STRUCTURAL PLAUSIBILITY WITHOUT BINDING SPECIFICITY: LIMITS OF AI-BASED ANTIBODY-ANTIGEN STRUCTURE PREDICTION SCORES

Anonymous authors
Paper under double-blind review

ABSTRACT

Antibodies are central to modern immunotherapy, yet accurately predicting antibody-antigen binding interfaces remains a major challenge for computational modeling. While recent AI-based structure prediction methods can generate plausible antibody-antigen complexes, it remains unclear whether they can reliably discriminate cognate binding partners and identify correct paratope-epitope interfaces in realistic discovery settings. Here, we introduce a controlled benchmarking framework to systematically evaluate publicly available state-of-the-art structure prediction models (AlphaFold3, Boltz-2, and Chai-1) in their ability to distinguish real antibody-antigen complexes-cognate pairs, extracted from experimentally solved structures (n=106), from shuffled complexes, which serve as artificial non-cognate negative controls (n=11,342). We assessed structural accuracy, interface correctness, and the discriminative power of commonly used confidence metrics to distinguish cognate from non-cognate complexes (including ipTM and DockQ) when applied without access to structural ground truth, reflecting real-world deployment scenarios. We further analyzed sequence-level, structural, and sequence-structure features associated with high or low prediction confidence, independent of pairing correctness, and evaluated the trade-offs between computational cost and performance gains from increased sampling. To support community-driven mining, benchmarking, and method development, we release a large-scale dataset of 1.8 million (561,800 VHH-antigen pairings predicted by 3 different tools) complex structures. Our results show that current confidence scores (ipTM) often fail to discriminate cognate from non-cognate interactions (high false positive rate), even with extensive sampling, highlighting key limitations in current antibody-antigen modeling pipelines. This work provides a biologically grounded benchmark for antibody-antigen interface prediction and outlines critical directions for improving computational screening strategies in antibody discovery.

1 INTRODUCTION

Antibodies are key immunotherapeutic biomolecules characterized by antigen-specific binding. This specificity enables antibodies to identify and bind molecular targets such as tumor- or pathogen-associated antigens (Meng et al., 2024), underpinning a wide range of therapeutic applications and positioning antibodies as the largest class of biotherapeutics (Bielska et al., 2025). With the continued growth of monoclonal antibody (mAb) therapeutics, there is increasing interest in developing *in silico* antibody discovery and design tools (Akbar et al., 2022; Bashour et al., 2024; Greiff et al., 2020; Levine et al., 2026; Overath et al., 2026). While experimental discovery pipelines still rely heavily on antibody libraries and screening (Laustsen et al., 2021), improved prediction of paratope-epitope interactions may enable *in silico* discovery strategies (Akbar et al., 2022).

Antigen recognition is primarily mediated by the antibody's six hypervariable complementarity-determining region (CDR) loops, which together form the paratope (Ruffolo et al., 2023). These loops engage antigen surface regions (the epitope) through a combination of shape complementarity, physicochemical interactions, and conformational adaptability (Smorodina et al., 2025a; Kuroda & Gray, 2016). Although recent advances in deep learning have revolutionized protein structure pre-

diction for many molecular systems (Abramson et al., 2024; Passaro et al., 2025; Discovery et al., 2024), accurately predicting antibody-antigen complexes in general, and identifying the correct paratope-epitope interface in particular, remains challenging (Smorodina et al., 2022; Yin & Pierce, 2024). Previous benchmarking studies report success rates of approximately 20% for antibody-antigen docking using AlphaFold-Multimer (v2.3.0) and Rosetta-based protocols (Hitawala & Gray, 2024; Harmalkar et al., 2023; Weitzner et al., 2017; Ambrosetti et al., 2020). More recent evaluations indicate improved performance, with approximately 35% success using a single stochastic seed and up to around 60% success when extensive sampling (up to 1000 seeds) is combined with confidence-based ranking (Abramson et al., 2024; Hitawala & Gray, 2025), albeit at substantial computational cost.

Recently, several state-of-the-art molecular structure prediction models have emerged, including AlphaFold2/3 (Abramson et al., 2024; Jumper et al., 2021), Boltz-1/2 (Passaro et al., 2025; Wohlwend et al., 2024), and Chai-1/2 (Boitreaud et al., 2024; Team et al., 2025). Performance evaluations on diverse benchmarks spanning protein monomers, multimers, and small-molecule interactions have suggested that some of these models outperform earlier generations such as AlphaFold-Multimer (Fromm et al., 2025; Unsal et al., 2025; Grieswelle et al., 2025; Bradley, 2023; Evans et al., 2021). However, systematic evaluation of their ability to identify correct antibody-antigen interfaces-and, critically, their inability to reject incorrect interactomes-remains absent. Evaluating computational complex-prediction models can highlight key metrics that discriminate cognate binding partners ("real" complexes) from non-cognate ones ("shuffled" complexes used as negative controls) (Overath et al., 2025). This distinction is particularly important in discovery and screening contexts, where large numbers of candidate antibody-antigen pairings must be evaluated to separate binders from non-binders and incorrect-epitope binders, with false positives remaining a persistent challenge (Hitawala & Gray, 2025; Holt et al., 2025; Dang et al., 2023).

To explicitly address the problem of partner discrimination, we introduce a benchmarking framework that enables ground-truth distinction between real and shuffled antibody (here, VHH) and antigen complexes (Figure 1). Real complexes are defined as cognate VHH-antigen pairs extracted from experimentally solved structures, where the VHH and antigen are known biological binding partners within the same Protein Data Bank (PDB) entry. In contrast, shuffled non-cognate complexes are artificially generated by pairing a VHH sequence from one experimental complex with an antigen sequence from a different experimental complex. These non-cognate shuffled pairs do not correspond to known biological interactions and serve as ground-truth decoys. This benchmark design allows direct comparison of computational predictions across cognate and non-cognate pairings under identical modeling conditions.

Using this framework, we benchmark AlphaFold3, Boltz-2, and Chai-1 for predicting VHH-antigen paratope-epitope interactions, quantifying their ability to distinguish real cognate from shuffled non-cognate pairs. This enables assessment of structural and interface accuracy, evaluation of whether confidence metrics (e.g., ipTM (Evans et al., 2021; Genz et al., 2025), DockQ (Basu & Wallner, 2016; Mirabello & Wallner, 2024)) can discriminate real from shuffled complexes, and identification of sequence or structural features associated with high or low confidence independently of pairing correctness. Overall, we present a ground-truth-based benchmarking pipeline that exposes current tool limitations and highlights opportunities to improve antibody-antigen complex prediction for practical screening workflows.

2 Computational predictors largely fail to discriminate real from "shuffled" antibody-antigen complexes

To assess whether state-of-the-art structure prediction tools can distinguish biologically observed antibody-antigen interactions from shuffled complexes, we evaluated three independent AI-based complex structure prediction tools (AF3, Boltz-2, and Chai-1) across a panel of 561,800 ($106^2 \times 50$) VHH-antigen pairings (Figure 2): 106 real systems (complexes) from 91 unique PDB IDs and 11,342 shuffled complexes with 50 replicates each, where a replicate is defined as a prediction sample generated by a model. The selection of 106 real VHH-antigen complexes was based on sequence and structural diversity, size, and uniqueness (see Methods). For all complexes, information on whether they were present in the training set of the respective AI tool was also available.

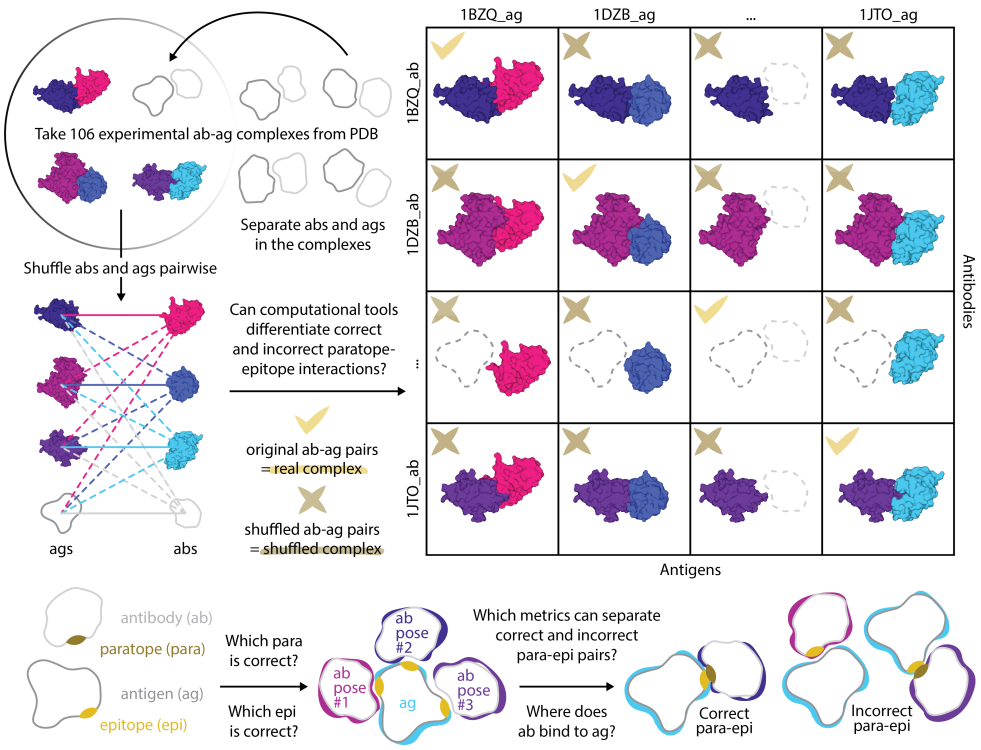


Figure 1: Paratope-epitope identification challenge. Schematic of the real-versus-shuffled antibody–antigen pairing strategy used to assess whether computational tools prioritize cognate paratope–epitope interactions. Starting from 106 experimentally resolved antibody–antigen complexes in the PDB, VHHs and antigens are separated and recombined pairwise. Native pairings define real complexes (Supplementary Table 1), while non-native combinations form shuffled (non-cognate) complexes. The central grid shows representative real (✓) and shuffled (×) pairings, illustrating correct and incorrect paratope–epitope matches. Antibodies are shown in purple/magenta and antigens in blue/cyan, with paratopes and epitopes highlighted.

Dataset splits were defined using the original train-test splits of each tool. Complexes were labeled as *train* when both the VHH and antigen were present as a cognate complex (real) in that model's training set (as of its training-set date cutoff); as *test* when both were present as a cognate complex (real) but not seen by the model; and as *both* when one binding partner was drawn from the training set and the other from the test set (shuffled). For this analysis, we used the following splits: Chai-1: 25 train and 82 test systems; Boltz-2: 64 train and 43 test systems; AF3: 30 train and 77 test systems.

First, we found that interface score landscapes (ipTM) largely lack specificity for real complexes (Figure 2A). All-VHH-vs-all-antigen heatmaps of predicted interface confidence scores revealed broadly similar interaction landscapes for real and shuffled complexes across all three tools. AF3 and Chai-1 produced sparse, heterogeneous score patterns with isolated high-scoring interactions distributed throughout the matrix, while Boltz-2 assigned uniformly high scores across most VHH-antigen combinations. While AF3 showed some enrichment of scores along the diagonal of the heatmap (real cognate pairs), none of the methods displayed systematic diagonal enrichment compared to the shuffled complexes, indicating a lack of specificity for biologically correct pairings.

To quantify enrichment on/off the diagonal in the all-vs-all comparison, we performed threshold-based classification (Figure 2B). Briefly, to test whether simple score thresholding could separate real from shuffled complexes, we evaluated multiple ipTM cutoffs (0.3, 0.6, and 0.8) chosen according to the scoring-metric literature (Genz et al., 2025; Weeratunga et al., 2024) and quantified precision, specificity, and F1 score. For AF3 and Chai-1, low thresholds resulted in high false-positive rates, while higher thresholds sharply increased false negatives, yielding poor overall performance ($F1 \leq 0.31$ for AF3 and $F1 \leq 0.13$ for Chai-1). Although Boltz-2 ($F1 = 0.02$) achieved high recall at low thresholds (≥ 0.9) due to uniformly elevated scores, this came at the cost of low specificity (≤ 0.18), and higher thresholds failed to balance precision and recall. Across all tools and thresholds, reliable discrimination between real and shuffled complexes was not achieved.

Next, we asked to what extent scores differ between pairs that belong to the train or test sets within each dataset split (Figure 2C). Violin plots summarizing ipTM score distributions confirmed substantial similarity of interface scores between real and shuffled complexes for all tools and across training, test, and mixed (*both*) splits. AF3 showed bimodal distributions for real complexes, with a subset of high-scoring interactions, but shuffled complexes still occupied overlapping score ranges. Boltz-2 assigned consistently high ipTM values to both real and shuffled interactions (mean $\sim$0.85-0.91), with minimal separation. Chai-1 yielded uniformly low scores for both classes, again with nearly identical distributions. Across all tools, mean scores and variances were comparable between real and shuffled complexes, indicating that score magnitude alone does not encode interaction authenticity.

