REVISITING STRUCTURE-BASED FILTERING FOR AI-DESIGNED ANTIBODIES: OPPORTUNITIES, LIMITS, AND WHAT MATTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI antibody design can generate far more candidates than can be experimentally tested, making post-generation filtering a central bottleneck. We study whether structure-prediction outputs can help prioritize AI-designed antibodies for experimental validation. We evaluate this problem on 4,184 experimentally characterized antibody–antigen pairs, combining newly generated wet-lab measurements with curated affinity data across diverse antigens and assay settings. We introduce two complementary filtering signals: Regional Interaction Contrast (RIC), which captures support for the predicted antibody–antigen interface, and Non-CDR Antibody Confidence (NAC), which measures antibody confidence outside the CDRs. Across AlphaFold 3, Protenix-v2, and ESMFold2, we find a striking difference in what information matters for filtering. Interface evidence is strongest when comparing candidates designed for different antigens, whereas antibody-level reliability is substantially more informative for ranking candidates within the same antigen. In particular, non-CDR antibody confidence consistently outperforms whole-antibody confidence across predictors, revealing an unexpected source of filtering signal beyond the predicted interface. Finally, we examine the information encoded in the predictor outputs themselves. Predicted structures contain substantial but incomplete binding-related signal, revealing both the opportunities and limits of structure-based filtering: better metrics can exploit current predictors more effectively, but their performance ultimately remains constrained by what those predictors can distinguish. Code is available in an anonymous repository.

1 INTRODUCTION

Antibodies are highly versatile molecular-recognition proteins whose ability to bind specific targets underlies their widespread use in therapeutics, diagnostics, and biological research. Conventional discovery relies on immunization, library screening, and affinity maturation (Chao et al., 2006). Computational methods increasingly complement these approaches by directly generating antigen-conditioned sequences, backbones, or joint sequence–structure designs (Luo et al., 2022; Kong et al., 2023). Recent diffusion-, flow-matching-, and all-atom generative systems have further extended this paradigm to experimentally validated epitope-specific antibodies, single-domain antibodies, and other protein binders (Bennett et al., 2025; Stark et al., 2025; Yu et al., 2026). As a result, large and structurally diverse libraries of candidate binders can now be generated entirely *in silico*.

This progress does not eliminate the experimental bottleneck. A computational campaign can readily produce hundreds or thousands of candidates, whereas synthesis, expression, binding assays, and downstream characterization remain comparatively costly and capacity limited. Deciding *which generated candidates should be tested* has therefore become a central post-generation problem. Candidate selection often begins with sequence-level criteria, such as removing developability liabilities, undesirable motifs, or poorly behaved sequences. Such filters are valuable, but provide limited evidence about the target-specific question that ultimately determines experimental success: whether a candidate actually binds its intended antigen.

Protein-complex predictors provide a natural source of such target-specific evidence. AlphaFold-Multimer and AlphaFold 3, together with more recent structure-prediction models, predict biomolecu-

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.

lar complexes while providing coordinates together with residue-, pair-, and interface-level confidence or interaction estimates (Evans et al., 2021; Abramson et al., 2024; Zhang et al., 2026; Candido et al., 2026; Passaro et al., 2025). Scores derived from these predictions, including ipTM, pDockQ, pDockQ2, and AntiConf, provide readily available summaries of predicted complex or interface confidence (Bryant et al., 2022; Zhu et al., 2023; Ünsal et al., 2026). Structural confidence and interface properties have consequently become important prioritization signals in computational binder-design pipelines (Pacesa et al., 2025).

For antibody discovery, however, the relevant decision is not simply whether a predicted complex appears credible. An effective *in silico* filter must balance two competing errors. False positives consume limited experimental capacity on candidates that do not bind, increasing the cost of discovery. False negatives are also costly: successful designs may be rare for difficult antigens, and an overly restrictive filter can discard the few true binders produced by a generator. Recovering such hits may require another full cycle of generation, prediction, and screening. The practical objective is therefore to retain as many true binders as possible while keeping the false-positive rate and experimental testing burden under control.

This decision problem raises a more fundamental question about the use of structure prediction for antibody filtering: given the predicted structures and confidence outputs for a candidate, how much information about binding is present, and how effectively can that information be organized into candidate-selection scores? We therefore study **structure-based *in silico* filtering of designed antibodies**: whether predictor outputs can provide structural evidence that more accurately prioritizes AI-designed antibody candidates by their likelihood of binding the intended antigen. We evaluate this question on a large and heterogeneous collection of experimentally characterized antibody–antigen pairs spanning diverse assay modalities, evidence sources, and quantitative affinity measurements.

