SEED2SUCCEED: TEST-TIME SCALING NEEDS BREADTH IN ANTIBODY–ANTIGEN STRUCTURE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time scaling in antibody–antigen structure prediction offers a route to improve its accuracy, but its benefits depend on how prediction budgets are allocated. We present Seed2Succeed, a systematic study of seed breadth and sampling depth in AlphaFold 3 using 113 FoldBench complexes and 2.26 million predicted structures. For each target, we compare 1,000 seeds $\times$ 10 diffusion samples with 10 seeds $\times$ 1,000 samples at a matched budget of 10,000 structures. We jointly analyze predicted epitope and paratope diversity, local structure diversity and global structure diversity. On 160 interfaces from 104 targets, broader seeding increases low-quality success rate from 43.1% to 60.0%, high-quality success rate from 16.3% to 25.0%. Broader seeding also expands contact coverage, while coordinate dispersion in the complementarity-determining regions remains nearly unchanged at full budget. However, additional sampling can reduce confidence-selected success, and full-pool oracle discovery numerically favors depth. Cluster analysis reveals structural populations with similar interface confidence but markedly different docking quality. These findings support seed breadth as a practical starting point on this benchmark and show that effective test-time scaling requires jointly considering candidate discovery, structural exploration, and ranking reliability.

1 INTRODUCTION

Predicting antibody–antigen complexes requires both determining the correct relative binding orientation and generating plausible molecular structures. While advancements in protein folding and biomolecular co-folding methods have made such predictions increasingly feasible (Jumper et al., 2021; Evans et al., 2022; Abramson et al., 2024), accurately predicting the antibody–antigen interface remains challenging in benchmarks based on experimental structures (Xu et al., 2026; Fromm et al., 2026). High-confidence structural predictions of the antibody alone are insufficient to determine which surface region recognizes the antigen or how the two molecules assemble (Fromm et al., 2026; Ünsal et al., 2026). Understanding how to generate and screen a set of candidate structures is key to effectively leveraging additional computational inference.

A natural approach to addressing uncertainty is to generate a larger number of predictions. Large-scale stochastic sampling has been shown to improve challenging multimer structure predictions (Wallner, 2023), while test-time scaling offers a broader framework for enhancing results by increasing the allocation of inference-time resources (Snell et al., 2025). However, a larger set of candidates is only truly valuable if it includes superior structures and makes them accessible to downstream users. These are distinct requirements: a set of structures might contain the correct binding pose yet fail to retain it during confidence-based ranking, or it might exhibit high structural diversity while completely failing to identify the native interface (Corso et al., 2023; Ünsal et al., 2026). Consequently, research into scaling must clearly define both the candidate generation budget and the selection rules used to transform generated outputs into final predictions.

This work adopts two approaches to expand the search scope: running more random seeds or generating more diffusion samples per seed. Samples within the same seed share upstream computations, whereas introducing a new seed alters the random inputs in the prediction pipeline (Abramson et al., 2024). Given a fixed total number of generated structures, allocating more candidates to one

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.

dimension inevitably means allocating fewer to the other. Concentrating a large number of samples within a few seeds allows for repeated exploration of similar structural populations through increased sampling; distributing samples across more seeds, however, may alter the structural populations covered by the search. Yet, it remains unclear which of these two approaches yields superior prediction performance. Therefore, empirical comparison is essential to evaluate them; one cannot equate an identical number of output structures with sample independence, equal computational costs, or equivalent search effectiveness.

Against this backdrop, we pose a specific question: *given a fixed budget of candidate structures, can allocating inference computation across a larger number of random seeds improve the quality of structures retained after confidence-based ranking?* We further examine the changes that occur in the resulting set of structures. Broader contact coverage, larger maximum displacements, and greater diversity within the complementarity-determining regions (CDRs) measure distinct properties; none of these metrics serves as definitive evidence of a more accurate interface.

Our contributions primarily span three areas: We constructed a dataset comprising over two million predicted antigen-antibody structures, which can serve as a resource for experimentally investigating potential new targets and binding modes. Experimental results demonstrate that, for a fixed total number of generated samples, "test-time scaling" significantly improves antigen-antibody structure prediction; specifically, conducting a broad search using a larger number of random seeds yields superior results. Analysis of the large-scale antigen-antibody structure dataset reveals that epitope diversity, paratope diversity, and global structural diversity of the complexes all increase markedly as the search scope expands; furthermore, there is a significant correlation between these increases in diversity and higher docking success rates. Within this benchmarking framework, broad exploration using multiple random seeds serves as an effective starting point for confidence-based search strategies, providing high-quality analytical data to support subsequent antigen-antibody structure prediction modeling and specific biological experiments.

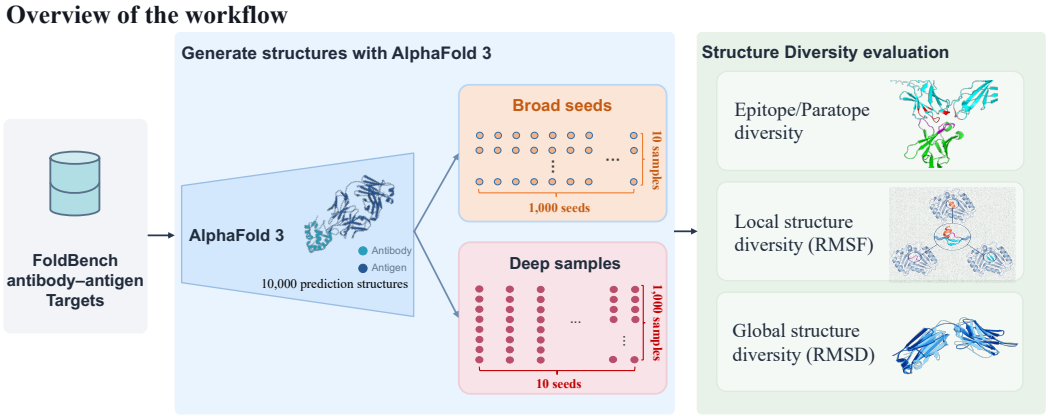


Figure 1: **Sampling and evaluation workflow.** Each of 113 targets receives two 10,000-structure AF3 ensembles: BROAD uses 1000×10 seeds and samples; DEEP uses 10×1000.

2 RELATED WORK

Sampling for structure prediction. AlphaFold2, AlphaFold-Multimer, and AF3 have advanced the field of structure and complex prediction (Jumper et al., 2021; Evans et al., 2022; Abramson et al., 2024), while Protenix (Zhang et al., 2026b) and OpenDDE (OpenDDE Project, Aureka AI Research, 2026) have significantly improved performance metrics for the specific task of predicting antigen-antibody complexes. AFsample demonstrated the improvements gained through large-scale stochastic multimer sampling (Wallner, 2023); evaluations focused on antibodies have also indicated that both sampling and model architecture are influential factors (Fromm et al., 2026; Ünsal et al., 2026).We examine the allocation between AF3 seeds and samples as distinct dimensions, while keeping the predictor and the total number of outputs constant.

Allocation and confidence-guided search. Test-time allocation and diffusion-based search can enhance inference effectiveness (Snell et al., 2025; Zhang et al., 2026a). While DiffDock decouples generation from ranking, Confidence Bootstrapping leverages confidence feedback to improve docking generalization (Corso et al., 2023; 2024). Proteína-Complexa applies generative modeling and inference-stage optimization to the design of binding molecules (Didi et al., 2026). Our study focuses on structure prediction for fixed sequences and evaluates seed allocation strategies, rather than training a generator or optimizing the sequences of binding molecules.

Evaluation and diversity. FoldBench provides a unified biomolecular benchmark (Xu et al., 2026), while DockQ measures the consistency of native contacts and geometric structures—a metric that has been extended to multi-chain systems (Basu & Wallner, 2016; Mirabello & Wallner, 2024). Our analysis integrates accuracy with contact and coordinate diversity (Ramaraj et al., 2012; Jing et al., 2024; Lu et al., 2024) while preserving their distinct statistical units, rather than treating the structural distribution range itself as sufficient evidence of prediction quality.