We used TopModel's clash score (Fernández-Quintero et al., 2023) to evaluate structural quality across real and shuffled antibody-antigen complexes (Figure 2D). The clash score accounts for both the number of steric clashes between all atoms and protein length, with lower values indicating better geometric quality. Across all tools and dataset splits, clash score distributions were broad and overlapped between real and shuffled complexes. For AF3, top-ranked (highest ipTM) models showed mean clash scores of 25.7 ± 9.1 (train), 26.7 ± 9.8 (test), and 26.8 ± 9.7 (*both*) for shuffled complexes, compared with 22.7 ± 9.4, 24.2 ± 9.9, and 26.8 ± 9.7 for real complexes across the same splits. Boltz-2 produced systematically higher clash scores than AF3, with shuffled complexes averaging 30.0 ± 9.5 (train), 36.1 ± 12.4 (test), and 33.0 ± 11.3 (*both*), while real complexes averaged 27.9 ± 10.7, 34.8 ± 12.2, and 32.9 ± 11.3, respectively. Chai-1 exhibited markedly poorer structural quality by this metric: mean clash scores were 4-5$\times$ higher and more variable, reaching $\sim$80-90 on average, with extreme dispersion (e.g., 62.0 ± 65.0 and 90.7 ± 125.2 in the test split). These elevated and highly variable scores were observed for both real and shuffled complexes, indicating frequent steric clashes in top-ranked predictions and a lack of effective internal filtering for geometric plausibility. Importantly, real complexes did not systematically achieve lower clash scores than shuffled complexes in any tool or dataset split. In several cases, shuffled complexes displayed comparable or even slightly lower average clash scores than real complexes, despite being biologically incorrect.

To assess whether different prediction tools agree on which antibody-antigen pairs are confident or uncertain, we directly compared ipTM score matrices generated by AF3, Boltz-2, and Chai-1 (Figure 2A). Global correlations between flattened score matrices were uniformly low (Pearson $r = 0.14$ for Chai-1 vs. Boltz-2, $r = 0.18$ for Chai-1 vs. AF3, and $r = 0.13$ for Boltz-2 vs. AF3), indicating low agreement across tools in their assessment of interaction confidence.

Per-system correlation analysis (testing whether tools rank the same systems as similarly confident or unconfident) further revealed substantial heterogeneity across both VHHs and antigens (Figure 2E). For each pair of tools, many systems exhibited weak or even negative correlations, highlighting disagreement in how individual VHHs or antigens are scored. These discrepancies were not confined to a small subset of problematic systems but were broadly distributed across the benchmark, suggesting that tool-specific inductive biases strongly shape confidence assignment. Outlier analysis reinforces this conclusion: high-confidence shuffled outliers identified by one tool rarely overlapped with those from another. Specifically, only four shuffled outliers overlapped between Chai-1 and Boltz-2, fourteen between Chai-1 and AF3, and just one between Boltz-2 and AF3, despite all tools being evaluated on the same set of VHH-antigen pairings. Representative examples illustrate the magnitude of these discrepancies. For instance, the system (4NC2, 7R24) was assigned high confidence by Boltz-2 (ipTM≈ 0.81) but substantially lower confidence by Chai-1 (ipTM≈ 0.60), while (9EMY, 8UKV) achieved high confidence in Chai-1 (ipTM≈ 0.73) but was scored markedly lower by Boltz-2 (ipTM≈ 0.44). Similarly, several systems such as (7TGF, 6OBO) and (8K4Q, 8EW6) were consistently high-confidence outliers in AF3 (ipTM≥ 0.8) yet did not emerge as outliers in the other tools (Supplementary Table 2).

Together, these results show that both ipTM confidence and structural accuracy remain limited across tools, with AF3 producing the lowest average clash scores, Boltz-2 intermediate values, and Chai-1 the poorest overall structural quality. However, even the best-performing method by this metric fails to reliably discriminate real antibody-antigen complexes from shuffled mismatches. This inability to distinguish real from shuffled complexes persists across training, test, and combined datasets, indicating that the observed overlap is unlikely to result from overfitting or data leakage. Instead, it reflects a limitation of current structure-based confidence scores and simple structural quality measures, which capture generic interface properties but fail to reflect biological correctness in antibody-antigen complex prediction.

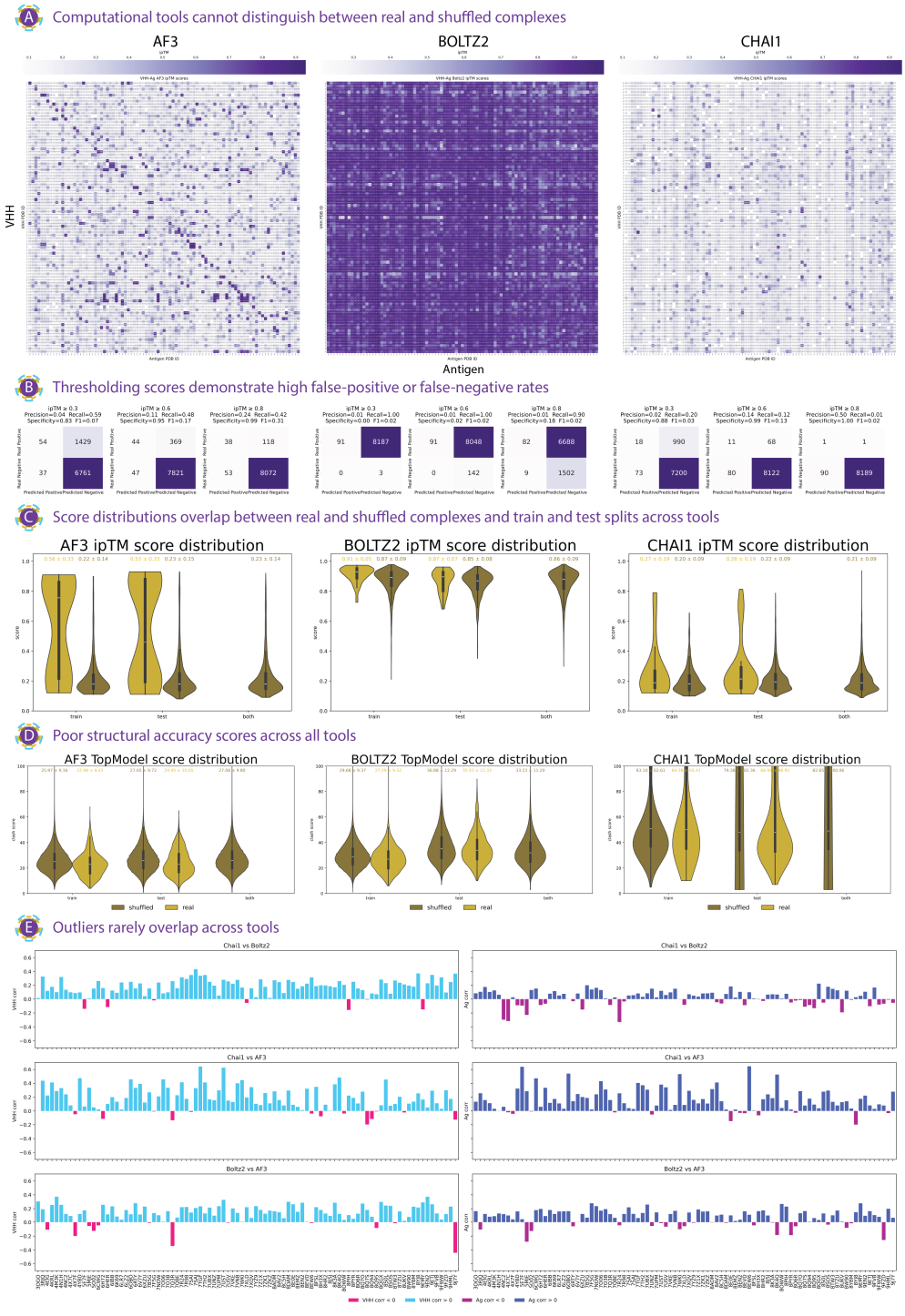


Figure 2: Current structure prediction tools and confidence scores fail to capture antibody-antigen specificity. **(A)** All-by-all interaction score landscapes. Heatmaps show the best interface confidence scores (ipTM) across 50 stochastic predictions for all VHH-antigen pairings generated by AF3, Boltz-2, and Chai-1. Rows denote VHHs and columns antigens; color scales indicate ipTM values. **(B)** Overlapping confidence distributions for real and shuffled complexes. Violin plots summarize ipTM scores across training, test, and mixed (*both*) splits. Real and shuffled distributions largely overlap for all models. **(C)** Threshold-based classification is unreliable. Confusion matrices report precision, specificity, and F1 scores at ipTM thresholds of 0.3, 0.6, and 0.8 for each tool. **(D)** Structural quality metrics are non-discriminative. Violin plots show clash score distributions for top-ranked predictions across dataset splits; real and shuffled complexes strongly overlap for all tools. **(E)** Tool-specific confidence outliers. Per-system correlations of ipTM scores between tool pairs reveal weak agreement (Pearson correlation) and inconsistent high-confidence outliers across VHHs and antigens.

A Appendix

A.1 Cross-tool score agreement is weak and sampling improves structure but not confidence alignment

We next asked whether models "know" when they fail-that is, whether confidence scores (ipTM) reliably indicate structural prediction quality (DockQ) under different sampling regimes (Figure 3). Confidence calibration-the correspondence between predicted confidence and actual quality (Guo et al., 2017; Kull et al., 2019)-is critical for practical screening workflows where ground-truth structural information is unavailable. Accordingly, confidence calibration analysis was performed only on real complexes, as shuffled pairings lack experimental reference structures against which DockQ can be computed (Mirabello & Wallner, 2024). We found that confidence calibration differs systematically across tools when examining the correlation between DockQ and ipTM for the best, initial (sample$_0$), and worst of 50 predictions (Figure 3A).

To interpret calibration patterns, we defined four diagnostic quadrants based on thresholds of DockQ$= 0.23$ (acceptable quality) and ipTM$= 0.5$ (confident prediction): Q1 (confident and correct), Q2 (overconfident failures), Q3 (uncertain and incorrect), and Q4 (underconfident success) (Supplementary Table 3). Among the three methods, AF3 exhibits the strongest calibration, with a Pearson correlation of $r = 0.736$ for best-DockQ structures and relatively few overconfident failures (Q2: 2%; Q4: 28%). AF3 maintained high correlation even at the single-sample level ($r = 0.888$ for sample$_0$), indicating comparatively robust confidence assignment without extensive sampling. In contrast, Boltz-2 was systematically overconfident. Although its best-DockQ calibration remained moderate (Pearson $r = 0.665$), a substantial fraction of predictions fell into the overconfident failure regime (Q2: 18%), where ipTM is high despite poor structural quality. This effect was exacerbated for single-sample predictions, where correlation dropped sharply (Pearson $r = 0.493$), highlighting the unreliability of Boltz-2 confidence scores in low-sampling screening workflows. Conversely, Chai-1 exhibited systematic underconfidence. Although a subset of predictions achieved high structural quality (DockQ$\approx$ 0.4-0.9), these models were frequently assigned low ipTM scores, resulting in a substantial fraction of underconfident successes (Q4: 24%). Consequently, the relationship between confidence and best-case structural quality was weaker for Chai-1 (best-DockQ vs. ipTM, $r = 0.612$), implying that confidence-based filtering would exclude many structurally accurate predictions. Extreme quadrant cases further illustrate observed failure modes. Overconfident failures (Q2) correspond to geometrically poor structures assigned high confidence (e.g., AF3 system_24_24_6V7Y), whereas underconfident successes (Q4) correspond to high-quality complexes that would be missed by thresholding (e.g., AF3 system_27_27_7B5G). These cases provide concrete examples of how confidence scores can diverge from structural reality in both directions (see Supplementary Table 3 for the classification of all systems across the four quadrants).

Next, we examined whether sampling improves structure but not confidence alignment. Across all three tools, saturation sampling improved best-case DockQ (Supplementary Figure 1A), confirming that additional sampling can refine geometry. DockQ improvement distributions were right-skewed for all models, with AF3 and Boltz-2 exhibiting the largest absolute gains, while Chai-1 showed smaller but consistent improvements. The proportion of systems achieving acceptable or higher quality (DockQ> 0.23) increased substantially after saturation sampling (Figure 3B; Supplementary Figure 1B). For AF3, incorrect predictions decreased by nearly 20%, with Boltz-2 showing a comparable reduction and a marked increase in medium- and high-quality outcomes; in contrast, Chai-1 remained dominated by incorrect predictions despite sampling. These changes reflect systems "rescued" above the quality threshold through sampling alone (the "improvers," Figure 3B).