Our contributions are fourfold:

- **A practical formulation of structure-based antibody filtering.** We study post-generation filtering of AI-designed antibodies using only structure-prediction outputs, with evaluation centered on recovering true binders under realistic false-positive and experimental-budget constraints.
- **A diverse experimental benchmark.** We assemble 4,184 experimentally characterized antibody–antigen candidates spanning multiple antigens and assay modalities, combining curated literature data with newly generated wet-lab measurements.
- **Two complementary filtering metrics.** We introduce Regional Interaction Contrast (RIC) to capture interface-level support across predicted structures, and Non-CDR Antibody Confidence (NAC) to measure the reliability of the predicted antibody outside the CDRs.
- **An unexpected finding on antibody reliability.** Across three structure predictors, antibody-level reliability emerges as the strongest signal for within-antigen filtering, while interface evidence remains more effective when comparing candidates designed for different antigens. Surprisingly, confidence outside the CDRs consistently outperforms whole-antibody confidence, revealing a robust filtering signal that is largely overlooked by existing interface-centric approaches.

2 Background and Problem Formulation

2.1 AI Antibody Design as a Generate–Predict–Filter Pipeline

From generation to filtering. An AI antibody design workflow begins with an antigen and constraints such as the antibody format, framework scaffold, CDR lengths, residues that may be redesigned, and a desired epitope or set of target hotspots. **Generate.** Current systems span several generative paradigms. RFantibody first generates a docked antibody backbone and then assigns sequences with ProteinMPNN (Bennett et al., 2025); PPIFlow couples backbone flow matching with sequence design and in silico maturation (Yu et al., 2026); and Chai-2 directly supports zero-shot antibody design (Chai Discovery Team et al., 2025). More general all-atom systems, including BoltzGen and ODesign, jointly reason over binder and target structure and can generate antibody-like or other binder modalities (Stark et al., 2025; Zhang et al., 2025). Despite these architectural differences,

generation is only the first stage. **Predict.** Candidate sequences are screened for sequence validity and developability, and their antibody–antigen complex structures are independently predicted. **Filter.** Designs are then assessed using the predicted complexes and physicochemical quantities such as interface confidence, buried surface area, hydrogen bonding, shape complementarity, and Rosetta interface-energy estimates. The final experimental set is chosen under a synthesis budget by combining predicted complex quality, interface quality, developability, and diversity.

In silico filtering with structural confidence. The central step in post-design filtering is to predict each designed antibody–antigen complex with a co-folding model. Alongside coordinates, co-folding models report confidence metrics used directly for screening: pLDDT measures local residue-level confidence, ipTM measures confidence in the relative placement of chains, PAE measures positional uncertainty after alignment, and PDE or gPDE summarizes uncertainty in predicted inter-residue distances. These metrics are attractive because they directly probe whether the antibody is well formed and placed against the antigen through a confident interface. Antibody–antigen prediction remains challenging, however, because these complexes often lack strong paired evolutionary information and concentrate binding determinants in flexible, hypervariable CDR loops. Errors in CDR geometry or docking therefore propagate directly into filtering and reduce downstream hit recovery (Ünsal et al., 2026; Erasmus et al., 2026). How best to use predicted complexes and their confidence outputs for in silico screening therefore remains an open question.

2.2 Quantifying the Utility of Structural Filters

To compare structural filters, we need to quantify how effectively they turn predicted complex confidence into useful experimental candidates. In silico filtering sits between a scalable computational pipeline and a much more constrained experimental stage. Two practical facts determine what makes a filter useful:

Experimental validation is expensive and capacity limited. Sequence generation and structure prediction can be parallelized over many candidates, whereas synthesis, expression, and functional testing remain limited by cost, time, and laboratory capacity. False positives therefore spend scarce experimental resources on candidates that passed the computational screen but fail in the laboratory.

Successful computational designs remain rare. Current antibody designers have limited and strongly target-dependent hit rates. A low-recall filter can reject the few genuine successes before they are tested, forcing additional cycles of generation, prediction, and screening.

An effective filter must therefore suppress false positives without doing so by discarding most true binders.

Filtering decision. Consider a designed antibody a in design context c, which includes its antigen target and design constraints. Let M_{ac} be a structural-confidence score such as ipTM. Given a prespecified threshold σ, the filter retains the candidate for subsequent analysis or experimental testing when

$$\widehat{A}_{ac}(\sigma) = \mathbb{1}\{M_{ac} \geq \sigma\}\,. \tag{1}$$

Here $\widehat{A}_{ac} = 1$ means that the candidate advances. In practice, M_{ac} may be a single metric or a composite score combining multiple structural criteria, such as interface confidence, antibody pLDDT, and inter-chain error. In this paper, we focus on individual metrics.