3 METHOD

3.1 SAMPLING DESIGN AND EVALUATION COHORTS

We selected AlphaFold3 as the generator for antigen-antibody structures. For the structure data, we utilized the 113 antigen-antibody complexes in Foldbench. Let t denote the target index, and let $X_{t,s,d}$ represent the predicted structure generated from seed s and diffusion sample d. We compared the BROAD sampling strategy—with $(S, D) = (1000, 10)$—against the DEEP sampling strategy—with $(S, D) = (10, 1000)$—ensuring that both satisfied the condition $B = S \times D = 10,000$. The complete dataset of antigen-antibody complexes comprised $113 \times 2 \times 10,000 = 2,260,000$ structures. Pairwise comparisons were defined by the target sequences, chain role assignments, and native reference structures. Figure 1 outlines the workflow. We matched the number of generated structures and recorded parameters such as execution time and confidence scores for each task. Selected success uses a fixed benchmark-result intersection of 160 interfaces from 104 targets, shared across configurations; diversity, ensemble DockQ, and oracle discovery use all 113 targets. Appendix A records cohort membership. Interfaces within a complex need not be independent.

3.2 CONFIDENCE SELECTION AND INTERFACE ACCURACY

For a prefix (s_0, d_0), we take the first s_0 numeric seeds and d_0 sample indices. Candidates are ordered by stored AF3 `ranking_score`, with ties broken by ascending seed and sample index, giving a Top-k set $\mathcal{T}_{t,k}$ of size $\min(k, s_0 d_0)$. Ranking uses neither native structures nor the cluster-average ipTM shown in Figure 5.

For an interface i and model m, DockQ combines the recovered fraction of native contacts f_{nat}, interface RMSD iRMSD, and ligand RMSD LRMSD (Basu & Wallner, 2016):

$$q_{t,i}(m) = \frac{1}{3}\left[f_{\text{nat}} + \frac{1}{1 + (\text{iRMSD}/1.5\text{Å})^2} + \frac{1}{1 + (\text{LRMSD}/8.5\text{Å})^2}\right]. \tag{1}$$

The selected-interface analysis uses OpenStructure chain mapping and a validated CUDA implementation of its DockQ definition. The all-model analysis uses declared-interface DockQ v2 results (Xu et al., 2026; Mirabello & Wallner, 2024). We use the strict thresholds $\tau \in \{0.23, 0.49, 0.80\}$ implemented in the displayed success figures. Thus,

$$\text{SR}_k(\tau) = \frac{1}{160} \sum_{(t,i)\in\mathcal{I}_{160}} \mathbf{1}\left[\max_{m\in\mathcal{T}_{t,k}} q_{t,i}(m) > \tau\right]. \tag{2}$$

For $k > 1$, this is the best quality among confidence-selected candidates for each interface; different interfaces may attain their maxima in different models. all reported configurations retain the same 160 interfaces.

For the all-model analysis, we instead require joint interface quality within one model:

$$q_t^{\min}(m) = \min_{i\in\mathcal{I}_t} q_{t,i}(m), \qquad \overline{q}_t = \frac{1}{10{,}000} \sum_{m=1}^{10{,}000} q_t^{\min}(m). \tag{3}$$

This *bottleneck DockQ* requires all interfaces to be accurate within one model. Its mean measures typical quality; its maximum measures oracle discovery using native knowledge unavailable at deployment.

3.3 CONTACT-SPACE COVERAGE

We use an asymmetric contact definition. Antigen residue r and antibody residue j contact when an antigen heavy atom lies within 4 Å of antibody Cα_j:

$$c_m(r,j) = \mathbf{1}\left[\min_{a \in \text{heavy}(r)} \|x_{m,a} - x_{m,\text{C}\alpha_j}\|_2 \le 4\text{Å}\right]. \tag{4}$$

Only declared antibody–antigen chain pairs are considered. Let $E_m = \{r : \exists j,\ c_m(r,j) = 1\}$ and $P_m = \{j : \exists r,\ c_m(r,j) = 1\}$. For a candidate subset $\mathcal{M}$, the two cumulative coverage metrics are

$$C_t^{\text{Ag}}(\mathcal{M}) = \frac{|\bigcup_{m \in \mathcal{M}} E_m|}{L_t^{\text{Ag}}}, \qquad C_t^{\text{Ab}}(\mathcal{M}) = \frac{|\bigcup_{m \in \mathcal{M}} P_m|}{L_t^{\text{Ab}}}. \tag{5}$$

The denominators are the full antigen-chain residue count and full antibody Cα residue count, respectively. The figure labels "epitope" and "paratope diversity" denote these predicted footprints, not native-site recovery or counts of distinct epitopes.

3.4 COORDINATE DIVERSITY AND STRUCTURAL CLUSTERING

Each predicted antigen is sequence matched and jointly aligned to native antigen Cα coordinates using the Kabsch transformation (Kabsch, 1976). Applying that same transformation to the antibody yields coordinates $z_{m,j}$ in a common antigen-fixed frame. No additional antibody fit is performed. For antibody Cα atoms $\mathcal{A}_t$ and a subset $\mathcal{M}$, global exploration is measured by the maximum pairwise distance:

$$d_t(m,n) = \sqrt{\frac{1}{|\mathcal{A}_t|} \sum_{j \in \mathcal{A}_t} \|z_{m,j} - z_{n,j}\|_2^2}, \qquad D_t^{\max}(\mathcal{M}) = \max_{\substack{m,n \in \mathcal{M} \\ m<n}} d_t(m,n). \tag{6}$$

This captures pose and conformational differences, including large translations of the antibody. It is not RMSD to the native antibody. Full-CDR ensemble RMSF is

$$\text{RMSF}_t(\mathcal{M}) = \sqrt{\frac{1}{|\mathcal{C}_t|\,|\mathcal{M}|} \sum_{j \in \mathcal{C}_t} \sum_{m \in \mathcal{M}} \|z_{m,j} - \bar{z}_j\|_2^2}, \qquad \bar{z}_j = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} z_{m,j}. \tag{7}$$

ANARCI with IMGT numbering defines $\mathcal{C}_t$ from CDR1 (27–38), CDR2 (56–65), and CDR3 (105–117), including numbered insertions (Dunbar & Deane, 2016). Equation 7 pools residue variances before taking the square root. It describes dispersion of independent predictions, not a time-resolved molecular-dynamics fluctuation. For accuracy, we separately compute CDR RMSD by replacing $\bar{z}_j$ with native CDR coordinates and averaging per-model RMSD over the ensemble.

For Figure 5, DBSCAN clusters the original pairwise RMSD matrix with $\epsilon = 2$ Å and `min_samples`=10 (Ester et al., 1996). Clusters are labeled by decreasing size, with noise unassigned. t-SNE is used only for display (van der Maaten & Hinton, 2008); cluster mean DockQ and ipTM average model-level values. Projection settings and numerical audits appear in Appendix D.

4 RESULTS

4.1 SEED BREADTH IMPROVES SELECTED SUCCESS, BUT MORE SAMPLES NEED NOT HELP

Figure 2 varies seeds and samples within the BROAD source ensemble. With one sample per seed, increasing seeds from 1 to 1,000 raises acceptable-quality Top-1 success from 43.1% to 60.0%, medium-quality success from 33.8% to 54.4%, and high-quality success from 14.4% to 25.0%. At ten samples per seed, increasing seed count from 1 to 1,000 raises the corresponding rates from 46.3%, 36.3%, and 16.3% to 56.9%, 51.9%, and 25.0%. Breadth contributes across the three thresholds, although the gains are not strictly monotone at every setting.