However, the relationship between structural improvement and confidence remained weak. Although all three models showed significant improvements in DockQ scores (Wilcoxon $p < 0.001$; Supplementary Figure 1C), paired analyses showed no significant improvement in ipTM for AF3 or Chai-1 ($p = 0.179$ and $p = 0.222$, respectively). While Boltz-2 exhibited a statistically significant change ($p < 0.001$), the ipTMs actually slightly decreased (Supplementary Figure 1D). In particular, correlations between changes in DockQ and changes in ipTM were near zero across all models ($r = -0.027$, -0.040, and -0.019 for AF3, Boltz-2, and Chai-1, respectively; Figure 3C), indicating that confidence scores were largely insensitive to structural improvements achieved through sampling and remained "locked in" to the initial structural hypothesis.

To assess whether confidence scores provide consistent rankings across models, we compared ipTM values for each system between model pairs (Supplementary Figure 1E). Cross-model correlations were modest (Pearson $r = 0.29$-0.39), and importantly, higher confidence in one model did not reliably correspond to higher structural quality as measured by DockQ. From a screening perspective, this behavior has important implications. Single-sample predictions from Boltz-2 are frequently overconfident relative to achieved DockQ, whereas AF3 displays comparatively better calibration even prior to saturation sampling. Nonetheless, for all tools, confidence does not reliably track structural refinement, limiting its utility as a post-sampling ranking criterion.

Finally, we examined prediction consistency using epitope-paratope variation across 15 antigens, each crystallized with two distinct VHHs (Supplementary Figure 2A; Methods). This provides a natural test of whether models that succeed with one VHH can generalize to alternative binders of the same antigen. AF3 produced acceptable predictions (DockQ> 0.23 for both pairs) in 10 out of 15 cases (67%), compared to 6/15 (40%) for Boltz-2 and 2/15 (13%) for Chai-1 (Supplementary Figure 2B). Boltz-2 and especially Chai-1 more frequently succeeded with only one VHH or failed on both partners (Supplementary Figure 2A), suggesting that AF3 better captures the underlying antigen structure independently of the specific paratope geometry.

Together, these results demonstrate that confident failures are largely tool-specific rather than reflecting a shared set of universally difficult or ambiguous systems. The low overlap in outlier systems and weak agreement in confidence-based rankings across predictors highlight the absence of a common confidence landscape, underscoring the risk of relying on any single model's confidence scores for candidate selection-particularly after saturation sampling, where structural quality improves but confidence remains misaligned.

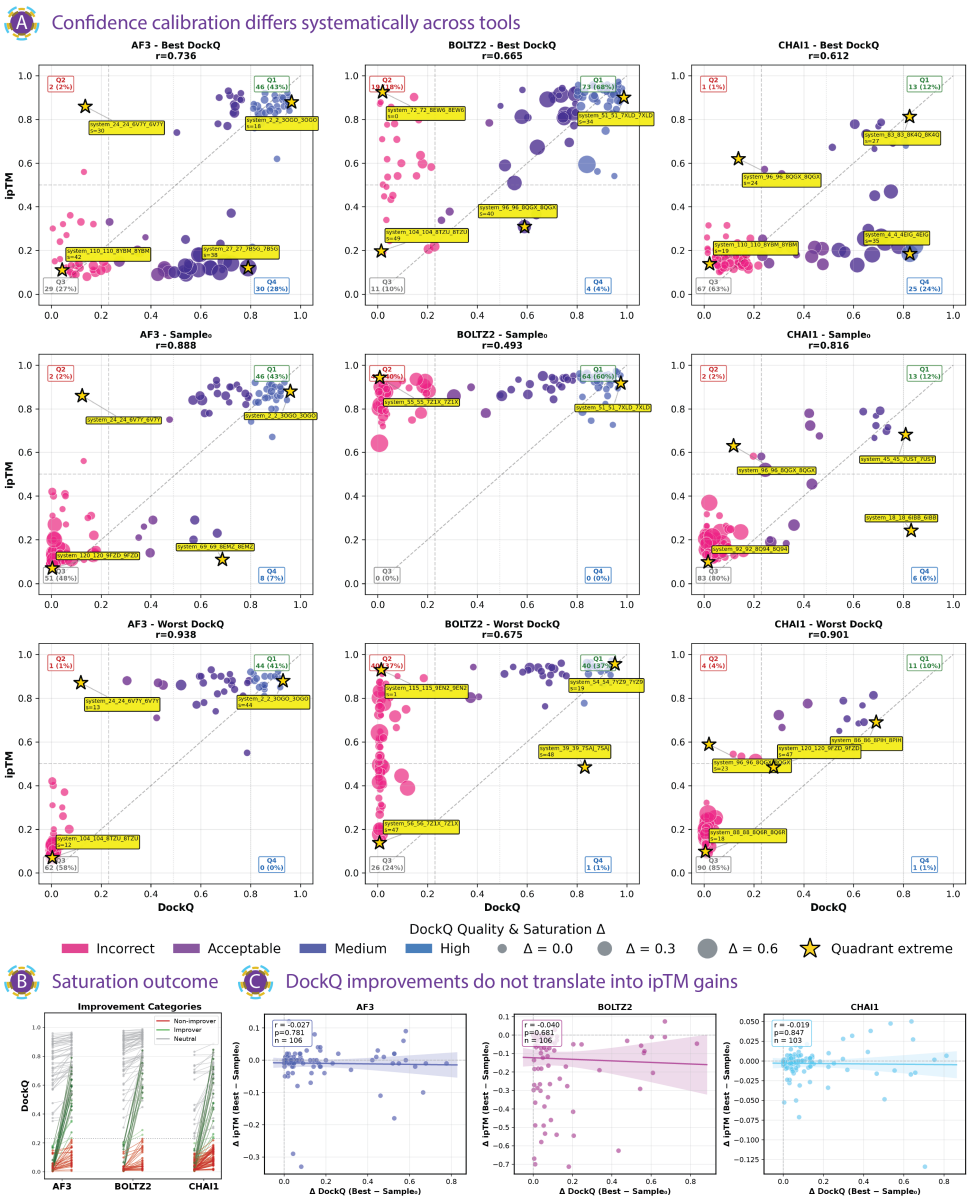


Figure 3: Cross-tool disagreement and confidence calibration of antibody-antigen docking predictions. **(A)** Confidence calibration differs systematically across ML structure prediction tools. Scatter plots show DockQ versus ipTM for AF3, Boltz-2, and Chai-1 under three sampling regimes: best DockQ, initial sample (sample$_0$), and worst DockQ. Points are colored by DockQ quality category (incorrect, acceptable, medium, high) and sized by saturation range (ΔDockQ from sample$_0$ to best). Dashed lines indicate DockQ$=0.23$ and ipTM$=0.5$. Pearson correlation coefficients are shown for each panel. **(B)** Saturation outcome categories reveal when sampling rescues docking quality. Line plots show per-system trajectories from sample$_0$ to the best sampled prediction, grouped into non-improvers, improvers, and neutral systems. **(C)** DockQ gains from sampling do not translate into confidence-score gains. Scatter plots show the relationship between ΔDockQ and ΔipTM for each model, with linear regression fits and 95% confidence intervals. Near-zero correlations indicate that confidence scores fail to track structural improvements achieved through sampling.

A.2 Structural accuracy scores are uniformly poor while interface contacts does not reflect specificity

We next evaluated whether predicted antibody-antigen complexes recovered the experimentally defined antigen epitope and whether this information can discriminate real from shuffled pairings (Fig-

ure 4). Epitope recall was quantified as the fraction of experimental epitope residues recovered by model predictions, where a residue was considered recovered if it contacted the VHH in at least 50% of stochastic replicate predictions ($n = 50$). Across all three models, epitope recall values were generally low and sparsely distributed, with substantial overlap between real and shuffled complexes (Figure 4A). Although real complexes lie along the diagonal of the epitope-recall matrices, off-diagonal (shuffled) complexes frequently exhibited comparable levels of epitope recovery. This indicates that models often identify plausible antigen contact regions even in non-cognate pairings, limiting the specificity of epitope-based discrimination. Consistent with this observation, direct comparison of epitope-recall distributions shows that shuffled complexes frequently overlap or exceed the recall observed for real complexes across all models and data splits (Figure 4B). Focusing on cognate pairings and whether the respective models recovered at least one residue from the true epitope (epitope recall > 0), AF3 recovered the most true epitopes in the test set of complexes ($n = 29$), followed by Chai-1 ($n = 26$) and Boltz-2 ($n = 17$).

To assess the impact of data leakage, we examined epitope-recall distributions for real complexes split into train-leakage and true test sets (Figure 4B). Across models, train-leakage complexes showed slightly higher median epitope recall than true test complexes, consistent with partial memorization or bias toward known interfaces. However, the overall distributions remained broad, and substantial overlap persisted between train-leakage and true test predictions. Notably, shuffled complexes spanned similar recall ranges, indicating that leakage alone does not explain the limited discriminatory power of epitope recall. Thus, high epitope recall does not reliably indicate cognate binding even in the absence of explicit train-test overlap.

We next tested whether increasing stochastic sampling improves epitope discrimination by assessing epitope-enrichment significance as a function of replicate count (Figure 4C). Enrichment significance was computed using a one-sided binomial test for each VHH-antigen pair. Let N denote the number of stochastic structure predictions and K the number of predictions containing at least one epitope contact. Under a null model in which contacts occur uniformly across the antigen surface with probability p_{epitope} (estimated from the fraction of solvent-accessible antigen residues belonging to the epitope), the probability of observing K or more epitope hits is given by the binomial survival function $P(X \geq K \mid X \sim \text{Binomial}(N, p_{\text{epitope}}))$. Resulting p-values were corrected across all VHH-antigen pairs using the Benjamini-Hochberg procedure, and pairs with FDR < 0.05 were considered significantly enriched. To assess the effect of ensemble size, we recomputed binomial p-values for hypothetical replicate counts m by rescaling the observed hit fraction (K/N) and repeating the same multiple-testing correction. For all three models, the fraction of VHH-antigen pairs declared significantly enriched increased rapidly with the number of stochastic predictions and saturated by approximately 10-15 replicates. Importantly, real and shuffled complexes exhibited highly similar saturation behavior, regardless of whether the shuffled pairs involved train leakage of the VHH, the antigen, both, or neither. This indicates that increasing ensemble size primarily reinforces existing contact preferences rather than sharpening discrimination between cognate and non-cognate interactions. AF3 and Boltz-2 retained a modest residual separation between real and shuffled complexes across replicate counts, with real pairs showing slightly higher significant-enrichment rates than any shuffled cohort. In contrast, Chai-1 showed near-complete convergence: shuffled complexes rapidly achieved significance frequencies comparable to, or exceeding, those of real complexes, consistent with a strong epitope-seeking bias that limits the interpretability of enrichment-based metrics for this model. Notably, the leakage-stratified shuffled cohorts largely overlapped within each model, further indicating that these effects are not primarily driven by explicit train-test leakage but instead reflect intrinsic model priors amplified by ensembling.

Finally, we examined the relationship between structural consistency across predictions and epitope recovery within each model (Figure 4D). Structural consistency was quantified as the mean pairwise RMSD across $n = 50$ stochastic replicates for each complex. Across all models and data splits, epitope recall was negatively correlated with replicate RMSD, indicating that complexes predicted with consistent orientations tend to recover a larger fraction of epitope residues. However, this relationship held for both train-leakage and true test complexes and did not differentiate real from shuffled pairings. Pearson correlation coefficients ranged from $r = -0.22$ to -0.30 for AF3, $r = -0.14$ to -0.27 for Boltz-2, and $r = -0.39$ to -0.59 for Chai-1, demonstrating that while structural consistency improves epitope recovery, it does not guarantee biological correctness. Notably, RMSD distributions of the 50 replicates showed slightly higher values for shuffled complexes compared to

real complexes (Figure 4E), indicating a potential-but limited-signal for distinguishing cognate from non-cognate interfaces.

Together, these analyses show that epitope recovery and enrichment metrics are dominated by generic contact biases and structural consistency rather than interaction specificity. Increasing stochastic sampling improves apparent epitope enrichment but does so similarly for real and shuffled complexes, limiting its utility as a discriminative signal. These findings further reinforce the conclusion that current structure prediction tools preferentially identify plausible antigen contact regions but struggle to distinguish true cognate paratope-epitope interactions from non-cognate alternatives.

A.3 Computational cost and sampling efficiency differ substantially across structure prediction tools

Using ML structure prediction tools for high-throughput antibody screening requires understanding not only prediction accuracy but also computational cost. With GPU resources unevenly distributed across institutions (ref, 2025) and growing concern over the energy footprint of large-scale computation (Morand et al., 2024; Lannelongue & Inouye, 2023), identifying cost-effective sampling strategies has practical importance. Saturation sampling and seed selection can both improve prediction quality (Abramson et al., 2024; Hitawala & Gray, 2025; Fromm et al., 2025; Johansson-Åkhe & Wallner, 2022), yet their relative contributions and associated costs remain uncharacterized. We therefore evaluated each system ($n = 106$, 91 unique PDBs; no deposition-year bias observed; Supplementary Figure 4A) across five random independent seeds with increasing numbers of diffusion samples ($N = 1, 10, 25, 50,$ and 100), monitoring GPU energy consumption throughout (Figure 5). For this analysis, we additionally included Boltz-1, the predecessor to Boltz-2, to check consistency of saturation effects across model versions. Because each saturation level uses an independent seed, this design enables analysis of combined seed and saturation effects; extracting only the first sample from each seed isolates seed-dependent variation.