Definition 1 (Filtering outcomes)**.** $\mathcal{D} = \{(a, c, Y_{ac})\}$ denotes a set of designed antibodies, their design contexts, and their **wet-lab** validation outcomes $Y_{ac} \in \{0, 1\}$. Here $Y_{ac} = 1$ means that candidate a satisfies the prespecified experimental binding criterion in context c and is treated as a binder; $Y_{ac} = 0$ denotes a negative outcome. A filter advances the candidate to synthesis or wet-lab validation when $\widehat{A}_{ac} = 1$ ($M_{ac} \geq \sigma$). Comparing $\widehat{A}_{ac}$ with Y_{ac} yields true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The precision P, recall R, false-positive rate F, and selected fraction Q are

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad F = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \quad Q = \frac{\mathrm{TP} + \mathrm{FP}}{|\mathcal{D}|}. \tag{2}$$

Definition 2 (Filtering utility)**.** Let $\underline{R}$ be the 95% Wilson lower confidence bound for recall, and let $\overline{F}$ be the corresponding upper confidence bound for the false-positive rate. Let f_{safe} be an acceptable

FPR and $f_{\text{severe}} > f_{\text{safe}}$ mark the onset of severe FPR penalty. Define

$$W(F) = \exp\left[-\left(\frac{F}{f_{\text{severe}}}\right)^2 - \left(\frac{(F - f_{\text{safe}})_+}{f_{\text{severe}} - f_{\text{safe}}}\right)^2\right], \qquad (x)_+ = \max(x, 0). \tag{3}$$

For fixed f_{safe} and f_{severe}, the utility at threshold σ and the best utility of metric m are

$$U(m, \sigma; \mathcal{D}) = \underline{R}\, W(\overline{F}), \qquad U^\star(m; \mathcal{D}) = \max_{\sigma \in \mathbb{R}} U(m, \sigma; \mathcal{D}). \tag{4}$$

Interpreting the utility. The first term in W gradually penalizes FPR even below f_{safe}, favoring cleaner filters when recall is similar. Above f_{safe}, the second term steepens the penalty: gains in recall must then offset a greater cost from false positives. The penalty remains smooth, with no hard cutoff. We set $f_{\text{safe}} = 0.10$ and $f_{\text{severe}} = 0.30$; W equals 0.97, 0.89, 0.50, and $e^{-2} \approx 0.14$ at FPRs of 5%, 10%, 20%, and 30%, respectively. Using the Wilson lower bound for recall and upper bound for FPR also accounts for uncertainty in the observed rates.

3 METHOD

In this section, we introduce two complementary structure-based filtering metrics, RIC and NAC, which capture interface-level and antibody-level evidence from predicted complex ensembles, respectively.

3.1 CANDIDATE EVIDENCE

For each antibody–antigen candidate i, let X_i be the ensemble of complexes predicted across independent runs (called seeds in the inference scripts) and their samples. These predictions provide two kinds of evidence worth examining. First, do the antigen regions predicted to contact the antibody CDRs receive stronger, reproducible support than the next-ranked antigen regions? Second, how confidently is the antibody itself modeled outside the more variable CDR loops? We capture these questions with Regional Interaction Contrast (RIC) and Non-CDR Antibody Confidence (NAC). We represent each candidate by these two complementary sources of evidence:

$$\mathbf{E}_i = (\text{RIC}(X_i), \text{NAC}(X_i)).$$

For simplicity, we omit the candidate index i below. We denote $\mathcal{A}$, $\mathcal{C}$, and $\mathcal{G}$ as the residue sets of the antibody, CDR regions, and antigen, respectively.

3.2 REGIONAL INTERACTION CONTRAST

High confidence at an entire interface need not single out the CDR contacts predicted for a candidate. RIC instead compares the most strongly predicted antigen contacts of each CDR residue with the next-ranked antigen region. It asks whether the first region also has stronger support from predicted aligned error (PAE) across two different runs. Comparing regions reduces the influence of uniformly high agreement, while comparing runs tests whether the contact pattern persists across predictions. The next-ranked contacts provide plausible alternatives to the proposed contact region. A reference run defines both regions; PAE from that run and another run tests their relative support.

For a reference run r, let $\bar{C}_r(a, b)$ be the mean bidirectional contact probability over its samples for CDR residue a and antigen residue b, and let $h_r(a) = \max_{b \in \mathcal{G}} \bar{C}_r(a, b)$. Ranking antigen residues by $\bar{C}_r(a, b)$ defines normalized weights $w^H_{ra}(b)$ and $w^L_{ra}(b)$ for the highest- and next-ranked regions. Let $A_{rtjk}(a, b)$ measure bidirectional PAE agreement between sample j in run r and sample k in another run t; larger PAE lowers agreement. At $\tau = 16$ Å, a directional PAE of 8 Å in each run gives agreement $1/2$. Write $\Delta_{rtjk}(a)$ for the difference in PAE agreement between the two regions of CDR residue a, and z_{rtjk} for the score of that sample pair. The candidate score is their nested median:

$$\Delta_{rtjk}(a) = \sum_{b \in \mathcal{G}} \left[w^H_{ra}(b) - w^L_{ra}(b)\right] A_{rtjk}(a, b), \qquad z_{rtjk} = \frac{\sum_{a \in \mathcal{C}} h_r(a)^2 \Delta_{rtjk}(a)}{\sum_{a \in \mathcal{C}} h_r(a)},$$
$$\text{RIC}(X) = \operatorname*{median}_{r} \operatorname*{median}_{t \neq r} \operatorname*{median}_{j,k} z_{rtjk}. \tag{5}$$

The squared contact support emphasizes CDR residues with stronger predicted contacts; the nested medians require the contrast to persist across sample pairs and runs. The contact average, PAE agreement, rank weights, and zero-support rule are specified in Appendix D.