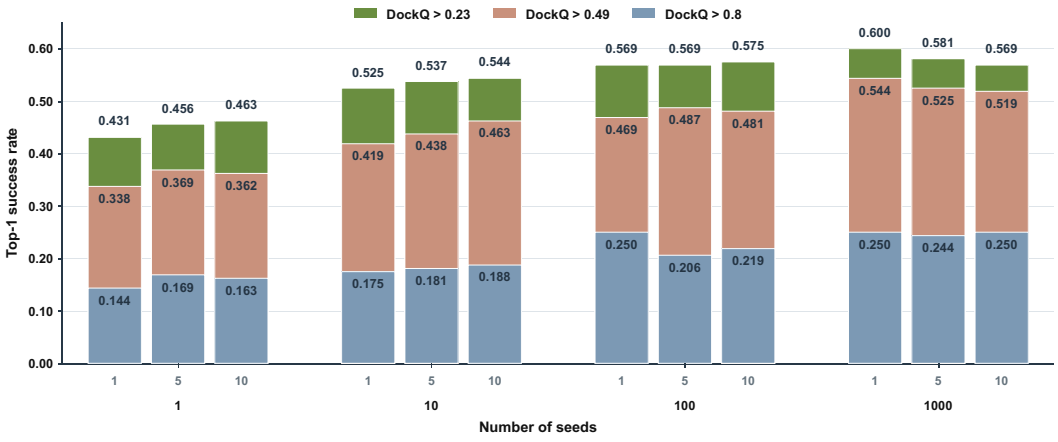


Figure 2: **Top-1 scaling across seed and sample budgets.** All configurations use prefixes of the 1000×10 ensemble and the same 160 interfaces from 104 targets.

Crucially, the figure also permits fixed-output comparisons. At 100 candidates, 100 seeds $\times 1$ sample gives 56.9%, 46.9%, and 25.0% success, compared with 54.4%, 46.3%, and 18.8% for 10×10. At 1,000 candidates, 1000×1 gives 60.0%, 54.4%, and 25.0%, compared with 57.5%, 48.1%, and 21.9% for 100×10. These comparisons support broader allocation within the available grid without assuming that every additional output has equal value.

Increasing depth at fixed breadth can even lower selected success. At 1,000 seeds, acceptable-quality success is 60.0% with one sample per seed, 58.1% with five, and 56.9% with ten. The candidate sets are nested, so their oracle best DockQ cannot decrease. Their confidence-selected Top-1 score can decrease when an added, incorrectly ranked candidate displaces a better structure. This separates a failure of selection from a failure to generate a useful candidate; it does not establish that one sample is optimal on unseen targets.

The full-budget comparison between the two separately generated allocations reinforces the selection result (Table 2). At 10,000 models, BROAD improves medium- and high-quality Top-1 success by 8.75 percentage points each. Expanding the retained set to Top-25 gives 60.0% versus 50.6% medium-quality success and 35.6% versus 25.6% high-quality success. These are descriptive paired-cohort rates, rather than significance claims based on independent interfaces.

It is worth noting that, in addition to the Top-1 selection, Top-10 and Top-25 results were also calculated for the two strategies discussed here. As the selection scope expanded, the success rates for both strategies increased; notably, under the BROAD strategy, the Top-25 result using 100×10 configuration reached an impressive 0.688 (110/160)—a figure that, while slightly lower than OpenDDE's 0.70 (OpenDDE Project, Aureka AI Research, 2026), surpassed the 0.65 result reported for Protenix-v2 (Zhang et al., 2026b). Full details of the data are presented in the appendix.

4.2 Analysis of the correlation between the number of models and diversity

The four measurements in Figure 3 probe different aspects of exploration. As the number of seeds/samples increases, epitope diversity, paratope diversity, and antigen-aligned RMSD all show significant improvement. Epitope diversity exhibits the most pronounced gain, rising from 8.8% and 10.4% to 51.3% and 49.6%, respectively. This clearly demonstrates that an exponential increase in the number of searches plays a crucial role in identifying antigen-antibody epitope structures, making it far more likely to discover potential binding sites for antigen-antibody complexes among the large pool of generated models.

Comparing the two strategies, the BROAD approach shows a 1.72 percentage point increase (95% CI: [0.84, 2.74]) in average antigen coverage relative to the DEEP pairing, while the increase in antibody contact coverage is more modest at 0.57 percentage points ([0.01, 1.23]). Specifically, regarding epitope results, BROAD outperformed DEEP on 63 targets, whereas DEEP outperformed BROAD

on 31 targets, with 19 targets showing no difference. These contact coverage ranges may include non-native contacts; thus, the increase in coverage measures the scope of the search and does not fully reflect the accuracy of the binding sites.

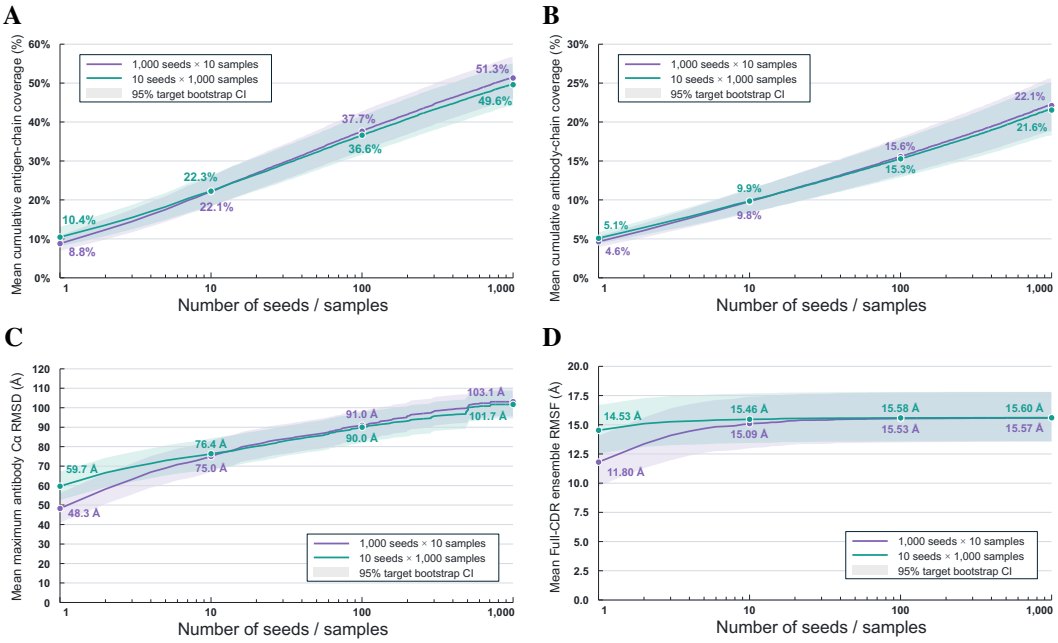


Figure 3: **Complementary diversity measurements across 113 targets.** A: epitope diversity. B: paratope diversity. C: maximum pairwise antibody Cα RMSD. D: full-CDR ensemble RMSF. Coordinate metrics use an antigen-fixed frame. Both strategies contribute $N = 10K$ structures at horizontal position K.

At $K = 1$, BROAD draws ten samples from one seed, while DEEP draws one from each of ten seeds. Its initially greater seed breadth accompanies higher coverage: 10.4% versus 8.8% on the antigen, and 5.1% versus 4.6% on the antibody. At $N = 100$, antigen coverage is nearly equal (22.3% versus 22.1%); at $N = 1000$, BROAD is ahead (37.7% versus 36.6%). This crossover reflects the grouping protocol, not a uniform advantage of within-seed depth.

Maximum antibody RMSD increases from 48.3 to 103.1 Å for BROAD and from 59.7 to 101.7 Å for DEEP. However, its full-budget paired difference is only 1.38 Å with interval $[-0.32, 3.08]$. A small number of distant, potentially incorrect poses can influence this extreme-value statistic. It supplies evidence of the extent of exploration, not of a uniformly broader useful distribution.

Full-CDR RMSF reaches a plateau much earlier. At 100 structures it is 15.09 Å for BROAD and 15.46 Å for DEEP; at full budget it is 15.572 and 15.602 Å. The paired difference is -0.030 Å ($[-0.335, 0.273]$). Consequently, increased contact coverage does not require increased overall CDR coordinate dispersion. The observations are compatible with a change in the relative occupancy of contact and pose populations, but do not identify the upstream stochastic component responsible for it.

4.3 ANALYSIS OF THE CORRELATION BETWEEN DIVERSITY AND SUCCESS RATE

Oracle discovery asks whether the full pool contains a jointly acceptable model; Table 2 evaluates selected individual interfaces on a different roster. Subtracting the percentages would not estimate a selection gap. The strongest evidence for breadth concerns retained candidates, not universal dominance in oracle discovery.