To characterize how computational cost scales with sampling depth, we computed the median energy consumption across all systems at each saturation level ($N = 1, 10, 25, 50,$ and 100). Energy consumption varied markedly between tools, revealing distinct cost profiles. At the highest saturation level ($N = 100$ samples), median energy usage per system ranged from 23.0 Wh for AF3 to 82.9 Wh for Chai-1, corresponding to a 3.6-fold difference (Figure 5A; Supplementary Figure 3A-C). Boltz-1 and Boltz-2 exhibited intermediate energy-usage profiles, with substantially lower cost than Chai-1 at high saturation but higher baseline costs than AF3. When aggregated across all 106 real systems and saturation levels, total energy consumption ranged from 13.6 kWh for AF3 to 22.7 kWh for Chai-1.

These differences reflect distinct architectural choices. AF3 incurred a high baseline energy cost for MSA computation (82.4 Wh per system, computed once regardless of sampling depth) but showed relatively low marginal cost for additional diffusion samples. In contrast, Chai-1 exhibited a low baseline cost (6.6 Wh) but a steep increase in energy consumption with increasing saturation, consistent with its embedding-based workflow that omits MSA computation (Methods). Boltz-based models showed intermediate behavior. Across all tools, both energy consumption and runtime scaled approximately linearly with system length, with Pearson correlation coefficients ranging from $r = 0.60$ for AF3 to $r = 0.98$ for Boltz-1 (Supplementary Figure 3D,E).

Increasing saturation sampling improved prediction quality for all models with diminishing returns, as measured by DockQ (Figure 5B). We quantified this by identifying, for each system, the maximum DockQ achieved at each saturation level and computing the median across systems. Median maximum DockQ increased from baseline ($N = 1$) to $N = 100$ by $\Delta = +0.43$ for AF3 (0.24 to 0.68), $\Delta = +0.23$ for Boltz-2 (0.57 to 0.80), $\Delta = +0.20$ for Boltz-1 (0.06 to 0.26), and $\Delta = +0.15$ for Chai-1 (0.04 to 0.19). While all models benefited from increased sampling, the magnitude of improvement differed substantially, with AF3 showing the largest gains.

Marginal quality gain was computed as the per-system difference in maximum DockQ between consecutive saturation levels and summarized as the median across systems. Across models, the largest marginal improvements occurred at low saturation levels. The transition from baseline to $N = 10$ samples captured the majority of achievable quality improvement, while additional sampling beyond $N = 25$ yielded progressively smaller gains (Figure 5C). Consistent with this pattern, the propor-

tion of incorrect predictions (DockQ< 0.23) decreased sharply from baseline to $N = 10$ but showed more modest ($\sim$5-15%) reductions at higher saturation levels ($N = 25$-100; Figure 5D).

Analyzing cumulative DockQ improvement as a function of cumulative energy expenditure revealed a consistent "efficiency frontier" across tools (Figure 5C). All models exhibited steep initial slopes, indicating high quality gains per unit energy at low sampling depths, followed by pronounced flattening as energy expenditure increased. Despite differing absolute costs, the efficiency frontiers converge across tools, indicating that early sampling dominates quality gains regardless of architecture. These trends indicate that $N \approx 10$-25 samples capture most attainable quality improvement at a fraction of the computational cost required for $N = 100$.

Beyond sampling depth, seed choice had a substantial independent effect on prediction outcomes (Figure 5E). To isolate seed effects from saturation, we extracted only the first sample (sample$_0$) from each seed and computed the per-system DockQ range (maximum minus minimum across seeds). The median range was 0.04-0.05, though extreme cases spanned nearly the full quality spectrum (e.g., system_70_70_8EN2: DockQ 0.01-0.93 across five AF3 seeds). Across models, approximately 10-15% of systems were strongly seed-sensitive and benefited disproportionately from deeper saturation sampling (Supplementary Figure 4D; Supplementary Figure 5A).

Cross-seed trajectory analysis further showed that deeper sampling can partially rescue poor initial seed choices. For each seed, we tracked cumulative maximum DockQ (the best quality achieved up to each sample number) and computed the median across systems (Figure 5F). Seeds producing high-quality structures early tended to maintain their advantage throughout saturation, whereas seeds that began in low-quality regions could still achieve substantial improvements with deeper sampling, though typically plateauing at lower absolute quality than favorable seeds (Figure 5F; Supplementary Figure 5B). Together, these results indicate that seed selection and saturation sampling act as orthogonal optimization mechanisms: seed choice determines which region of the solution landscape is explored, while saturation sampling refines predictions within that region and can partially mitigate but rarely fully overcome suboptimal initial trajectories (approximately 85-90% of systems remained below their best-seed ceiling).

Finally, we assessed whether confidence metrics reflect structural improvements achieved through saturation sampling. Across all models, changes in ipTM from baseline to $N = 100$ showed near-zero correlation with corresponding changes in DockQ (Pearson $r = 0.00$-0.16; Supplementary Figure 4B,C). This indicates that confidence scores are largely determined by properties of the initial prediction and remain insensitive to substantial quality gains obtained through additional sampling. Reliable confidence tracking would enable early stopping without ground-truth structures, but the observed ipTM-DockQ decoupling precludes such use. This behavior mirrors the DockQ-ipTM misalignment observed above, reinforcing that current confidence metrics do not reflect convergence toward higher-quality docking solutions.

Together, these results demonstrate that moderate sampling ($N \approx 10$-25) captures most achievable quality gains at a fraction of full saturation cost, seed selection and saturation act as orthogonal optimization mechanisms, and current model confidence metrics fail to track these improvements.

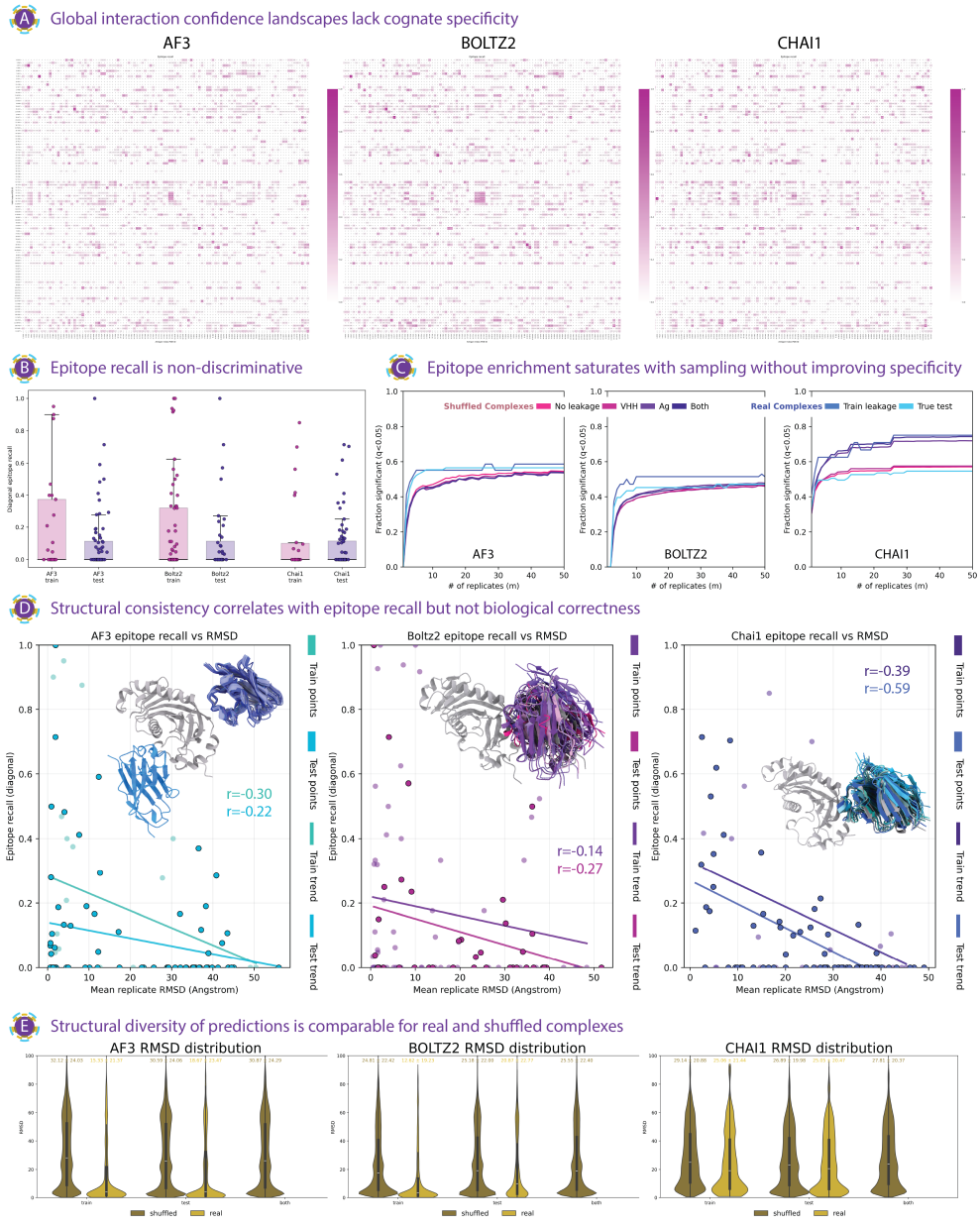


Figure 4: Epitope recovery and enrichment fail to distinguish cognate from non-cognate antibody-antigen interactions. **(A)** Epitope recall across real and shuffled complexes. Heatmaps show epitope recall for predicted antibody-antigen complexes, defined as the fraction of experimentally determined epitope residues recovered by the model. An epitope residue is considered recovered if it contacts the VHH in at least 50% of the $n = 50$ stochastic replicate predictions. Diagonal entries correspond to real (cognate) pairs, while off-diagonal entries represent shuffled (non-cognate) complexes. **(B)** Distribution of epitope recall across real and shuffled complexes. Boxplots summarize epitope recall values stratified by train-leakage and true test sets. Shuffled complexes frequently overlap or exceed recall observed for real complexes. Numbers of systems with epitope recall > 0 are indicated for each model and split. **(C)** Epitope-enrichment significance saturates rapidly with stochastic sampling. Line plots show the fraction of VHH-antigen pairs declared significantly enriched (FDR < 0.05) as a function of the number of stochastic predictions. Real and shuffled complexes exhibit similar saturation behavior across models. **(D)** Structural consistency correlates with epitope recovery but not interaction correctness. Scatter plots show epitope recall versus mean pairwise RMSD across $n = 50$ replicates, with linear fits shown separately for train-leakage and true test sets. Pearson correlation coefficients are reported. **(E)** Structural diversity of predicted complexes. Violin plots show RMSD distributions for real, shuffled, and mixed ("both") complexes across train-leakage and true test sets. Similar distributions indicate that structural consistency alone is insufficient to discriminate cognate from non-cognate interactions.