3.3 NON-CDR ANTIBODY CONFIDENCE

NAC measures confidence in the antibody outside the CDRs. CDR loops are structurally variable, so excluding them gives a broader view of how confidently the rest of the antibody is modeled within the complex. For each sample, we average pLDDT over the non-CDR residues, then take the median over samples within a run and the median across runs. With p^{nc}_{rj} denoting the non-CDR confidence of sample j in run r, this gives

$$\mathrm{NAC}(X) = \underset{r}{\mathrm{median}} \underset{j \in \mathcal{K}_r}{\mathrm{median}} p^{\mathrm{nc}}_{rj}, \qquad p^{\mathrm{nc}}_{rj} = \frac{\sum_{a \in \mathcal{A} \setminus \mathcal{C}} \mathrm{pLDDT}_{rja}}{100|\mathcal{A} \setminus \mathcal{C}|}. \tag{6}$$

NAC is computed from the predicted complex, so its confidence values also reflect the context in which the antibody and antigen were modeled.

4 EXPERIMENTS

4.1 SETUP

Public benchmarks. We curate two benchmarks from public sources. IgDesign is an experimentally validated inverse-folding method that redesigns heavy-chain CDRs on a native antibody–antigen backbone and measures binding by surface plasmon resonance (SPR) (Shanehsazzadeh et al., 2023). We also separately curate antibody–antigen pairs with direct affinity measurements reported in papers and patents.

Wet-lab validation. We additionally evaluate display-selected antibodies against seven antigens. TfR, spike S2, and GPC3 were screened by yeast display; CD98hc, EGFR, PLXNA1 ectodomain, and TNFRSF23 by phage display. Sequencing across selection rounds tracks clone enrichment. For each antigen, after sequence quality control and deduplication, we classify 100 late-enriched candidates as putative binders and 100 baseline or early low-enrichment candidates as putative non-binders. Table 1 summarizes the three benchmarks; Appendix B describes their construction, eligibility criteria, and filtering in detail.

Co-folding model training-set exposure. AlphaFold 3, Protenix-v2, and ESMFold2 each use a 30 September 2021 cutoff for PDB training structures (Candido et al., 2026). ESMFold2 also trains on AlphaFold DB-predicted monomers. An exact-match audit against SAbDab found no pre-cutoff structure matching both the antibody variable chain(s) and antigen construct among the 4,184 candidates; thus, no exact complex was identified in the models' PDB structural training data.

Structure prediction protocol. For each of the 4,184 antibody–antigen pairs, we independently predict the complex with AlphaFold 3, Protenix-v2, and ESMFold2. Every predictor uses MSAs, no structural templates, 10 random seeds, and 5 samples per seed, yielding up to 50 predicted complexes per pair and predictor. All subsequent metrics are computed from these predicted structures and their associated confidence outputs. Predictor configurations and inference details are given in Appendix C.

Baselines. Baselines include predictor-native ranking confidence, pair-iPTM, antibody and interface pLDDT, directional inter-chain PAE, and PDE/gPDE. We also evaluate pDockQ (Bryant et al., 2022), pDockQ2 (Zhu et al., 2023), three variants of the interface score derived from predicted aligned error (ipSAE) (Dunbrack, 2025), and AntiConf (Ünsal et al., 2026). Further baseline implementation details are provided in Appendix C.2.

Threshold selection. Each metric advances a candidate when its score satisfies $M \geq \sigma$, where σ is a threshold. To compare metrics on equal terms, we sweep all attainable thresholds and select the value that maximizes the filtering utility in Definition 2. Portfolio-wide (*global*) filtering uses one threshold shared across all antigens, whereas campaign-specific (*target-specific*) filtering optimizes the decision rule separately within each antigen. We report the resulting utility and precision, recall at FPR levels of 10%, 20%, and 30%, and recall when the selected fraction is fixed at 5% or 20%. The last view directly models a limited synthesis budget: candidates are ranked by the metric and only the highest-scoring fraction is tested. Further details on threshold selection are given in Appendix C.4.

Table 1: **Benchmark dataset statistics.** Unique-antigen counts refer to distinct antigen constructs within each benchmark.

Dataset	Unique antigens	Positive	Negative	Total	Positive rate	Label source
Display-based enrichment	7	700	700	1,400	50.0%	round-wise selection enrichment
IgDesign SPR	7	249	922	1,171	21.3%	replicated SPR binding
Literature/patent affinity	165	1,510	103	1,613	93.6%	direct SPR/BLI/ITC K_D
Total	179	2,459	1,725	4,184	58.8%	

4.2 RESULTS

Tables 2 and 3 evaluate structure-based filtering from two practical perspectives. The first examines portfolio-level selection, where candidates designed for different antigens compete under a shared threshold, while the second examines within-antigen ranking under target-specific thresholds. The main observations from these complementary evaluations are summarized below.