Figure 4 illustrates the relationship between four metrics of integrated exploration and the "oracle" discovery results across 113 targets. Each budget size comprises K groups of structures, with 10 structures per group. At budget sizes of 100 and 1,000 structures, the expected discovery rates for

the BROAD and DEEP strategies were 67.6% versus 66.7% and 80.8% versus 78.8%, respectively. When the budget reached the full 10,000 structures, this ranking reversed: the BROAD strategy identified at least one model with a bottleneck DockQ value greater than 0.23 for 99 of the 113 targets (87.6%), whereas the DEEP strategy achieved this for 102 targets (90.3%). At this final scale, results differed for only five targets, yielding a calculated paired difference of -2.65 percentage points (95% confidence interval: $[-7.08, 0.88]$) between BROAD and DEEP.

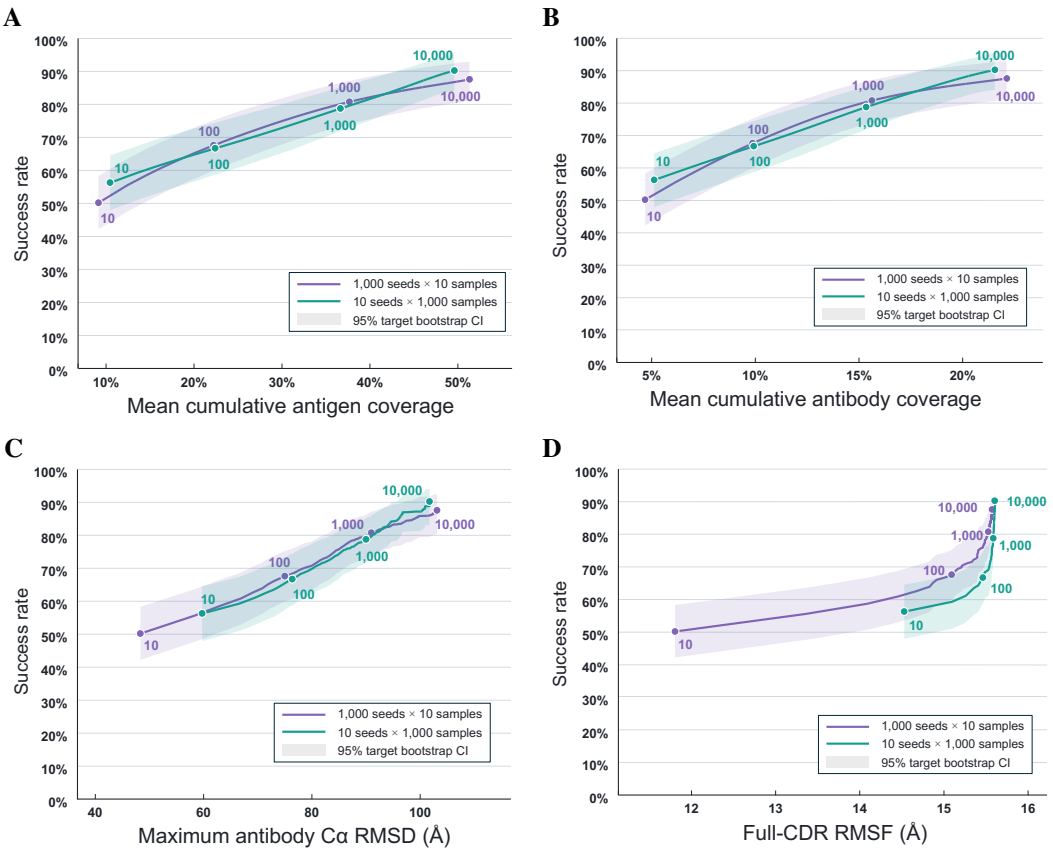


Figure 4: **Exploration and oracle discovery as budget increases.** Vertical values are expected fractions of 113 targets with bottleneck DockQ > 0.23 in a random K-group subset; labels give $N = 10K$ structures. Horizontal metrics are antigen coverage (A), antibody coverage (B), maximum antibody RMSD (C), and full-CDR RMSF (D). A uses expected coverage; B–D use subset medians.

Across the four checkpoints, both predicted antigen coverage and antibody coverage increased alongside the "oracle discovery rate" (Figure 4A–B). The Pearson correlation coefficients for these metrics were $r = 0.981/1.000$ and $r = 0.972/0.999$, respectively. The maximum RMSD of antibody Cα atoms also increased with the discovery rate (Figure 4C; $r = 0.999/0.995$). Thus, continued discovery does not require a commensurate increase in the overall dispersion of CDR coordinates. The RMSF of the entire CDR region shows relatively weak correlation across these four checkpoints (Figure 4D; $r = 0.916/0.824$). The increase in these values occurs primarily during the generation of the first 100 structures: as the number of structures rises from 100 to 10,000, the RMSF shifts only from 15.09 Å to 15.57 Å (BROAD strategy) and from 15.46 Å to 15.60 Å (DEEP strategy), respectively; meanwhile, the Oracle discovery rate rises from 67.6% to 87.6% and from 66.7% to 90.3%, respectively. Thus, continued discovery may not require a comparable increase in overall CDR coordinate dispersion.

4.4 POSE POPULATIONS REVEAL LIMITS OF CONFIDENCE AND TARGET DEPENDENCE

Figure 5 shows two target–strategy examples, each containing 10,000 models. These are illustrative cases rather than a paired strategy comparison; cluster quality averages model-level bottleneck DockQ.