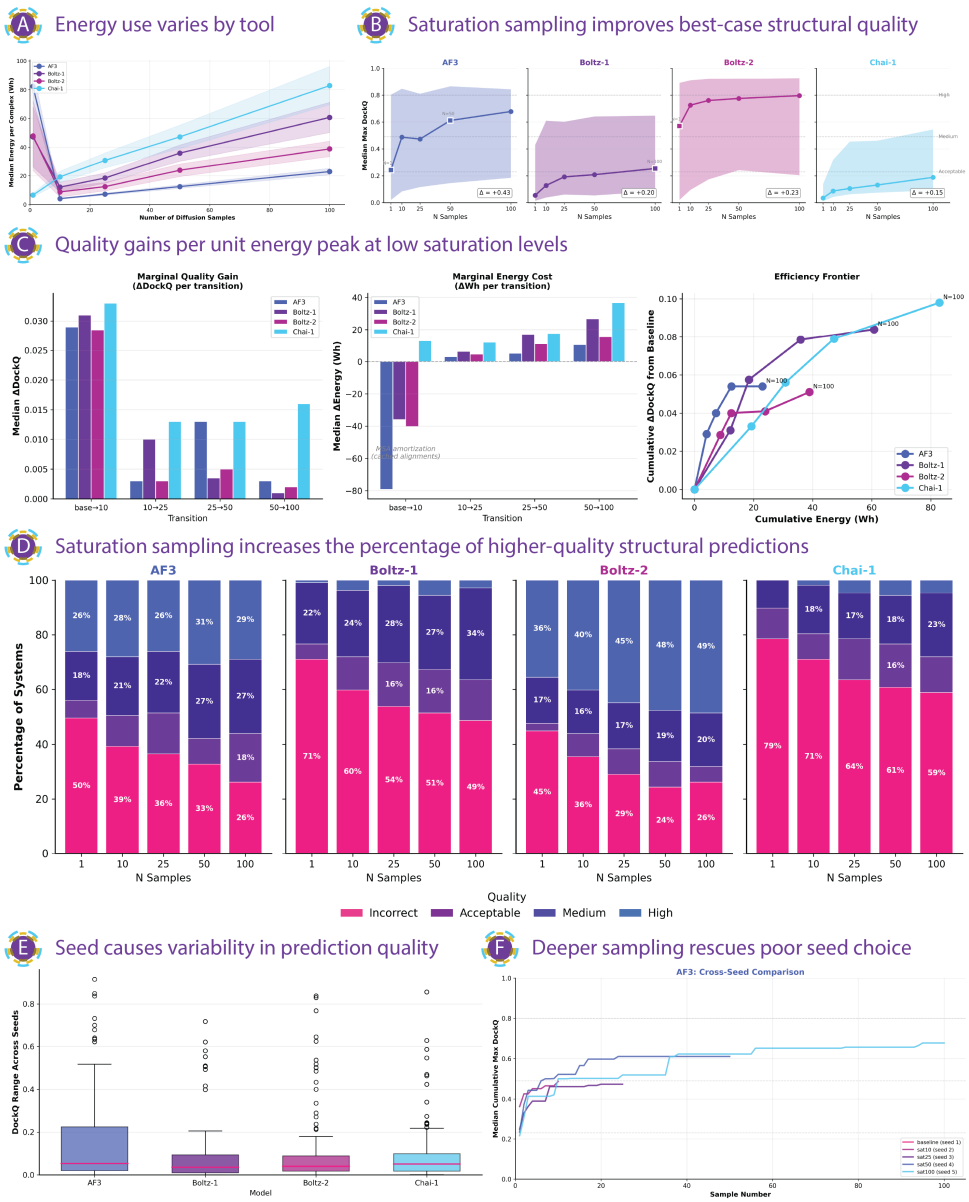
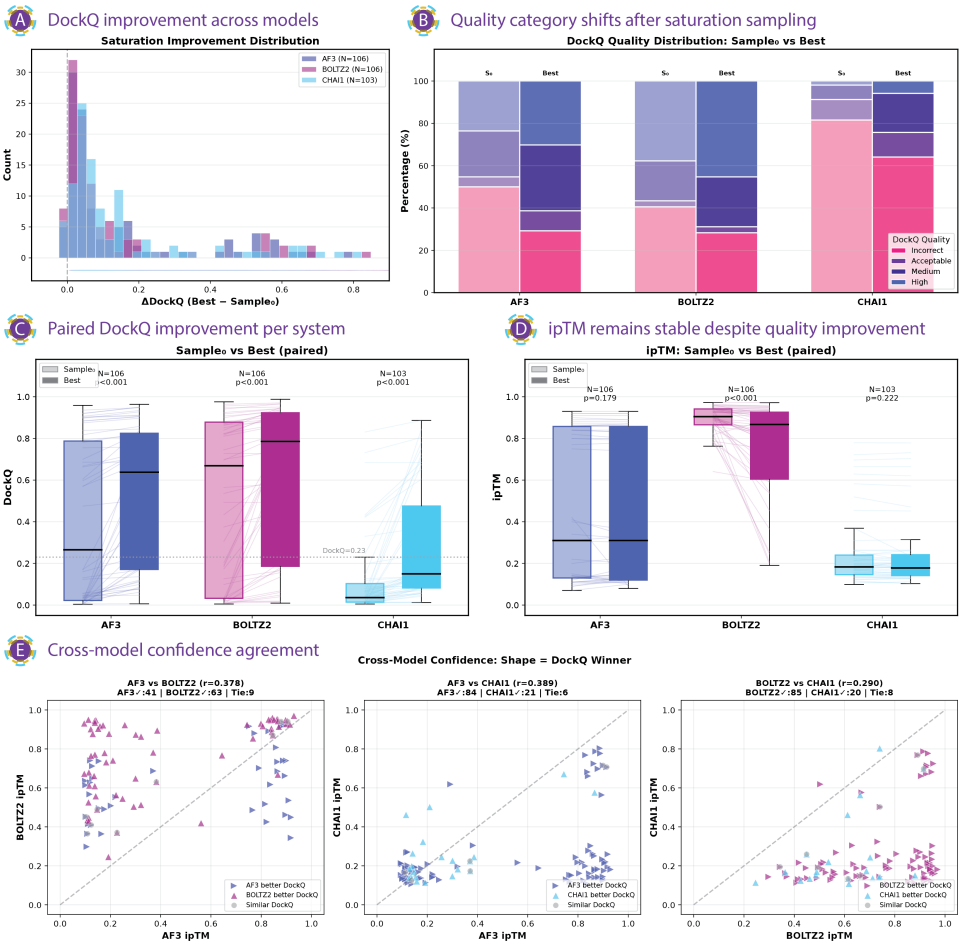


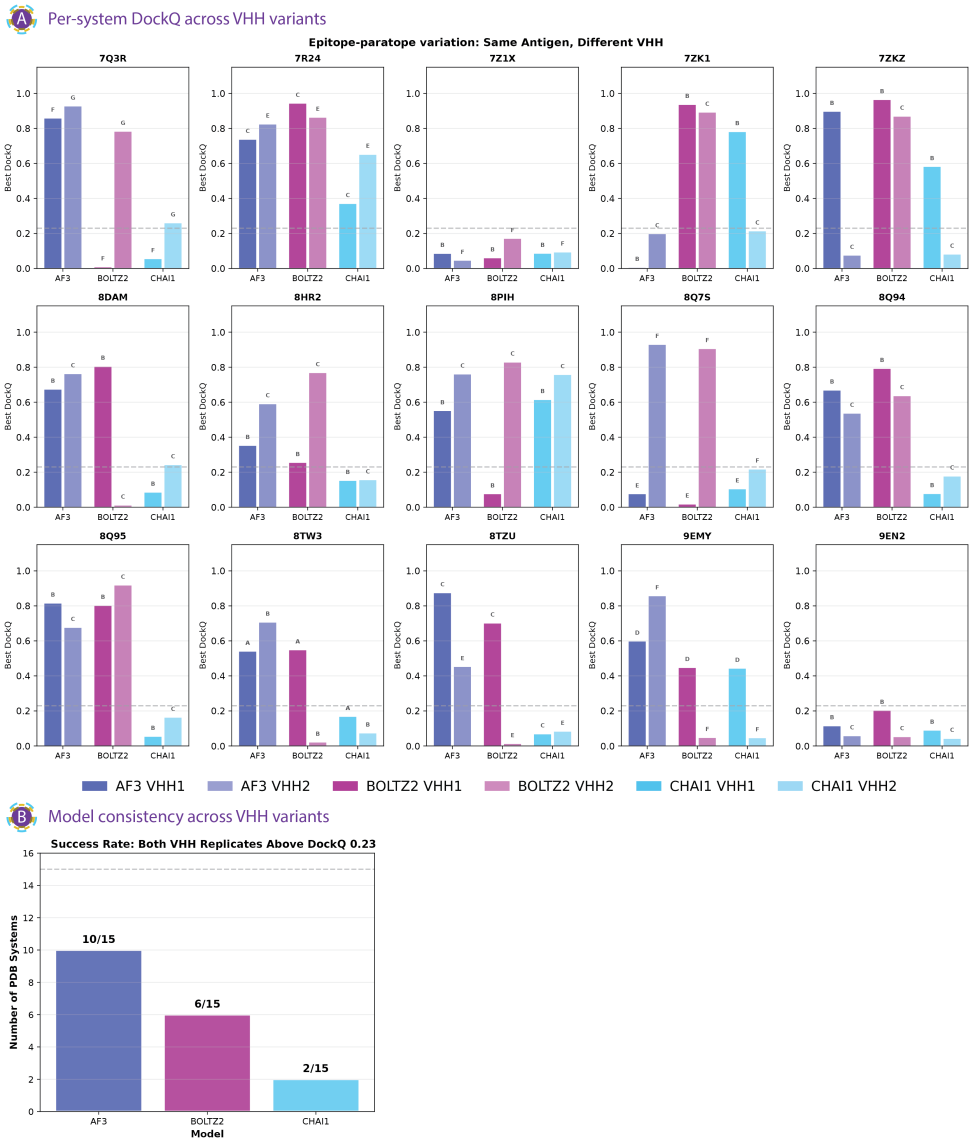
Figure 5: Computational cost, sampling efficiency, and seed-dependent optimization in antibody-antigen structure prediction. **(A)** Energy consumption differs substantially across tools and sampling depth. Median energy usage per predicted complex is shown as a function of diffusion sampling depth ($N = 1, 10, 25, 50, 100$), with each saturation level corresponding to an independent random seed. Shaded regions indicate interquartile ranges across 106 systems. **(B)** Saturation sampling improves best-case structural quality with model-dependent magnitude. Median maximum DockQ is shown as a function of sampling depth. Horizontal dashed lines mark DockQ quality thresholds separating incorrect (< 0.23), acceptable (0.23-0.49), medium (0.49-0.80), and high (≥ 0.80) quality regimes. **(C)** Quality gains per unit energy peak at low saturation levels. Bar plots show median gains in maximum DockQ between consecutive saturation levels. Line plots show efficiency frontiers relating cumulative DockQ improvement to cumulative energy expenditure. **(D)** Saturation sampling reduces incorrect predictions but plateaus at higher depth. Stacked bar plots show DockQ quality-category distributions across saturation levels. **(E)** Seed selection introduces substantial variability in prediction quality. Boxplots show the DockQ range obtained from five independent seeds using only the first sample (sample$_0$). **(F)** Different seeds explore distinct solution landscapes. Cross-seed saturation trajectories show median cumulative maximum DockQ across systems as a function of sample number. Early advantages persist, while poor initial seeds plateau at lower quality ceilings.

B SUPPLEMENTARY FIGURES

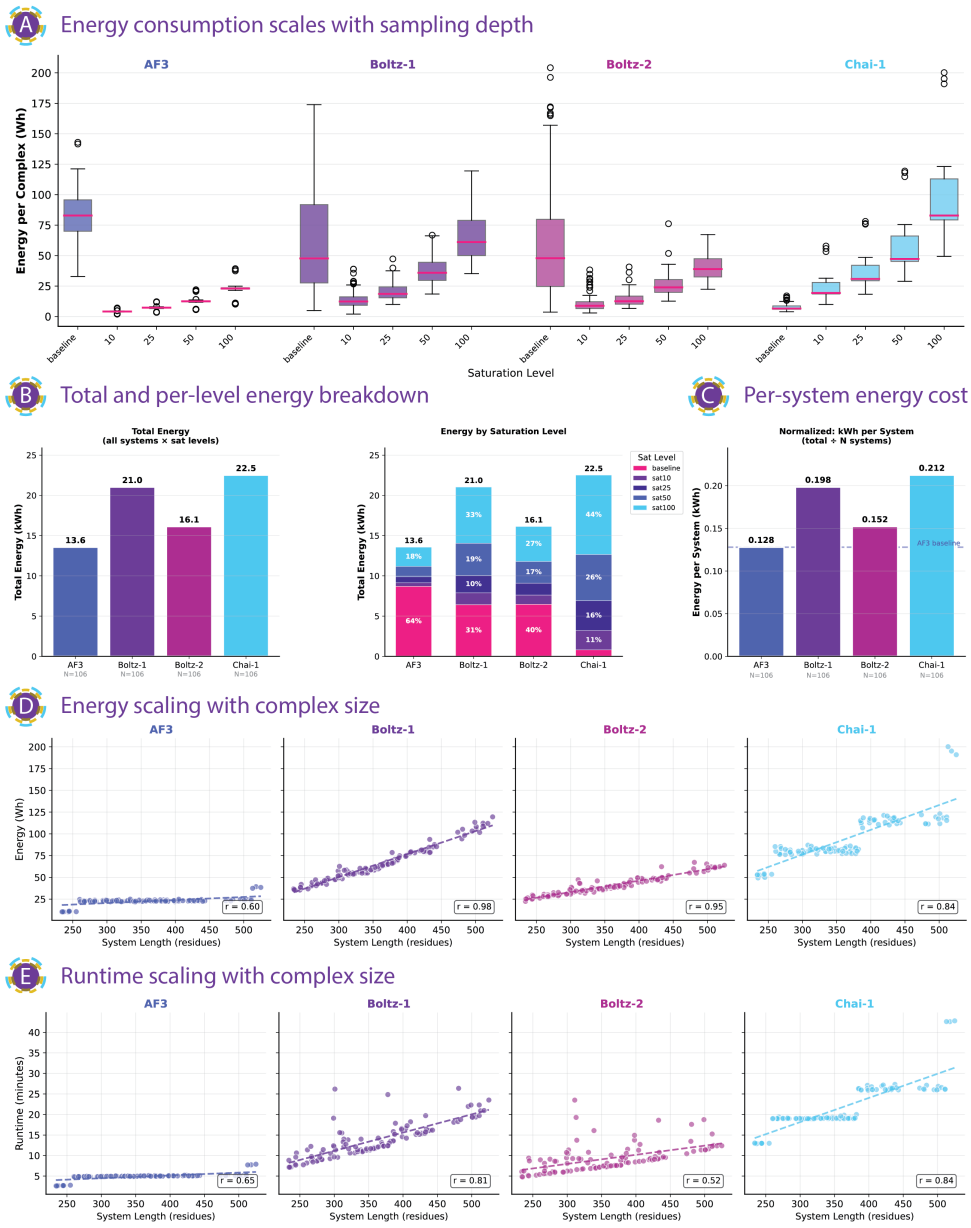
Supplementary Figures 1–5 provide additional analyses that support and extend the results presented in the main text and Appendix. These figures elaborate on sampling behavior, confidence calibration, seed dependence, and computational cost across models, offering deeper quantitative insight into trends discussed earlier.