Table 2: **In silico filtering using a single global threshold shared across antigen targets.** Recall and precision at $U^\star$ use the threshold that maximizes filtering utility. Recall and precision are expressed as percentages. The benchmark includes 4,184 candidates per predictor, comprising 2,459 positives and 1,725 negatives. Best values are bold and second-best values are underlined. For ESMFold2, 275 out-of-memory complexes are excluded. *gRIC is used for ESMFold2 because the model does not provide predicted contact probabilities; it approximates RIC using 8-Å geometric contacts from predicted coordinates.

Metric	Recall @ 10% FPR	Recall @ 20% FPR	Recall @ 30% FPR	Recall @ $Q = 5\%$	Recall @ $Q = 20\%$	Utility $U^\star$	Recall @ $U^\star$	Precision @ $U^\star$	Average rank
AlphaFold 3									
Pair-iPTM	14.64	19.03	**31.35**	8.48	19.34	0.1150	14.84	67.47	3.12
Interface PAE ↓	9.88	15.45	26.68	5.21	17.89	0.0815	13.58	56.71	8.12
Antibody pLDDT	7.40	16.10	18.71	5.53	17.32	0.0794	13.42	56.12	9.25
Ranking score	14.92	16.63	25.82	**8.50**	18.58	0.1229	14.92	72.53	3.62
pDockQ	10.33	14.03	30.42	6.67	17.97	0.0791	10.33	59.76	7.44
pDockQ2	12.48	15.41	26.88	5.12	18.06	0.0972	12.77	63.56	6.94
ipSAE-d0res	9.72	16.31	27.45	5.33	18.06	0.0831	13.62	57.17	6.94
ipSAE-d0dom	8.74	16.92	27.17	4.96	17.85	0.0856	15.21	57.98	6.75
ipSAE-d0chn	14.07	19.28	30.46	**8.50**	19.40	0.1119	13.91	69.09	3.12
AntiConf	2.81	6.99	19.15	2.56	15.86	0.0284	13.95	44.66	10.88
RIC (Interface)	**17.93**	**21.43**	28.95	**8.50**	**21.27**	**0.1465**	**17.61**	**75.83**	**1.50**
NAC (Antibody)	9.80	14.52	17.61	4.80	16.27	0.0767	10.33	57.86	10.31
Protenix-v2									
Pair-iPTM	11.43	20.86	29.89	7.97	20.33	0.0904	11.14	66.67	3.00
Pair-gPDE ↓	2.89	5.57	13.79	2.72	13.42	0.0223	3.38	29.96	12.56
Interface PAE ↓	9.92	15.90	27.12	4.60	18.02	0.0824	12.24	57.44	5.81
Antibody pLDDT	3.90	10.90	16.96	3.13	14.84	0.0432	9.39	43.10	9.25
Ranking score	3.13	8.38	27.12	2.81	17.16	0.0415	15.62	49.42	8.69
pDockQ	6.71	9.48	21.27	4.96	15.13	0.0514	7.08	48.74	8.25
pDockQ2	9.96	15.94	27.00	4.27	17.97	0.0852	12.08	58.81	5.88
ipSAE-d0res	6.75	16.84	29.16	5.57	18.83	0.0639	16.63	54.53	5.25
ipSAE-d0dom	6.67	18.10	29.20	4.92	19.36	0.0685	18.42	56.62	5.00
ipSAE-d0chn	11.39	22.12	32.98	7.89	20.37	0.0892	11.02	66.10	3.25
AntiConf	3.21	5.49	16.51	2.93	14.07	0.0227	3.17	34.21	11.62
RIC (Interface)	**27.90**	**32.53**	**33.75**	**8.46**	**27.65**	**0.2309**	**27.49**	**82.14**	**1.00**
NAC (Antibody)	3.50	8.09	14.68	2.52	13.42	0.0321	6.71	36.03	11.44
ESMFold2									
Pair-iPTM	1.57	3.80	11.05	1.48	11.63	0.0132	4.16	21.38	10.62
Interface PAE ↓	3.18	10.20	15.07	2.95	14.18	0.0379	8.77	39.20	7.25
Antibody pLDDT	3.40	7.29	11.85	2.82	12.21	0.0314	6.26	33.98	8.31
pDockQ	8.09	14.18	20.93	5.28	17.35	0.0646	9.79	50.34	4.38
pDockQ2	3.62	11.58	18.56	2.86	16.23	0.0457	14.40	47.49	6.00
ipSAE-d0res	4.87	22.00	27.59	3.67	20.44	0.0833	**22.00**	59.56	3.00
ipSAE-d0dom	4.79	20.71	27.46	3.35	20.21	0.0799	19.86	57.81	4.25
ipSAE-d0chn	7.69	24.46	31.98	5.14	21.33	0.0984	21.78	61.41	2.25
AntiConf	1.79	7.29	16.32	1.79	14.40	0.0258	7.78	33.59	8.56
gRIC*	**23.03**	**28.58**	**33.68**	**8.72**	**25.36**	**0.1868**	21.69	**81.38**	**1.25**
NAC (Antibody)	2.42	4.65	7.74	2.24	11.00	0.0205	3.80	25.68	10.12

Key Messages from Tables 2 and 3.