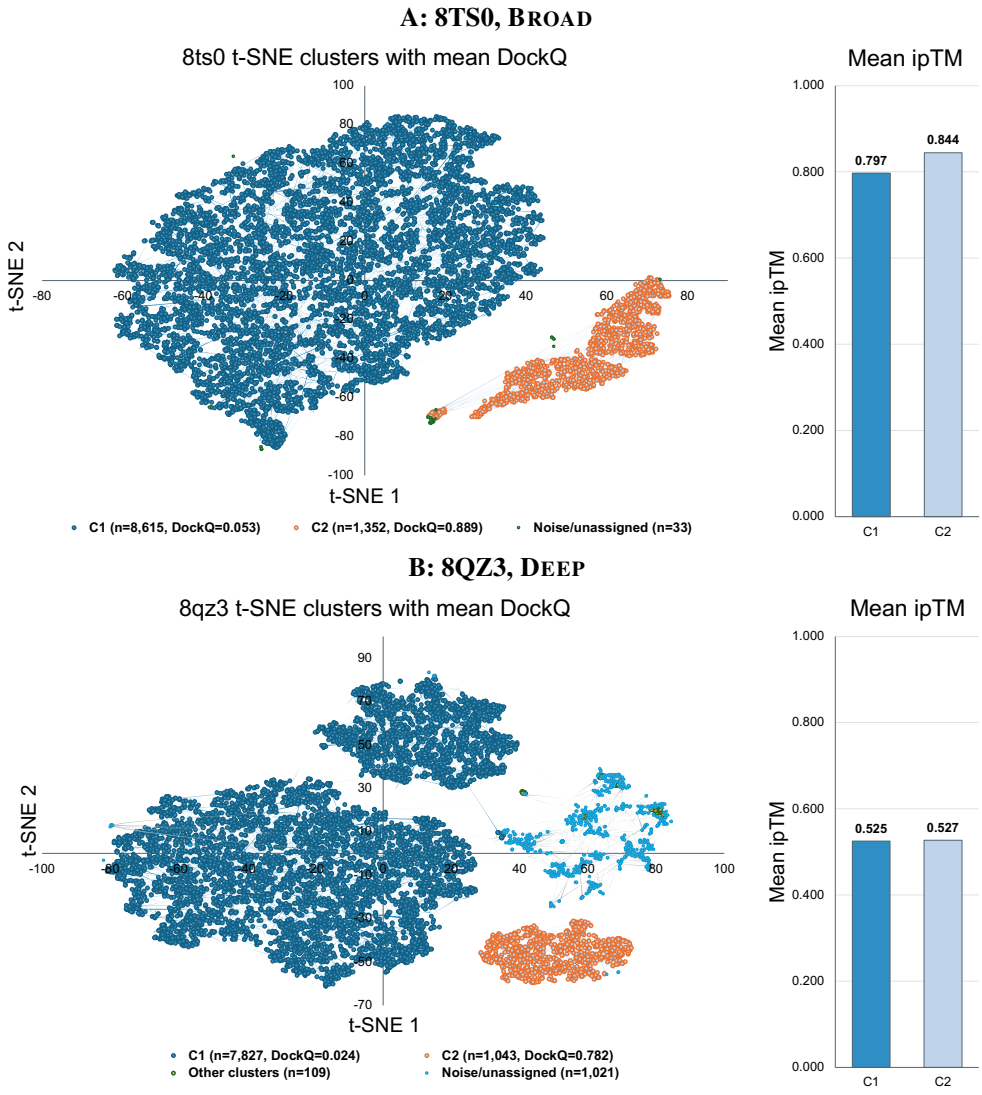


Figure 5: **Structural populations and interface confidence.** A: 8TS0/BROAD. B: 8QZ3/DEEP. DBSCAN uses antigen-aligned antibody RMSD; t-SNE supplies the display. Bars give mean ipTM of the two largest clusters. The dominant cluster has low mean DockQ in both examples. Cluster identities are local; plotted traces and separations are neither physical trajectories nor calibrated structural distances.

For 8TS0 under BROAD, C1 contains 8,615 models (86.15%) with mean DockQ 0.053, whereas C2 contains 1,352 models (13.52%) with mean DockQ 0.889; 33 models are unassigned. Their mean ipTM values are 0.797 and 0.844. The confidence difference is modest compared with the quality difference, and a high cluster-average ipTM does not guarantee a correct interface. For 8QZ3 under DEEP, C1 contains 7,827 models with mean DockQ 0.024 and C2 contains 1,043 with mean DockQ 0.782. Their mean ipTM values are nearly identical: 0.525 and 0.527. Population frequency and average confidence therefore provide incomplete evidence of docking correctness.

The corresponding unplotted ensembles reveal target dependence. In 8QZ3 under BROAD, a high-quality population contains 1,685 models with mean DockQ 0.777, compared with 1,043 at 0.782 under DEEP. Conversely, 8TS0 under DEEP has a high-quality population of 2,989 models with mean DockQ 0.901, larger than BROAD's 1,352-model population. These quality-defined comparisons do not assume that identically numbered clusters across separate fits are matched. They show that breadth can increase useful-population occupancy for one target while depth does so for another. Mean cluster ipTM alone also does not determine model-level ranking performance.

5 CONCLUSION

Seed2Succeed shows that effective test-time scaling in antibody–antigen structure prediction depends on sampling allocation and candidate selection, rather than ensemble size alone. Across 2.26 million AF3 predictions for 113 FoldBench complexes, we compared broad seeding with deep within-seed sampling at matched budgets of 10,000 structures per target. Broader seeding increased medium- and high-quality Top-1 success by 8.75 percentage points each on the shared evaluation cohort, with improvements extending to larger retained sets. It also expanded predicted epitope and paratope coverage while leaving full-CDR coordinate dispersion nearly unchanged. Together, these findings support seed breadth as a practical starting point for confidence-based prediction on this benchmark, rather than a universally superior sampling strategy.

However, full-cohort oracle discovery numerically favored depth, with an uncertainty interval including zero, while additional sampling sometimes reduced confidence-selected accuracy despite preserving previously generated correct candidates. These results distinguish discovering accurate structures from reliably selecting them. Moreover, contact coverage showed no stable positive association with oracle discovery at full budget, demonstrating that greater exploration is not itself evidence of better prediction. Effective scaling should therefore jointly address candidate generation and ranking. Breadth-first exploration followed by selective deepening and improved reranking warrants prospective evaluation on held-out targets under measured computational budgets; the present study does not establish a universally optimal allocation.

Limitations. Our study evaluates one predictor with limited cases, lacking investigation of generalization and leaving the globally optimal allocation unresolved. Timing measurements cover only 25 paired targets and do not establish computational equivalence across the full cohort. Confidence-selected success and oracle discovery use different cohorts and endpoints, so their rates cannot be directly subtracted to quantify a selection gap. Diversity measures also have interpretive limits: predicted contacts may be non-native, antigen-aligned RMSF combines pose and conformational variation, and maximum RMSD is sensitive to outliers. These analyses therefore do not establish a causal relationship between diversity and prediction accuracy.

AI USE STATEMENT

In this work, we did not use generative AI tools to generate synthetic datasets; develop theoretical models or conceptual frameworks; propose or refine hypotheses; design or provide feedback on research methodology or experiments; implement methods; assist with translation; clean or reformat datasets; support qualitative and thematic data analysis; or interpret results. Activities related to formulating mathematical claims, providing essential elements for their proofs, and writing proofs are not applicable to this work.We used generative AI tools to create or modify scientific figures or images, create or edit software code, summarize or analyze existing literature, and identify relevant literature.We reviewed all AI-assisted work. Two authors checked the AI-assisted scientific figures for accuracy, tested the AI-assisted code for correctness, and verified the literature sources and summaries.We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

Sankar Basu and Björn Wallner. DockQ: A quality measure for protein–protein docking models. *PLOS ONE*, 11(8):e0161879, 2016. doi: 10.1371/journal.pone.0161879. URL https://doi.org/10.1371/journal.pone.0161879.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kKF8_K-mBbS.

Gabriele Corso, Arthur Deng, Nicholas Polizzi, Regina Barzilay, and Tommi Jaakkola. Deep confident steps to new pockets: Strategies for docking generalization. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=UfBIxpTK10.

Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M. Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qmCpJtFZra.

James Dunbar and Charlotte M. Deane. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016. doi: 10.1093/bioinformatics/btv552. URL https://doi.org/10.1093/bioinformatics/btv552.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996. URL https://cdn.aaai.org/KDD/1996/KDD96-037.pdf.

Richard Evans, Michael O'Neill, Alexander Pritzel, et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2022. doi: 10.1101/2021.10.04.463034. URL https://doi.org/10.1101/2021.10.04.463034. Preprint, revised March 10, 2022.

Samuel Fromm, Marko Ludaic, and Arne Elofsson. Evaluating deep learning based structure prediction methods on antibody–antigen complexes. *Bioinformatics*, 42(4):btag136, 2026. doi: 10.1093/bioinformatics/btag136. URL https://doi.org/10.1093/bioinformatics/btag136.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold meets flow matching for generating protein ensembles. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22277–22303. PMLR, 2024. URL https://proceedings.mlr.press/v235/jing24a.html.

John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32:922–923, 1976. doi: 10.1107/S0567739476001873. URL https://doi.org/10.1107/S0567739476001873.

Jiarui Lu, Bozitao Zhong, Zuobai Zhang, and Jian Tang. Str2str: A score-based framework for zero-shot protein conformation sampling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=C4BikKsgmK.

Claudio Mirabello and Björn Wallner. DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 40(10):btae586, 2024. doi: 10.1093/bioinformatics/btae586. URL https://doi.org/10.1093/bioinformatics/btae586.

OpenDDE Project, Aureka AI Research. Folding, reasoning, and scaling with open-source drug discovery engine. Technical report, Aureka AI Research, 2026. URL https://github.com/aurekaresearch/OpenDDE/blob/main/assets/OpenDDE_Technical_Reports.pdf. Accessed 23 September 2026.

Thiruvarangan Ramaraj, Thomas Angel, Edward A. Dratz, Algirdas J. Jesaitis, and Brendan Mumey. Antigen–antibody interface properties: Composition, residue interactions, and features of 53 non-redundant structures. *Biochimica et Biophysica Acta—Proteins and Proteomics*, 1824(3): 520–532, 2012. doi: 10.1016/j.bbapap.2011.12.007. URL https://doi.org/10.1016/j.bbapap.2011.12.007.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Serbülent Ünsal, Benjamin Holland, Inci Sardag, and Emel Timucin. Confidence scoring for deep learning-predicted antibody–antigen complexes: AntiConf as a precision-driven metric. *Briefings in Bioinformatics*, 27(2):bbag137, 2026. doi: 10.1093/bib/bbag137. URL https://doi.org/10.1093/bib/bbag137.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL https://www.jmlr.org/papers/v9/vandermaaten08a.html.