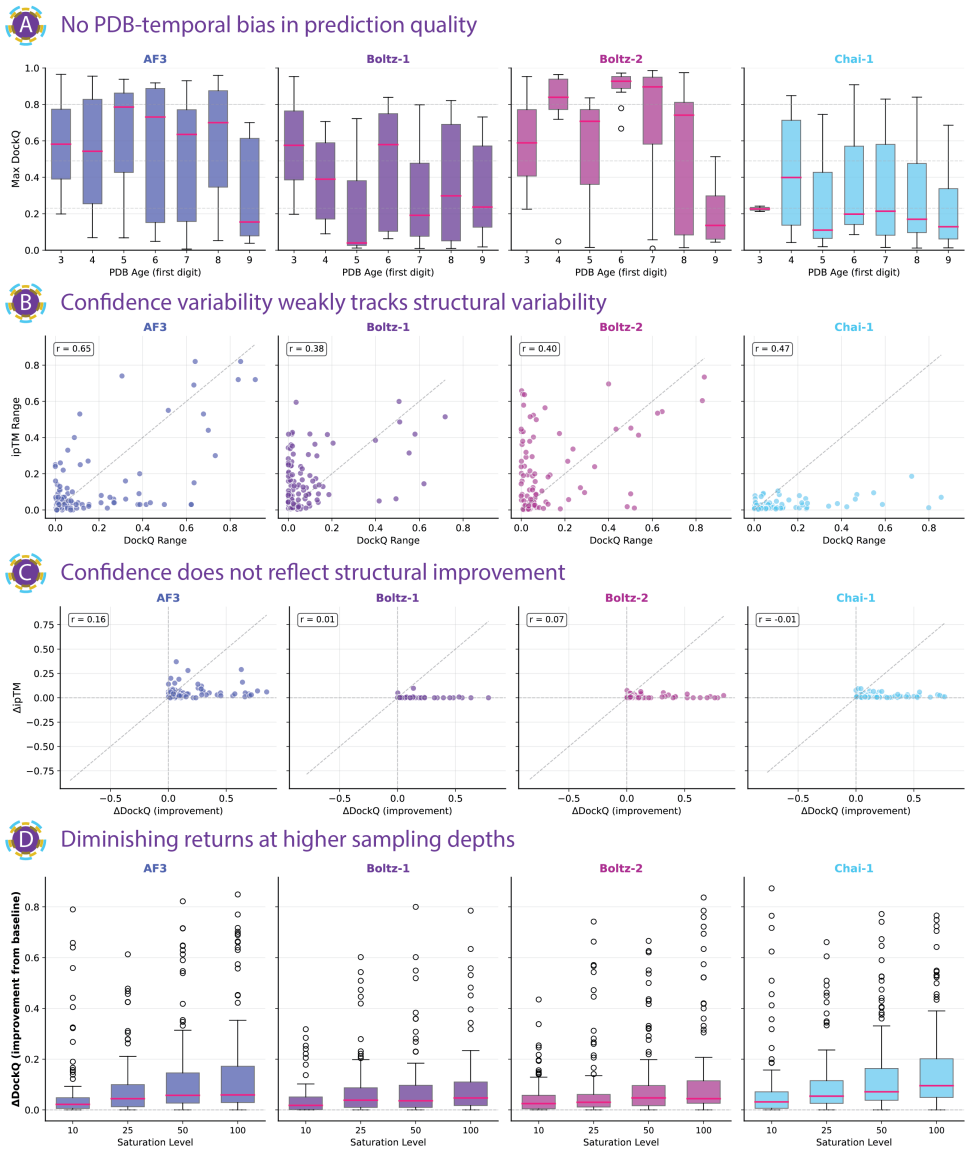
Supplementary Figure 1: **Saturation sampling improves docking quality but does not realign confidence scores across models. (A)** DockQ improvement across models. Distribution of DockQ improvement following saturation sampling, shown as ΔDockQ (Best $-$ Sample$_0$) for AF3 ($N = 106$), Boltz-2 ($N = 106$), and Chai-1 ($N = 103$). All three models exhibit right-skewed improvement distributions, indicating that additional sampling frequently refines docking geometry. **(B)** Quality category shifts after saturation sampling. DockQ quality category distributions (Incorrect, Acceptable, Medium, High) for the initial single-sample prediction (Sample$_0$) and the best structure obtained after saturation sampling (Best). Saturation sampling substantially increases the fraction of acceptable or higher-quality models for AF3 and Boltz-2, with more limited gains for Chai-1. **(C)** Paired DockQ improvement per system. Paired comparison of DockQ scores between Sample$_0$ and Best predictions for each system. Lines connect paired predictions; boxplots summarize score distributions. All three models show significant improvements in DockQ after saturation sampling (Wilcoxon signed-rank test, $p < 0.001$). The dashed line indicates the DockQ $= 0.23$ acceptability threshold. **(D)** ipTM remains stable despite quality gains. Paired comparison of ipTM confidence scores between Sample$_0$ and Best predictions. In contrast to DockQ, ipTM shows no consistent improvement with saturation sampling for AF3 or Chai-1, and only a variable decrease for Boltz-2 (p-values shown), indicating weak coupling between confidence and structural refinement. **(E)** Cross-model confidence agreement. Scatter plots show pairwise ipTM values between model pairs (AF3 vs. Boltz-2, AF3 vs. Chai-1, Boltz-2 vs. Chai-1), with point shapes indicating the model achieving higher DockQ for each system and dashed lines indicating equal confidence. Pearson correlation coefficients and winner counts are reported for each comparison, revealing modest cross-model agreement and frequent mismatches between higher confidence and better structural quality.



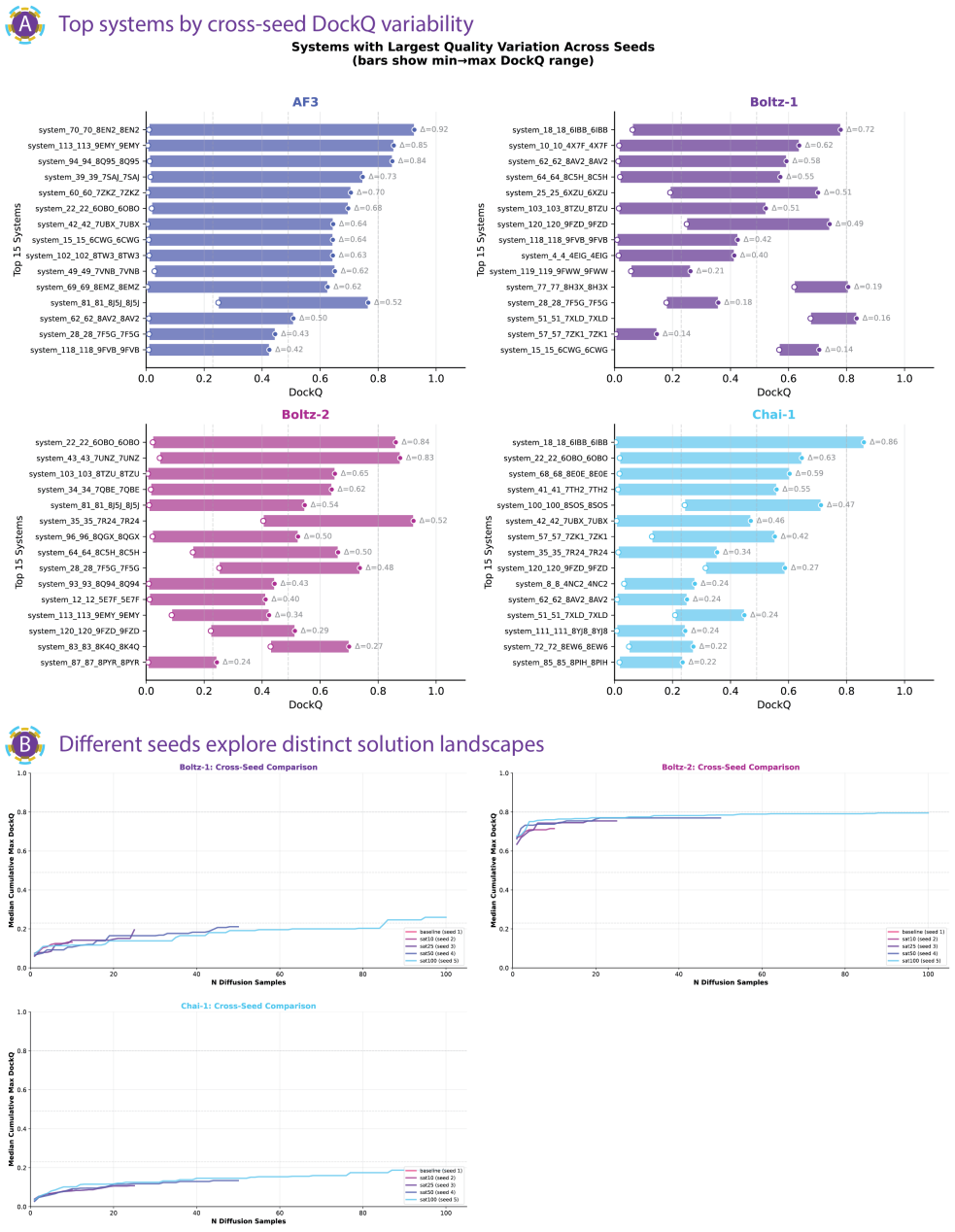
Supplementary Figure 2: **Robustness of docking predictions across VHH variants. (A)** Per-system DockQ across VHH variants. Best DockQ scores obtained after saturation sampling for 15 antigen systems, each evaluated with two distinct VHH binders (VHH1 and VHH2). For each antigen (PDB ID shown above each panel), bars show the best DockQ achieved by AF3, Boltz-2, and Chai-1 for each VHH replicate. The dashed horizontal line indicates the DockQ $= 0.23$ threshold for acceptable docking quality. Letter annotations denote relative ranking of predictions within each antigen system. Consistent high DockQ across both VHHs indicates robust recovery of the antigen epitope despite paratope variation, whereas discordant outcomes highlight sensitivity to VHH identity. **(B)** Model consistency across VHH variants. Bars show the number of antigen systems (out of 15) for which both VHH replicates achieved acceptable docking quality (DockQ ≥ 0.23). AF3 succeeds in 10/15 systems, compared to 6/15 for Boltz-2 and 2/15 for Chai-1, indicating superior consistency of AF3 in recovering correct antibody-antigen interactions across VHH variants.



Supplementary Figure 3: **Computational cost of diffusion-based antibody-antigen docking across models.** **(A)** Energy consumption per complex as a function of sampling depth. Boxplots show GPU energy usage (Wh) per system for AF3, Boltz-1, Boltz-2, and Chai-1 across increasing numbers of diffusion samples (baseline, 10, 25, 50, and 100). Each point corresponds to a single system; boxes summarize medians and interquartile ranges. **(B)** Aggregate energy consumption across all systems and sampling levels. Left: total GPU energy consumed (kWh) across all 106 systems and all sampling depths for each model. Middle: contribution of each sampling level to total energy usage, shown as stacked bars with percentages indicating relative contribution. **(C)** Normalized energy usage per system. Bars show total energy expenditure per system (kWh) aggregated across all sampling levels, enabling direct comparison of average computational cost between models. **(D)** Scaling of energy consumption with system size. Scatter plots show GPU energy usage (Wh) versus total system length (number of residues) for each model, with dashed lines indicating linear fits. Pearson correlation coefficients quantify scaling strength. **(E)** Scaling of runtime with system size. Scatter plots show wall-clock runtime (minutes) versus system length for each model, with linear fits and Pearson correlation coefficients shown. Runtime scaling mirrors energy scaling, with model-specific differences in slope and variance.