Interface evidence leads in across-antigen filtering. With one threshold shared across antigens in Table 2, RIC has the best average rank for AlphaFold 3 (1.50) and Protenix-v2 (1.00), with large gains in low-FPR recall and utility. Its geometric-contact approximation, gRIC, also ranks first for ESMFold2 (1.25), whereas NAC ranks near the bottom (10.12).

Antibody reliability is strongest within an antigen. The ordering reverses when candidates compete only within the same antigen. NAC ranks first for AlphaFold 3 and ESMFold2 (1.38 for both) and remains among the strongest scores for Protenix-v2 (2.12). In contrast, RIC falls to average ranks of 10.25 and 13.00 for AlphaFold 3 and Protenix-v2, and gRIC ranks 10.62 for ESMFold2.

Confidence outside the CDRs is more informative than whole-antibody confidence. In campaign selection, NAC outperforms antibody pLDDT in 23 of 24 predictor–measure comparisons: all eight for AlphaFold 3 and ESMFold2, and seven of eight for Protenix-v2. Removing the CDRs from the confidence summary therefore almost always improves within-antigen ranking, despite discarding the residues most directly involved in binding.

Predictor comparisons depend on the evaluation context. Among ten shared baselines, AlphaFold 3 has higher values in 54 of 80 baseline–measure comparisons in across-antigen filtering (Table 2), whereas Protenix-v2 leads in 77 of 80 comparisons under campaign-specific evaluation (Table 3).

Within-antigen filtering offers greater fixed-budget gains. At a 20% testing budget, most scores in Table 2 fall below the roughly 20% recall expected from random selection. In Table 3, NAC reaches 35.82%, 36.21%, and 39.22% recall with AlphaFold 3, Protenix-v2, and ESMFold2. Within-antigen filtering therefore appears more effective for the predictors tested.

Reading precision and recall. In the full 4,184-candidate benchmark, 58.8% are binders, so testing at $Q = 20\%$ caps recall at 34.0% even with perfect ranking. At $U^\star$, a large fraction of selected candidates comes from the mostly positive literature/patent benchmark. Such concentration can yield high precision while leaving many binders elsewhere unselected. RIC/gRIC achieve FPRs of 8.0%, 8.5%, and 6.6% for AlphaFold 3, Protenix-v2, and ESMFold2, respectively, at $U^\star$.

Interpreting the contrasting signals. The recall gap between across-antigen and within-antigen filtering may arise from target-dependent prediction errors. Candidates for one antigen can share docking and confidence biases, which affect their scores similarly and partly cancel in within-antigen ranking. Across antigens, different biases can shift score scales, weakening a shared threshold. On the same candidates as Table 3, a shared global ranking still favors RIC/gRIC over NAC (Table 4), although its fixed-budget recall remains below random selection. Table 7 also helps explain the strength of NAC: non-CDR pLDDT has the highest mean within-antigen AUC in all six predictor–benchmark comparisons. In five, AUC decreases from non-CDR to whole-antibody, CDR, and interface pLDDT. This ordering suggests a trade-off between binding relevance and prediction reliability: confidence in the more reliably modeled non-CDR region can filter better than confidence in the binding-proximal but harder-to-model CDRs or predicted interface.

Table 3: **Campaign-specific evaluation of 2,400 candidates (944 binders, 1,456 non-binders) across 13 display-based and IgDesign antigens.** Candidates compete only within an antigen; literature/patent antigens have too few negatives for low-FPR evaluation and are excluded. Appendix C.3 details eligibility, weighting, and results aggregation. For ESMFold2, 107 out-of-memory complexes are excluded.

Metric	Recall @ 10% FPR	Recall @ 20% FPR	Recall @ 30% FPR	Recall @ $Q = 5\%$	Recall @ $Q = 20\%$	Utility $U^\star$	Recall @ $U^\star$	Precision @ $U^\star$	Average rank
AlphaFold 3									
Pair-iPTM	15.44	25.99	38.74	7.49	25.57	0.0807	20.68	51.54	8.00
Interface PAE ↓	17.30	28.12	41.88	6.23	25.12	0.0847	22.51	46.82	7.88
Antibody pLDDT	27.85	42.27	51.00	8.91	33.39	0.1367	34.12	60.36	2.38
Ranking score	14.83	29.24	36.43	6.82	28.05	0.0870	23.48	49.33	6.25
pDockQ	19.63	33.50	44.64	6.60	26.00	0.1021	25.97	51.36	5.25
pDockQ2	16.51	29.20	41.04	5.99	26.69	0.0869	23.42	48.75	6.75
ipSAE-d0res	12.01	17.69	21.78	7.58	26.64	0.0557	**47.41**	58.59	7.12
ipSAE-d0dom	10.62	17.69	22.54	8.52	26.32	0.0538	45.55	58.40	7.62
ipSAE-d0chn	10.61	17.69	22.54	8.52	26.02	0.0539	45.85	57.53	7.75
AntiConf	17.31	29.20	40.99	6.81	25.79	0.0842	20.88	48.45	7.38