Björn Wallner. AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics*, 39(9):btad573, 2023. doi: 10.1093/bioinformatics/btad573. URL https://doi.org/10.1093/bioinformatics/btad573.

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*, 17:442, 2026. doi: 10.1038/s41467-025-67127-3. URL https://doi.org/10.1038/s41467-025-67127-3.

XiangCheng Zhang, Haowei Lin, Haotian Ye, James Y. Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2505.23614.

Yuxuan Zhang, Chengyue Gong, Jinyuan Sun, Jiaqi Guan, Milong Ren, Song Xue, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, and Wenzhi Xiao. Protenix-v2: Broadening the reach of structure prediction and biomolecular design. *bioRxiv*, 2026b. doi: 10.64898/2026.04.10.717613. URL https://doi.org/10.64898/2026.04.10.717613. Preprint.

APPENDIX

A COHORTS, COMPUTATION, AND PROVENANCE

Three different units. A *target* is one reference assembly, a *task* is a target–allocation pair, and an *interface* is a declared antibody–antigen chain pair. There are 113 targets, 226 tasks, 2,260,000 predicted structures, and 3,440,000 all-model interface DockQ rows. Each target–allocation task contributes exactly 10,000 unique models. The full all-model interface roster has 172 interfaces per allocation. Neither 3.44 million interface scores nor 2.26 million models constitute independent target replicates.

Fixed selected-success roster. Figure 2 and Table 2 use the 160-interface roster recorded in `topk_cohort_160.csv`. This is the existing benchmark-result intersection called `protenix160` in the analysis code; it is a roster name and does not indicate that the predictions evaluated here were generated by Protenix. The retained interfaces span 104 targets. The nine targets outside this roster are 7YV1, 8DTN, 8F60, 8IM0, 8OUD, 8SFZ, 8TS0, 8XNH, and 8ZP2. Thus 8TS0 remains valid for full-cohort structural analysis and Figure 5, despite not contributing to the selected-success table. Comparisons in that table always preserve the same denominator.

Generation and scoring records. The AF3 ensembles are distinguished by the seed list and `num_diffusion_samples`: 1,000 seeds with ten samples or ten seeds with 1,000 samples. Existing target manifests specify sequences and chain roles; output keys identify target, allocation, seed, and sample. The analysis consumes stored confidence and coordinate outputs and does not regenerate structures. The selected-candidate pipeline uses OpenStructure 2.8 chain mapping and residue alignments, followed by a validated PyTorch/CUDA implementation of the OpenStructure DockQ formula. Full-ensemble bottleneck quality is obtained from audited declared-interface DockQ v2 outputs. In a 1,265-interface validation against stored OpenStructure results, all acceptable-threshold decisions agreed; quality-bin agreement was 99.92%. Mean absolute DockQ discrepancy was 0.000323 and maximum discrepancy was 0.01019. Thus numerical agreement is strong but not exact, and borderline decisions can differ. These protocols are not treated as interchangeable. Candidate scoring failures are omitted before the Top-k interface maximum; interfaces with no valid selected candidate are flagged. All reported configurations retain 160 interfaces. Exact AF3 software commit, parameter release, effective recycle count, diffusion-step count, and complete run-environment provenance were not recoverable from the local analysis bundle used for this manuscript and are not inferred from current defaults.

Alignment and atom selection. For each task, sequence matching identifies native and predicted antigen Cα correspondences. One joint antigen fit supplies the transformation applied to antibody coordinates. CDR membership follows ANARCI's IMGT numbering, including heavy and light chains when present. Contacts are evaluated between antigen heavy atoms and antibody Cα atoms, within declared chain pairs only. Antigen coverage is normalized by complete antigen-chain residue counts; antibody coverage uses the complete antibody Cα residue counts. Native CDR, antibody, and complex RMSDs also use this antigen alignment. The contact asymmetry is part of the implemented metric.

Independent audit of manuscript summaries. The accompanying `audit_results.py` checks the constant 160-interface denominator, recomputes all full-budget paired summaries, and records input SHA-256 hashes. It verifies that cluster and noise counts sum to 10,000 models for every task. Full-ensemble mean DockQ can be recovered exactly from cluster summaries because clusters plus noise partition the ensemble:

$$\bar{q}_t = \frac{1}{10{,}000} \sum_{c \in \mathcal{G}_t} n_{t,c} \bar{q}_{t,c}, \tag{8}$$

where $\mathcal{G}_t$ includes the noise group. Native quality is not used to assign DBSCAN clusters. These weighted means reproduce the all-model endpoint without treating cluster averages as equally weighted observations. Full-budget discovery is recovered by checking whether the maximum across all cluster/noise maxima exceeds 0.23. It agrees with the 99- versus 102-target endpoints in Figure 4.

B EXACT SELECTED-SUCCESS COUNTS AND BUDGET CHECKPOINTS

Table 1: Exact numerator counts for Figure 2. Every denominator is 160; all rows are numeric prefixes of the BROAD ensemble.

Seeds	Samples/seed	Candidates	$q > 0.23$	$q > 0.49$	$q > 0.80$
1	1	1	69	54	23
1	5	5	73	59	27
1	10	10	74	58	26
10	1	10	84	67	28
10	5	50	86	70	29
10	10	100	87	74	30
100	1	100	91	75	40
100	5	500	91	78	33
100	10	1000	92	77	35
1000	1	1000	96	87	40
1000	5	5000	93	84	39
1000	10	10000	91	83	40

Table 2: Full-budget selected-success counts, out of 160 interfaces. Thresholds are strict throughout this table.

Retained set	Allocation	$q > 0.23$	$q > 0.49$	$q > 0.80$
Top-1	DEEP	83	69	26
Top-1	BROAD	91	83	40
Top-10	DEEP	90	75	37
Top-10	BROAD	105	93	53
Top-25	DEEP	94	81	41
Top-25	BROAD	110	96	57

Table 3: Figure 4A checkpoints. Coverage here is the exact expected fraction of antigen residues visited by a random group subset, rather than the median-subset coverage of Figure 3A. Discovery requires one model to have bottleneck DockQ > 0.23.

K groups	N structures	Allocation	Coverage (%)	Discovery (%)
1	10	BROAD	9.139	50.225
1	10	DEEP	10.461	56.302
10	100	BROAD	22.218	67.625
10	100	DEEP	22.410	66.718
100	1000	BROAD	37.660	80.803
100	1000	DEEP	36.646	78.781
1000	10000	BROAD	51.318	87.611
1000	10000	DEEP	49.596	90.265

At full budget, all subset constructions converge to the complete ensemble, so the difference between an expectation and a median disappears. At intermediate budgets they need not agree. In particular, Figure 3A's 8.8% at ten BROAD structures and Figure 4A's 9.1% are consistent summaries of different functionals of the same group-subset distribution. For DEEP, grouping by a sample index across ten seeds is essential: using K seeds would give $1000K$ structures and would not match the displayed $10K$ budgets.

C PAIRED FULL-BUDGET EFFECTS AND EXPLORATORY UNCERTAINTY

Table 4 reports newly recomputed descriptive paired effects using the completed sources. All differences are BROAD minus DEEP; for RMSD-to-native, negative differences favor BROAD. Each target is weighted once. Intervals are percentile intervals from 10,000 paired target resamples with fixed random seed 20260922. They are pointwise, exploratory, and not adjusted for multiple comparisons. They quantify uncertainty across the observed target collection; they do not include uncertainty from retraining AF3 or rerunning the full generation experiment.

Table 4: Full-budget paired effects over 113 targets. Coverage and discovery differences are percentage points; coordinate metrics are in Å.

