complexes and does not extend to units without co-crystals, so pose statements remain case-level. And the scores are not probabilities under any of the ten definitions in Table S7: a mapped score supports a ranking and a calibration diagnostic, and it does not support an expected hit rate that would replace the assay.

H RELEASE CONTRACTS AND REPRODUCIBILITY

The data described below are submitted as supplementary material, in one archive separate from this PDF. The archive contains the DARPinDB records; Mode 1 sequences, labels, and frozen scores, with the identifier of the selected sample and AF3 per-structure ipTM and pTM beside the ranking score; Mode 2 and Mode 3 sequences and labels without scores; the DockQ coordinates and chain maps for the 13 cases; and the tables read by the figures in this appendix. It also contains the expert-chosen predicted reference complexes for Mode 1 positives that lack their own co-crystal, with a manifest of the cases for which no model was acceptable. It does not contain plotting scripts.

Record and label contract. The release unit is a $(\mathrm{DARPin}, \mathrm{target})$ record. Every record carries a binding label together with the evidence class that produced it, so that a consumer can separate a literature positive from a SYNBIP entry, a co-crystal, and an SPR measurement. A label of 0 inside a Mode 1 library is a task negative and not an assay result unless its type says otherwise; Appendix B.3 gives the counts. Empty cells mean absent. The strings `NA`, `None` and − are not used as missing-value sentinels anywhere in the release, so a consumer must not treat them as such.

Structure tiering and chain maps. Structures are tiered rather than flagged: complex, target-only, or none, with 50, 50 and 88 records respectively. A DARPin-only monomer stays in the structure tree as scaffold reference and does not enter the complex tier, because treating a monomer as a complex would put later docking and pose evaluation on the wrong premise. Chain correspondence is frozen separately from the coordinate files: after cleaning, the DARPin is chain A and the target is chain B in every released complex, the pre-clean author chain identifiers are recorded alongside because they differ for several entries including 4HRL, 9F24 and 7P3I, and any further chain in the deposited entry is outside the binary complex used here. The DockQ subset releases, for each of its 13 unique identifiers, the native complex, the Boltz-2 top-1 prediction, the AF3 top-1 prediction, and that case's chain map.

Score semantics. Each method's score is released under its own definition and direction, and the definitions are not interchangeable. The quantity that orders the library was verified against the per-model outputs rather than taken from documentation: ESMFold2 and Boltz-2 both rank by ipTM, a structural proxy and not a binding probability, while AF3 ranks by its model ranking score, which is neither ipTM nor a calibrated probability. AF3 per-structure ipTM and pTM are released alongside the ranking score so that the two structure models can be compared on one scale. None of the sequence methods is a K_{D} regressor. Where scores are mapped to $[0, 1]$ for a calibration diagnostic, the mapping is per method and per unit, and the mapped quantity is a score-derived diagnostic rather than an affinity or a probability. Reliability bins are released with an explicit empty-bin convention: an empty bin contributes zero to the weighted sum, the denominator remains the full pair count of the unit, and no statistic is fabricated for a bin that holds no pairs.

Figure provenance. Each figure in this appendix is backed by the tables it reads, and those tables are in the supplementary archive, so that a consumer can check the plotted values rather than trust the drawing. Figure S1 reads the candidate-level recall table and the first-positive rank table; Figure S2 the top-window composition table; Figure S3 and Figure S4 the DockQ table that also produces Table S13 and Table S14; and Figure S5 the candidate-level ranks together with the timing ledger.

What is not released and what is not claimed. Predicted reference complexes are expert selections among sampled candidates; they are labeled as predictions and never serve as DockQ natives. Modes 2 and 3 release sequences and labels only, so neither can be read as a completed cross-target generalization experiment. The crystal subset covers 26 comparisons and does not extrapolate to the $5{,}000$ library members. No wet-lab guidance is included.

Score provenance. Every released score is matched back to the model run that produced it, pair by pair, against that model's own output files rather than inherited from an intermediate summary. Sequence inputs carry no non-standard residue placeholders: crystallographic waters are not written into a chain, and each score comes from the run on the cleaned input. Where a model emits several samples per pair, the released score is the best-ranked sample under that model's own criterion, and

the identifier of that sample is released beside it, so that a reported number can be traced to one specific prediction rather than to a pair.

Compute environment. Every run reported here was executed on a shared GPU cluster, one job per method and unit, with each job confined to a single NVIDIA A40 GPU (48 GB). The wall-clock hours in Appendix F.2 and in panel (b) of Figure S5 are therefore the time of one such job on one device rather than aggregate GPU hours, and no run was distributed across devices, so the size of the cluster does not enter any reported number.

Coordinate-to-row binding. The structure manifest binds each coordinate file to the DockQ row it supports. A consumer can therefore check that the coordinates behind a reported DockQ value are the ones the number was computed from, rather than inferring the correspondence from file names.