Continued on next page

Table 3 continued

Metric	Recall @ 10% FPR	Recall @ 20% FPR	Recall @ 30% FPR	Recall @ $Q = 5\%$	Recall @ $Q = 20\%$	Utility $U^\star$	Recall @ $U^\star$	Precision @ $U^\star$	Average rank
RIC (Interface)	15.55	23.80	33.72	5.78	21.14	0.0702	18.20	44.28	10.25
NAC (Antibody)	**33.97**	**44.99**	**57.41**	**11.09**	**35.82**	**0.1791**	34.49	**69.14**	**1.38**
Protenix-v2									
Pair-iPTM	30.31	43.53	50.32	13.01	33.30	0.1674	33.77	63.50	6.88
Pair-gPDE ↓	27.60	38.97	46.50	11.44	32.82	0.1529	27.72	60.23	11.38
Interface PAE ↓	34.00	43.32	49.74	**14.04**	34.63	0.1866	33.80	64.35	4.25
Antibody pLDDT	32.94	44.38	52.92	12.95	36.18	0.1837	34.44	67.44	3.50
Ranking score	29.82	44.15	51.60	13.64	33.52	0.1671	33.83	61.50	6.00
pDockQ	23.91	40.18	50.72	8.29	33.25	0.1280	30.49	57.61	10.62
pDockQ2	34.03	42.91	49.72	13.74	34.15	0.1914	34.37	64.03	4.62
ipSAE-d0res	30.15	42.93	50.58	13.08	33.37	0.1667	32.53	62.19	7.50
ipSAE-d0dom	30.96	43.21	49.69	13.15	33.30	0.1635	31.72	61.69	8.06
ipSAE-d0chn	31.95	41.87	49.75	13.15	32.59	0.1682	31.64	65.18	7.31
AntiConf	31.50	41.93	51.98	13.10	34.11	0.1792	35.89	60.80	5.75
RIC (Interface)	14.28	23.31	33.42	5.46	18.65	0.0652	18.31	53.07	13.00
NAC (Antibody)	**36.54**	**49.40**	**59.26**	11.70	**36.21**	**0.1983**	**37.77**	**67.59**	**2.12**
ESMFold2									
Pair-iPTM	24.60	33.95	46.83	8.92	33.61	0.1385	29.14	56.04	6.00
Interface PAE ↓	25.50	34.09	46.59	9.23	33.80	0.1514	29.52	49.89	5.25
Antibody pLDDT	32.78	43.49	51.40	10.19	37.21	0.1889	37.08	58.40	2.62
pDockQ	18.41	29.50	41.46	8.00	26.14	0.0934	22.94	47.79	9.25
pDockQ2	24.79	34.50	44.92	9.38	33.78	0.1407	28.04	50.55	5.88
ipSAE-d0res	22.30	29.02	30.69	10.06	33.21	0.1210	39.56	52.86	7.38
ipSAE-d0dom	23.83	28.42	30.99	10.44	33.89	0.1268	39.81	54.32	5.75
ipSAE-d0chn	23.08	28.72	30.39	10.22	33.82	0.1299	**40.11**	57.02	5.75
AntiConf	24.63	33.62	44.02	9.63	33.38	0.1451	29.32	52.15	6.12
gRIC*	11.88	22.53	35.15	4.69	21.19	0.0566	17.70	42.90	10.62
NAC (Antibody)	**37.92**	**47.40**	**58.13**	**11.59**	**39.22**	**0.2165**	39.45	**62.96**	**1.38**

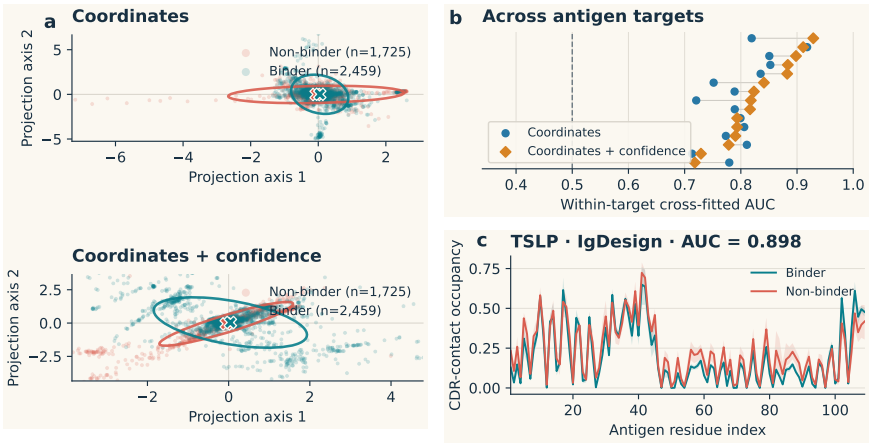


Figure 1: Binding information in AlphaFold 3 outputs. **(a)** Projections of 4,184 candidates from coordinates (top) and coordinates plus confidence (bottom), each over 10 seeds and 5 samples. **(b)** Cross-fitted within-target AUCs across 15 antigen–dataset strata with at least five binders and five non-binders; confidence improves separability for many targets. **(c)** Median CDR-contact occupancy (interquartile range) for 114 binders and 49 non-binders against IgDesign TSLP. Contact regions largely overlap despite a within-target AUC of 0.898.