Metric	BROAD	DEEP	Difference	95% interval
Antigen coverage (%)	51.318	49.596	1.722	$[0.839, 2.738]$
Antibody coverage (%)	22.123	21.552	0.572	$[0.013, 1.227]$
Full-CDR RMSF	15.572	15.602	-0.030	$[-0.335, 0.273]$
Maximum antibody RMSD	103.102	101.718	1.385	$[-0.317, 3.078]$
Mean bottleneck DockQ	0.2955	0.2895	0.0060	$[0.0010, 0.0108]$
Oracle discovery (%)	87.611	90.265	-2.655	$[-7.080, 0.885]$
CDR RMSD to native	23.140	23.465	-0.325	$[-0.659, -0.023]$
Antibody RMSD to native	38.548	39.000	-0.451	$[-0.947, 0.028]$
Complex RMSD to native	29.200	29.496	-0.296	$[-0.668, 0.083]$

For antigen coverage, BROAD is higher on 63 targets, lower on 31, and equal on 19. For antibody coverage, those counts are 60, 39, and 14. Full-CDR RMSF is higher for BROAD on 57 targets and for DEEP on 56. The maximum-RMSD comparison splits 62 versus 51. Native CDR RMSD is lower for BROAD on 64 targets and for DEEP on 49. These counts make clear that an average gain is not universal target-wise dominance.

Difference from earlier analysis snapshots. Earlier project summaries used an incomplete 100-target pairing for ensemble DockQ. The all-model cluster/DockQ annotation now covers 226 complete tasks, allowing the 113-target calculation in Table 4. The former mean DockQ values and their confidence interval are therefore not used in this manuscript. Likewise, the full-budget confidence intervals in Table 4 are generated by the stated common bootstrap procedure and need not match intervals from earlier snapshots with different random seeds. The supplied figure curves and their original bands are preserved.

Table 5: Fixed-budget, across-target Spearman correlation between expected antigen coverage and oracle discovery. Existing 5,000-resample target-bootstrap intervals are shown. These are exploratory associations across 113 targets, not causal effects within a target.

Structures	Allocation	Spearman ρ	95% interval
10	BROAD	-0.001	$[-0.187, 0.192]$
10	DEEP	0.035	$[-0.155, 0.229]$
100	BROAD	-0.204	$[-0.385, -0.011]$
100	DEEP	-0.166	$[-0.352, 0.029]$
1000	BROAD	-0.255	$[-0.434, -0.054]$
1000	DEEP	-0.188	$[-0.371, 0.007]$
10000	BROAD	-0.023	$[-0.181, 0.135]$
10000	DEEP	0.080	$[-0.077, 0.226]$

At intermediate budgets, correlations are computed with exact probability ordering rather than rounded display percentages. At full budget, discovery becomes binary. Dependence among nested budgets prevents the cumulative trajectory coefficients from supporting ordinary independent-observation significance tests; the manuscript therefore reports no such p-values. Target difficulty, antigen length, and binding-site geometry can still confound fixed-budget correlations.

D CASE-STUDY DETAILS AND REPRODUCIBILITY BOUNDARIES

Table 6: The two largest structural populations for both case-study targets and both allocations. Cluster identities are local to each separate fit. Means are computed over all members; noise and smaller clusters are included in complete-ensemble analyses but omitted here for space.

Target	Allocation	Cluster	Models	Mean bottleneck DockQ	Mean ipTM
8TS0	BROAD	C1	8615	0.0531	0.7968
8TS0	BROAD	C2	1352	0.8891	0.8444
8TS0	DEEP	C1	6904	0.0531	—
8TS0	DEEP	C2	2989	0.9011	—
8QZ3	BROAD	C1	6161	0.0235	0.5235
8QZ3	BROAD	C2	1685	0.7765	0.5216
8QZ3	DEEP	C1	7827	0.0237	0.5251
8QZ3	DEEP	C2	1043	0.7820	0.5272

The dashes indicate that cluster ipTM is not reported in the supplied case-study confidence summary for that allocation, rather than a value of zero. The main figure retains the supplied panels: 8TS0 under BROAD and 8QZ3 under DEEP. The other rows document target dependence without selecting only the allocation that supports breadth. In 8QZ3, BROAD also has 1,914 unassigned models compared with 1,021 under DEEP. More scattered points or more clusters therefore cannot be equated with a larger useful population.

Clustering protocol and audits. The production protocol constructs all $10{,}000 \times 9{,}999/2 = 49{,}995{,}000$ unique pairwise antibody RMSDs per task in the antigen-fixed frame. DBSCAN uses that distance space, not t-SNE coordinates. The t-SNE display uses perplexity 40, 1,000 iterations, random initialization, the Barnes–Hut method, and random state 20260921. The RMSD matrix is checked for symmetry and consistency with previously computed per-model and global mean distances. In 8TS0/BROAD, the maximum discrepancy from the reference per-model mean is below 10^{-4} Å; the corresponding 8QZ3/DEEP discrepancy is also below 10^{-4} Å. These audits establish numerical agreement, not biological validity of the clusters. A 2 Å density threshold is an analysis choice, and cluster labels do not establish metastable physical states.

How to reproduce the reported summaries. The source bundle contains the unchanged figure PDFs, LaTeX sources, bibliography, `audit_results.py`, the portable `verify_bundled.py`, and a compact `evidence/` directory. The latter includes the selected-success roster and counts, target-level coverage and coordinate summaries, all-task cluster DockQ summaries, timing records, budget checkpoints, fixed-budget correlations, and a source-hash manifest. The tables are sufficient to recompute the numerical claims reported here without rerunning AF3. The complete-coordinate and per-interface sources identified by the analysis manifests are necessary to rerun all original metric computations. The summary bundle is not a substitute for those larger raw data.

Raw-generation reproducibility. Reproducing AF3 generation additionally requires its model parameters, installation, inputs, databases, and applicable access terms (Abramson et al., 2024). A complete run manifest should record the exact code revision, parameter release, database snapshot, MSA/template provenance, effective inference settings, accelerator type, batching, seed lists, and sample indexing. The present manuscript does not invent absent values or assume that current defaults equal the original run settings. Any future adaptive policy or learned reranker must be evaluated with a held-out design rather than tuned and tested on the same 113 complexes.

Scope of claims. All reported outcomes are computational structure metrics against benchmark references. No experimental binding assay is performed. Contact footprints are predicted residue coverage, model confidence is a learned score, and oracle success uses native knowledge. Keeping these quantities separate is necessary to interpret the practical value of test-time scaling.

E ADDITIONAL SELECTED-SUCCESS SCALING CURVES

Figures 6–10 complement the BROAD Top-1 result in Figure 2. All rates use the same 160 interfaces from 104 targets and the confidence-ranked Top-k rule in Equation 2; when fewer than k candidates are available, all available candidates are retained. Within each retained set, an interface succeeds if at least one model exceeds the indicated DockQ threshold. These are interface-level selected-set outcomes, distinct from the 113-target bottleneck-DockQ discovery curves in Figure 4.

The figures follow the typography, colors, nested columns, and numerical annotations of Figure 2. Each colored upper boundary is a cumulative success rate; the three rates must not be added. Large horizontal labels indicate seed counts; smaller labels immediately below individual bars indicate samples per seed. BROAD uses the four seed groups shown in Figure 2; DEEP shows all three seed budgets and five sample budgets. Vertical limits expand for Top-10 and Top-25 to accommodate their higher success rates. The complete counts, including the additional five-seed BROAD configurations, are available in `evidence/topk_all_configs.csv`.