Supplementary Figure 4: **Temporal effects, confidence variability, and sampling-dependent improvements across models.** **(A)** Prediction quality as a function of target age. Boxplots show maximum DockQ distributions for AF3, Boltz-1, Boltz-2, and Chai-1 grouped by the first digit of the PDB deposition year. No systematic degradation in performance is observed for older structures. **(B)** Relationship between structural variability and confidence variability. Scatter plots show the range of ipTM values versus the range of DockQ scores across stochastic replicates for each system. Pearson correlation coefficients indicate weak-to-moderate coupling depending on the model. **(C)** Coupling between structural improvement and confidence change. Scatter plots show changes in ipTM (ΔipTM) versus improvements in DockQ (ΔDockQ) from baseline to best sampled structure. Near-zero correlations indicate confidence scores do not track refinement. **(D)** Distribution of DockQ improvement as a function of sampling depth. Boxplots show ΔDockQ relative to baseline across increasing sample counts (10, 25, 50, 100). Deeper sampling expands the upper tail of improvement, while median gains show diminishing returns.



Supplementary Figure 5: **Seed-dependent variability and cross-seed convergence of docking quality. (A)** Systems with the largest seed-dependent variability. Horizontal bar plots show the top 15 systems per model (AF3, Boltz-1, Boltz-2, Chai-1) ranked by the DockQ range across five independent seeds ($\Delta = \max - \min$ DockQ). Some systems span nearly the full DockQ scale. **(B)** Cross-seed convergence of sampling trajectories. Line plots show median cumulative maximum DockQ as a function of diffusion sample count for each seed. Curves illustrate how seeds explore distinct solution landscapes and converge to different quality plateaus, revealing partial convergence and persistent seed dependence.

C METHODS

C.1 VHH-ANTIGEN DATASET CURATION

For benchmarking antibody-antigen complex prediction, VHH-antigen structures were curated from two complementary sources: SAbDab-nano (Schneider et al., 2022) (downloaded March 2025; 90%

sequence redundancy cutoff; bound complexes only) and the Antigen-Antibody Complex Database (AACDB) (Zhou et al., 2025). To minimize overlap with the training data of AlphaFold3 (AF3), Boltz-2, and Chai-1, only post-October 2021 depositions were retained from SAbDab-nano, corresponding to the earliest training cutoff among the evaluated tools (Boltz-1). To explicitly probe potential memorization effects, pre-cutoff structures from AACDB were retained and later used to define train-leakage subsets.

All structures were filtered to include protein antigens only, with crystallographic resolution $\leq$ 3.0 Å, VHH length between 110-150 amino acids, and antigen length between 100-400 amino acids. The two datasets were merged and exact PDB-chain duplicates were removed. For epitope-paratope variation analysis, we defined a set of distinct VHHs binding the same antigen within a single PDB entry (Supplementary Table 1). This yielded 15 PDB entries containing such replicates (30 systems).

To maintain computational feasibility, the total number of unique systems was capped below 125. After manual inspection, 17 systems were excluded due to structural artifacts, VHHs contacting multiple antigen chains, or ambiguous chain annotation. The final benchmark comprised 106 VHH-antigen systems from 91 unique PDB entries.

Antigen secondary structure was assigned using DSSP, amino-acid composition was computed directly from sequences, and PDB novelty was inferred from the first character of the PDB accession code.

C.2 Real vs. shuffled complex generation

To assess interaction specificity, we constructed a full combinatorial pairing matrix between the curated VHHs and antigens. For each VHH, the experimentally observed pairing with its cognate antigen was designated as the *real* complex, while all other VHH-antigen pairings were designated as *shuffled* complexes. This design preserves realistic molecular interfaces while systematically breaking biological specificity.

Unless otherwise stated, all analyses were performed on the complete 106×106 VHH-antigen matrix, enabling direct comparison between real and shuffled predictions under identical modeling conditions.

C.3 Structure prediction

Structure predictions for the full combinatorial dataset (106 VHHs $\times$ 106 antigens; 50 samples per complex) were performed using three independent structure prediction frameworks: AlphaFold3 (AF3), Boltz-2, and Chai-1. Predictions were executed across multiple high-performance computing environments. Exact hardware specifications, runtime stacks, and software versions were logged at job start for reproducibility.

C.3.1 Chai-1

Chai-1 predictions were performed using version 0.6.1. All complexes were run on the Immunohub cluster using one GPU per job, starting on August 14, 2025 at 12:00 PM. Each system was predicted with 50 models using the following parameters: `num_trunk_samples=5`, `num_diffn_samples=10`, `num_trunk_recycles=3`, `num_diffn_timesteps=200`, `seed=42`. No MSAs were computed; instead, ESM-based embeddings were used under Chai-1's default inference mode.

C.3.2 Boltz-2

Predictions were performed using the Boltz CLI v2.2.0. Boltz-1 and Boltz-2 were selected via `--model boltz1` or `--model boltz2`, respectively. Unless otherwise noted, all runs used `--use_msa_server`, `recycling_steps=3`, `sampling_steps=200`, `diffusion_samples=50`, and a fixed random seed (42). Outputs were written in PDB or mmCIF format, with full per-residue and interaction-level scores exported using `--write_full_pde`. Boltz-2 jobs ran on Immunohub and Boltz-1 on VEGA.

C.3.3 ALPHAFOLD3

AlphaFold3 predictions were performed using `alphafold3-3.0.1` via Singularity containers. Each system was predicted with 50 diffusion samples using `num_diffusion_samples=50`, `run_data_pipeline=true`, and `run_inference=true`. Each job was run with a fixed random seed (`seed=1`). MSAs were computed using the built-in data pipeline.

C.4 TRAIN, TEST, AND MIXED LABELING

Each VHH-antigen system was assigned to *train*, *test*, or *both* categories based on the training cutoffs of the corresponding model: (i) AF3: ∼30 September 2021, (ii) Chai-1: ∼12 January 2021, and (iii) Boltz-2: ∼1 June 2023. A system was labeled *train* if both the VHH and antigen were present in the model's training set, *test* if neither was present, and *both* if one component was present while the other was not.

C.5 STRUCTURAL SIMILARITY BETWEEN REPLICATES

To quantify structural reproducibility, geometric and contact-based similarity metrics were computed across independently generated predictions. All analyses were performed on Cα atomic coordinates extracted from PDB files using custom Python scripts (NumPy, SciPy, pandas). For each complex, all pairwise comparisons among replicate structures were computed and aggregated to yield per-system metrics.

C.5.1 ROOT MEAN SQUARE DEVIATION (RMSD)

Global structural similarity was quantified using RMSD after superposition on the antigen via the Kabsch algorithm:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{X}_i - (\mathbf{R}\mathbf{Y}_i + \mathbf{t})\|^2},$$

where $\mathbf{X}$ and $\mathbf{Y}$ are corresponding Cα coordinates, and $\mathbf{R}$ and $\mathbf{t}$ are the optimal rotation and translation. Lower RMSD values indicate higher structural consistency. RMSD values were averaged across all replicate pairs to obtain mean ± SD estimates per system.

C.6 ENERGY MONITORING

GPU telemetry (power draw, utilization, memory usage) was sampled every 5 seconds using `nvidia-smi` and written to per-run CSV files. Wall-clock timing and system metrics were captured via `/usr/bin/time -v`. Total energy consumption (Wh) was calculated as:

$$E_{\text{run}} = \sum_i P_i \Delta t / 3600,$$

with $\Delta t = 5$ s. Runs were aggregated by model and diffusion sample count and reported as mean ± SD.

C.7 COMPUTATIONAL COST AND SAMPLING EFFICIENCY

Structural quality was assessed using DockQ (Mirabello & Wallner, 2024), which integrates interface RMSD, ligand RMSD, and fraction of native contacts into a single score ranging from 0 (incorrect) to 1 (perfect). Quality categories followed CAPRI conventions.

For saturation analysis, each system was evaluated using five independent random seeds, with each seed assigned to a single saturation level ($N = 1, 10, 25, 50, 100$ diffusion samples). Marginal quality gains were computed as per-system differences in maximum DockQ between consecutive saturation levels and summarized as medians across systems. Efficiency frontiers were generated by plotting cumulative median ΔDockQ against cumulative median energy expenditure.

Correlations between confidence-score changes (ΔipTM) and quality changes (ΔDockQ) were assessed using Pearson correlation coefficients.

D DATA AVAILABILITY

Supplementary tables will be released upon publication.

E DISCUSSION

This work benchmarks whether modern AI complex-prediction tools can discriminate cognate antibody-antigen binding and recover correct paratope-epitope interfaces. Our central findings are twofold. First, across tools, high-level confidence scores frequently fail to separate real from mismatched ("shuffled") complexes. Second, while increased sampling can improve best-case geometry, it does not resolve the upstream problem of selecting the correct binding mode. Importantly, even when structural quality improves substantially under aggressive sampling (as measured, for example, by DockQ for real complexes), model confidence scores often remain unchanged, revealing a disconnect between structural refinement and confidence calibration. Below, we place these results in the broader AI biologics landscape and outline practical implications for drug discovery and model development.

E.1 WHERE THE FIELD STANDS: DISCONNECT BETWEEN *de novo* GENERATION AND *de novo* PREDICTION

The AI biologics ecosystem is now shaped by strong competition among major players (e.g., Latent, Chai, Nabla, Isomorphic, Boltz), with rapid iteration cycles. Despite headline claims, although *de novo* generation success rates are increasing into the double-digit range (approximately 10-15% on difficult interface tasks (Team et al., 2025; Stark et al., 2025)), our results indicate a disconnect between *de novo* design ("generate an antibody for a given target without exploiting other information") and *de novo* prediction ("identify all antigens that can be bound by a given antibody, and vice versa"). While such success rates may appear encouraging in isolation, they remain low in absolute terms for practical screening scenarios, where even modest false-positive rates translate into substantial experimental burden. Our real-versus-shuffled discrimination benchmark demonstrates that models can generate interfaces that appear plausible across many pairings, producing an abundance of false positives that are challenging to triage.

An emerging industry screening workflow is to generate thousands to millions of candidates and then filter using internal confidence scores (e.g., pLDDT, ipTM, PAE). Our results suggest that this strategy is fragile for antibody-antigen binding, because confidence does not equate to accurate biology. In a many-VHH-versus-many-antigen screening setting, shuffled complexes frequently achieve confidence comparable to real complexes, while some high-quality structures remain underconfident depending on the tool. This breaks the assumption that "high-confidence docking implies correct binding" and helps explain why confidence-guided selection can yield many experimental failures even when predicted structures appear polished. Consistent with recent proposals for alternative evaluation metrics (Xu et al., 2025; Almeida et al., 2025), our results underscore the need for scores that more directly reflect biological meaning rather than internal structural self-consistency.

Crucially, drug viability depends on properties that these scores do not directly encode, including induced-fit compatibility, entropic penalties, off-target propensity, developability constraints, and tolerance to antigen dynamics or conformational selection. Confidence metrics were not designed as surrogates for molecular functionality, and our data reinforce that treating them as such inflates false positives, especially in interface-dependent problems.

E.2 WHY EVOLUTIONARY SIGNALS AND SAMPLING HELP LESS THAN HOPED FOR ANTIBODIES

A common intuition is that better MSAs and richer evolutionary information should improve AI docking and design. However, antibody paratopes-especially CDR loops-are among the most variable regions in biology, and much of binding specificity arises from flexibility, conformational diversity, and context-dependent loop rearrangements rather than sequence conservation (Smorodina et al., 2025b; Spoendlin et al., 2025; Fernández-Quintero et al., 2021). This supports the view that the utility of MSAs in antibody-focused tasks is more nuanced than in general protein structure prediction. While MSAs are critical for accurate monomer modeling (removal of MSAs in

AF3 leads to substantial degradation in structural accuracy (Ali et al., 2025)), their contribution to antibody-antigen docking is constrained by the extreme diversity of CDRs, which limits informative evolutionary signal. As a result, MSA-derived confidence estimates may be inherently less reliable for antibody-antigen interfaces than for other protein complexes. Alternative MSA strategies may therefore provide better guidance, particularly for evolutionary complex or rapidly mutating systems such as antibodies.

On the other hand, sampling helps but exposes a deeper bottleneck: path selection versus refinement. Saturation or diffusion sampling improves best-case DockQ in many systems, confirming that sampling is valuable as a refinement mechanism. However, we observe essentially zero correlation between changes in DockQ (ΔDockQ) and changes in ipTM (ΔipTM) across all three models (Pearson $r = -0.03$, -0.04, and -0.02 for AF3, Boltz-2, and Chai-1, respectively), indicating that confidence scores do not track structural improvement. The persistence of "non-improver" systems supports a two-stage view: (i) path or seed selection determines the qualitative docking mode (often incorrect), and (ii) saturation refines within that chosen mode.