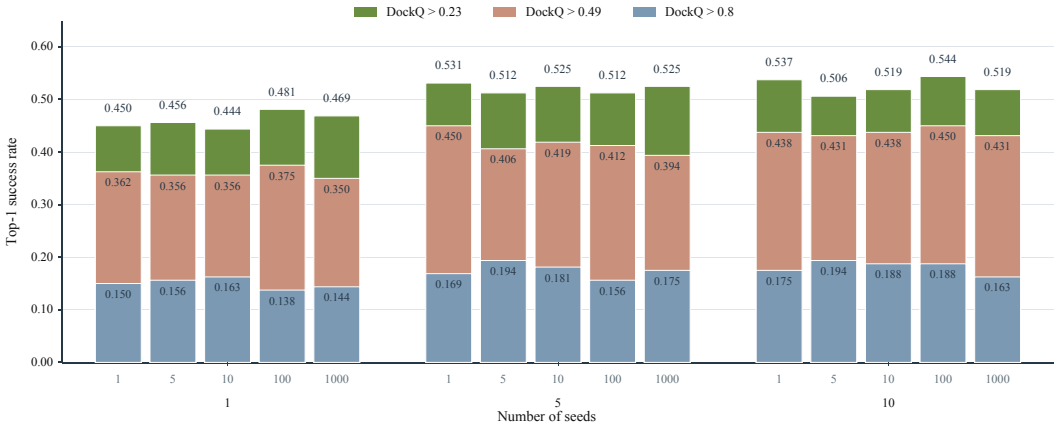


Figure 6: **DEEP Top-1 scaling.** Prefixes of the 10×1000 ensemble, with 1, 5, or 10 seeds and 1, 5, 10, 100, or 1,000 samples per seed. At full budget, success is $83/160$, $69/160$, and $26/160$ at DockQ > 0.23, > 0.49, and > 0.80, respectively.

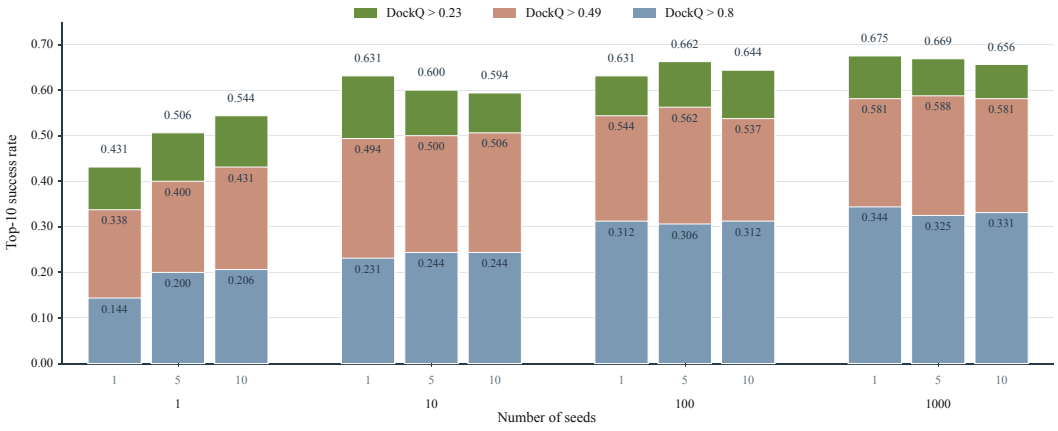


Figure 7: **BROAD Top-10 scaling.** Prefixes of the 1000×10 ensemble, with 1, 10, 100, or 1,000 seeds and 1, 5, or 10 samples per seed. At full budget, success is $105/160$, $93/160$, and $53/160$ at the three thresholds.

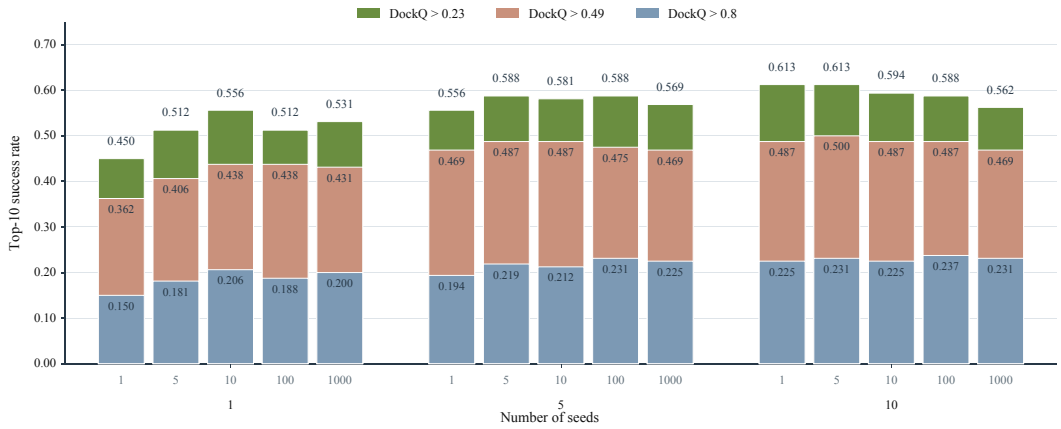


Figure 8: **DEEP Top-10 scaling.** The seed and sample grid is identical to Figure 6, using confidence-ranked Top-10 retention. At full budget, success is $90/160$, $75/160$, and $37/160$ at DockQ > 0.23, > 0.49, and > 0.80, respectively.

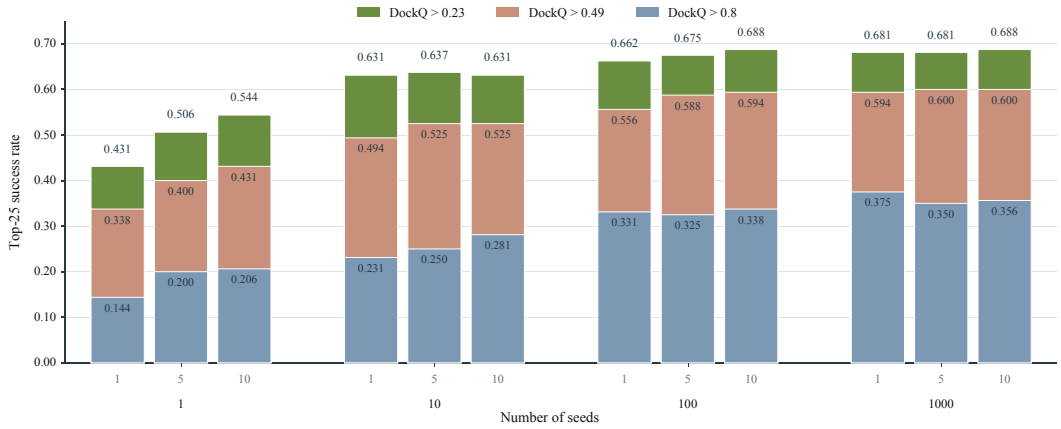


Figure 9: **BROAD Top-25 scaling.** The seed and sample grid is identical to Figure 7, using confidence-ranked Top-25 retention. At full budget, success is $110/160$, $96/160$, and $57/160$ at the three thresholds.

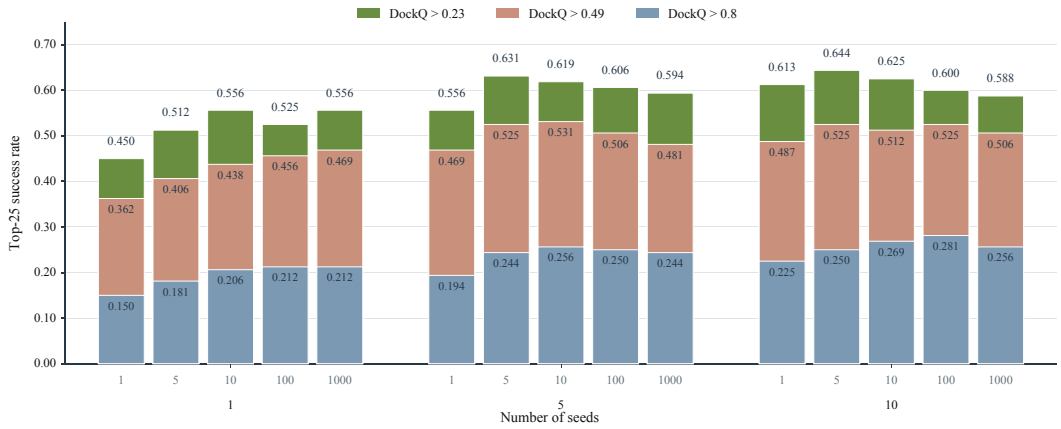


Figure 10: **DEEP Top-25 scaling.** The seed and sample grid is identical to Figure 6, using confidence-ranked Top-25 retention. At full budget, success is $94/160$, $81/160$, and $41/160$ at DockQ > 0.23, > 0.49, and > 0.80, respectively.