In this sense, current models appear to have a "fixed mindset": once a confidence level is assigned to a given seed or trajectory, it remains largely invariant even when the resulting structure improves substantially. This likely reflects an architectural constraint: confidence scores such as ipTM are derived from MSA-based pairwise representations computed in the trunk network, which remains fixed during diffusion sampling. If the model commits early to an incorrect binding mode, additional diffusion samples often cannot rescue it. This has immediate practical implications: simply sampling more may be an expensive path to diminishing returns, and large seed sweeps, while sometimes recommended, can be prohibitive at screening scale without guaranteeing specificity or providing confidence-based validation that sampling helped. Developing confidence measures sensitive to binding-mode quality, or orthogonal strategies such as consensus across independent seeds, remains an important challenge.

E.3 Structural biology perspective: flexibility is not optional

Antibody recognition is not merely geometric matching; it is often a dynamic process in which CDR loops adapt to the antigen surface and exploit transient conformations. This is particularly relevant for targets with disordered regions or multiple accessible states. A major unresolved challenge is therefore distinguishing hallucinated (geometrically plausible but physically or functionally implausible) complexes from viable ones that can maintain binding under realistic dynamics.

This points to a missing ingredient in many AI pipelines: standardized representations of flexibility and experimentally grounded dynamics. A key practical bottleneck in current AI biologics workflows is selection. Generating candidates is increasingly easy; selecting the biologically functional subset is not. We argue that molecular-dynamics-informed filtering is a promising bridge between AI-generated hypotheses and functional plausibility (Bashour et al., 2024; Park & Izadi, 2024). However, today's AI-generated "MD-like" outputs are often not physically faithful enough to serve as ground truth.

Standardized, cross-method training data for dynamics therefore matter. Efforts such as the DINO proposal (Smorodina et al., 2025c; Anonymous, 2025), aimed at becoming a "PDB-equivalent for dynamics," directly address the lack of unified MD data across force fields and simulation protocols. By enabling consistent learning from flexibility and improving computational-experimental alignment, such resources could underpin a new generation of screening criteria that go beyond static confidence scores, including interface stability under perturbation, persistence of key contacts, accessibility of binding-competent states, and robustness to antigen motion.

As many AI startups increasingly partner with pharmaceutical companies, the core industrial question is not "which model has the highest benchmark score?" but "which pipeline yields validated candidates efficiently?" Our benchmark suggests that confidence-score-based filtering alone is insufficient for antibody-antigen interface prediction. The field may need to shift from (i) generating more candidates and ranking by confidence to (ii) ranking by biophysical plausibility and functional robustness, potentially using MD-informed filters and interface-specific mechanistic constraints.

E.4 Limitations and future directions

Our study focuses on VHH-antigen complexes under a controlled real-versus-shuffled pairing design that explicitly tests specificity rather than docking plausibility alone. As such, it quantifies the feasibility of generalized binding prediction. This design may overestimate performance limitations in narrower settings where the antigen epitope is known, constraints are available, or the target is rigid. Additionally, similar benchmarks should be extended to more complex antibody formats to assess generalizability beyond VHHs.

Large-scale resources such as AbSet (Almeida et al., 2025), which contains hundreds of thousands of antibody structures, highlight an additional challenge: volume does not imply reliability. Systematic assessment of how many computational antibody structures are structurally sound, interface-correct, or functionally meaningful remains limited. Our results support the view that antibody-antigen complex prediction is not yet a solved problem despite massive *in silico* datasets.

Looking forward, progress will likely require (1) interface-specific training objectives incorporating hard, mutation-diverse negatives across affinity ranges (Ursu et al., 2024); (2) confidence estimates calibrated to specificity and functional plausibility rather than structural self-consistency; and (3) standardized dynamic datasets enabling models and filters to account for flexibility and induced fit. The most impactful near-term advance may not be a new confidence score, but a robust, scalable post-prediction filtering layer grounded in biophysics and dynamics.

In summary, this study addresses a central challenge in immunotherapy: accurate modeling of antibody-antigen interactions. Despite major advances in structure prediction, current tools fall short in reliably identifying correct paratope-epitope interfaces. Through systematic benchmarking, we expose key failure modes in interface identification and binding-mode selection. These findings suggest that the next leap in AI biologics may come not from increasingly sophisticated architectures alone, but from moving beyond static confidence proxies toward biochemical, mechanistic, and flexibility-aware evaluation of binding-and from building the data infrastructure needed to make that possible.

References

Website. https://arxiv.org/html/2512.11892v1, 2025.

J. Abramson et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493–500, 2024.

R. Akbar et al. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *MAbs*, 14:2008790, 2022.

M. Ali et al. Improving nanobody structure prediction with self-distillation. *bioRxiv*, 2025. doi: 10.64898/2025.12.01.691162.

D. S. Almeida et al. Abset: A standardized data set of antibody structures for machine learning applications. *Journal of Chemical Information and Modeling*, 65:4767–4774, 2025.

F. Ambrosetti, B. Jiménez-García, J. Roel-Touris, and A. M. J. J. Bonvin. Modeling antibody-antigen complexes by information-driven docking. *Structure*, 28:119–129.e2, 2020.

Anonymous. Probabilistic modeling of antibody structural dynamics. In *2nd edition of Frontiers in Probabilistic Inference: Learning meets Sampling*, 2025. URL https://openreview.net/forum?id=26C7kYxxlp.

H. Bashour et al. Biophysical cartography of the native and human-engineered antibody landscapes quantifies the plasticity of antibody developability. *Communications Biology*, 7:922, 2024.

S. Basu and B. Wallner. Dockq: A quality measure for protein-protein docking models. *PLoS One*, 11:e0161879, 2016.

W. Bielska et al. Applying computational protein design to therapeutic antibody discovery – current state and perspectives. *arXiv*, 2025. arXiv:q-bio.BM.

J. Boitreaud et al. Chai-1: Decoding the molecular interactions of life. *Synthetic Biology*, 2024.

P. Bradley. Structure-based prediction of t cell receptor:peptide-mhc interactions. *eLife*, 12, 2023.

X. Dang et al. Epitope mapping of monoclonal antibodies: a comprehensive comparison of different technologies. *MAbs*, 15:2285285, 2023.

C. Discovery et al. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955.

R. Evans et al. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2021. doi: 10.1101/2021.10.04.463034.

M. L. Fernández-Quintero, G. Georges, J. M. Varga, and K. R. Liedl. Ensembles in solution as a new paradigm for antibody structure prediction and design. *MAbs*, 13:1923122, 2021.

M. L. Fernández-Quintero et al. Challenges in antibody structure prediction. *MAbs*, 15:2175319, 2023.

S. Fromm, M. Ludaic, and A. Elofsson. Evaluating deep learning based structure prediction methods on antibody-antigen complexes. *bioRxiv*, 2025. doi: 10.1101/2025.07.11.662141.

L. R. Genz, S. Nair, N. Nagar, and M. Topf. Assessing scoring metrics for alphafold2 and alphafold3 protein complex predictions. *Protein Science*, 34:e70327, 2025.

V. Greiff, G. Yaari, and L. G. Cowell. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Current Opinion in Systems Biology*, 24:109–119, 2020.

M. Grieswelle et al. A new benchmark for deep learning based affinity prediction: Solving the inter-protein scoring noise problem. *ChemRxiv*, 2025. doi: 10.26434/chemrxiv-2025-sf3cs.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. *arXiv*, 2017. doi: 10.48550/ARXIV.1706.04599. arXiv:1706.04599 [cs.LG].

A. Harmalkar, S. Lyskov, and J. J. Gray. Reliable protein-protein docking with alphafold, rosetta, and replica-exchange. *bioRxiv*, 2023. doi: 10.1101/2023.07.28.551063.

F. N. Hitawala and J. J. Gray. What does alphafold3 learn about antigen and nanobody docking, and what remains unsolved? *Bioengineering*, 2024.

F. N. Hitawala and J. J. Gray. What does alphafold3 learn about antibody and nanobody docking, and what remains unsolved? *MAbs*, 17:2545601, 2025.

C. M. Holt et al. Contrastive learning enables epitope overlap predictions for targeted antibody discovery. *bioRxiv*, 2025. doi: 10.1101/2025.02.25.640114.

I. Johansson-Åkhe and B. Wallner. Improving peptide-protein docking with alphafold-multimer using forced sampling. *Frontiers in Bioinformatics*, 2:959160, 2022.

J. Jumper et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021.

M. Kull et al. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *arXiv*, 2019. doi: 10.48550/ARXIV.1910.12656. arXiv:1910.12656 [cs.LG].

D. Kuroda and J. J. Gray. Shape complementarity and hydrogen bond preferences in protein-protein interfaces: implications for antibody modeling and protein-protein docking. *Bioinformatics*, 32: 2451–2456, 2016.

L. Lannelongue and M. Inouye. Environmental impacts of machine learning applications in protein science. *Cold Spring Harbor Perspectives in Biology*, 15, 2023.

A. H. Laustsen, V. Greiff, A. Karatt-Vellatt, S. Muyldermans, and T. P. Jenkins. Animal immunization, in vitro display technologies, and machine learning for antibody discovery. *Trends in Biotechnology*, 39:1263–1273, 2021.

S. Levine et al. Origin-1: a generative ai platform for de novo antibody design against novel epitopes. *bioRxiv*, 2026. doi: 10.64898/2026.01.14.699389.

F. Meng et al. A comprehensive overview of recent advances in generative models for antibodies. *Computational and Structural Biotechnology Journal*, 23:2648–2660, 2024.

C. Mirabello and B. Wallner. Dockq v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 40, 2024.

C. Morand, A.-L. Ligozat, and A. Névéol. How green can ai be? a study of trends in machine learning environmental impacts. *arXiv*, 2024. doi: 10.48550/ARXIV.2412.17376. arXiv:2412.17376 [cs.LG].

M. D. Overath et al. Predicting experimental success in de novo binder design: A meta-analysis of 3,766 experimentally characterised binders. *bioRxiv*, 2025. doi: 10.1101/2025.08.14.670059.

M. D. Overath et al. Accelerating multi-objective vhh discovery via integrated high-throughput selection and alphafold3-guided structure prediction. *bioRxiv*, 2026. doi: 10.64898/2026.01.19.700436.

E. Park and S. Izadi. Molecular surface descriptors to predict antibody developability: sensitivity to parameters, structure models, and conformational sampling. *MAbs*, 16:2362788, 2024.

S. Passaro et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

J. A. Ruffolo, L.-S. Chu, S. P. Mahajan, and J. J. Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14:2389, 2023.

C. Schneider, M. I. J. Raybould, and C. M. Deane. Sabdab in the age of biotherapeutics: updates including sabdab-nano, the nanobody structure tracker. *Nucleic Acids Research*, 50:D1368–D1372, 2022.

E. Smorodina et al. Structural informatic study of determined and alphafold2 predicted molecular structures of 13 human solute carrier transporters and their water-soluble qty variants. *Scientific Reports*, 12:20103, 2022.

E. Smorodina et al. Structural modeling of antibody variant epitope specificity with complementary experimental and computational techniques. In *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2025a.

E. Smorodina et al. Structural modeling of antibody variant epitope specificity with complementary experimental and computational techniques. In *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2025b.

Eva Smorodina, Victor Greiff, and Rahmad Akbar. DINO: dynamics-informed dataset to overcome the limitations of static molecular data in AI-driven drug discovery. In *NeurIPS 2025 AI for Science Workshop*, 2025c. URL `https://openreview.net/forum?id=ENvitzhEhh`.

F. C. Spoendlin et al. Predicting the conformational flexibility of antibody and t cell receptor complementarity-determining regions. *Nature Machine Intelligence*, 7:1755–1767, 2025.

H. Stark et al. Boltzgen: Toward universal binder design. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689494.

Chai Discovery Team et al. Zero-shot antibody design in a 24-well plate. *bioRxiv*, 2025. doi: 10.1101/2025.07.05.663018.

S. Unsal, B. Holland, I. Sardag, and E. Timucin. Confidence scoring for ai-predicted antibody-antigen complexes: Anticonf as a precision-driven metric. *bioRxiv*, 2025. doi: 10.1101/2025.07.25.666870.

E. Ursu et al. Training data composition determines machine learning generalization and biological rule discovery. *bioRxiv*, 2024. doi: 10.1101/2024.06.17.599333.

S. Weeratunga et al. Interrogation and validation of the interactome of neuronal munc18-interacting mint proteins with alphafold2. *Journal of Biological Chemistry*, 300:105541, 2024.

B. D. Weitzner et al. Modeling and docking of antibody structures with rosetta. *Nature Protocols*, 12:401–416, 2017.

J. Wohlwend et al. Boltz-1: Democratizing biomolecular interaction modeling. *Biophysics*, 2024.

X. Xu, I. Coratella, V. Reys, and A. M. Bonvin. Deeprank-ab: a scoring function for antibody-antigen complexes based on geometric deep learning. *bioRxiv*, 2025. doi: 10.64898/2025.12.03.691974.

R. Yin and B. G. Pierce. Evaluation of alphafold antibody-antigen modeling with implications for improving predictive accuracy. *Protein Science*, 33:e4865, 2024.

Y. Zhou et al. A comprehensive antigen-antibody complex database unlocking insights into interaction interface. *eLife*, 14, 2025.