4.3 Discussion

Predictor identifiability sets the computational ceiling. Here, identifiability refers to a simple limit: a downstream metric cannot recover distinctions that are absent from the predictor outputs. In the extreme, if binders and non-binders produce the same distribution of predicted coordinates and confidence values, no deterministic score computed from those outputs can distinguish them. Additional seeds and samples can reduce stochastic variation, but they cannot correct systematic errors that are reproduced across the ensemble.

We first examine how much binding-related information is encoded in predicted coordinates alone. Using AlphaFold 3 prediction ensembles, we represent each candidate by coordinate-derived features,

including CDR–antigen distances and geometric contacts. A cross-fitted kernel witness analysis estimates binder and non-binder separability within each antigen, achieving a target-macro AUC of 0.801. Thus, predicted coordinates contain substantial binding-related signal, but the broad overlap between the two classes indicates that this information is incomplete. The qualitative projection in Figure 1a illustrates this limited separation.

Adding confidence tensors, including contact probabilities, bidirectional PAE, and residue-level pLDDT, further improves separability (target-macro AUC from 0.801 to 0.827), showing that confidence tensors provide additional information beyond coordinates. However, this improvement is not uniform across antigens: some targets become substantially more separable, whereas others show little gain or even reduced separation (Figure 1b).

For TSLP in the IgDesign benchmark, coordinate-derived contact patterns of binders and non-binders overlap broadly across antigen regions (Figure 1c), suggesting that predicted interaction locations alone are insufficient to explain binding outcomes. Nevertheless, the complete structure-prediction outputs achieve a within-target AUC of 0.898, indicating that additional information remains in confidence tensors and interface-related properties. Together, these results highlight both the opportunity and limit of structure-based filtering: better metrics can extract more of the binding-related information encoded in predictor outputs, but their ultimate performance is constrained by distinctions that the predictors themselves do not capture.

Biological separability sets the benchmark ceiling. A benchmark is informative only when binder and non-binder labels represent distinct biological outcomes for matched antibody–antigen constructs. Our data combine display-based enrichment, replicated SPR, and literature-derived affinity measurements, which differ in assay conditions, detection limits, antigen formats, and binder definitions. Quality control reduces but cannot eliminate this heterogeneity. The evaluation ceiling therefore also depends on label fidelity. Table 6 demonstrates that both display-based enrichment and IgDesign show within-antigen AUCs significantly above chance for all three predictors.

5 Conclusion and Limitations

We revisit post-design antibody filtering as a practical selection problem under limited experimental capacity, where candidates may be compared across different antigens or ranked within a single antigen. We identify two complementary signals from predicted complexes: interaction support for the CDR–antigen interface, captured by RIC, and antibody structural confidence outside the CDRs, captured by NAC. Across 4,184 labeled antibody–antigen pairs and three structure predictors, we find that interface evidence is most useful when candidates designed for different antigens compete, whereas antibody-level confidence is more informative for ranking candidates within the same antigen. Surprisingly, non-CDR confidence consistently outperforms whole-antibody pLDDT for within-antigen ranking, revealing an underexplored source of filtering signal beyond the predicted interface. More broadly, predicted geometry and confidence provide useful but incomplete binding information: better metrics can extract more from current predictors, but their ultimate performance is limited by distinctions that the predictors themselves do not capture.

Limitations. Only part of our benchmark comes from AI design; it also includes display-selected antibodies and examples curated from the literature, which may have different sequence and structural distributions from AI-generated candidates. Still, these experimentally characterized antibodies allow us to test whether co-folding outputs carry binding-related signals that can guide filtering after AI design. Moreover, in the predictor training data available for audit, we found no exact pre-cutoff antibody–antigen complexes, reducing concern about direct memorization of the evaluated pairs. The benchmark combines display-based enrichment, replicated binding measurements, and published affinity data, which capture related but not identical notions of binding. Enrichment should be interpreted cautiously, because antibody expression, clone growth, and sampling may also influence round-to-round frequency changes. Across all three predictors, the best cutoff varies by antigen (Appendix A.5); predictor-derived scores should therefore not be assumed to share a calibrated absolute scale across targets. We optimize each metric's cutoff on labeled candidates for a fair benchmark comparison. A new campaign instead needs independent calibration or a prespecified top-Q budget. Limited computational resources restricted us to ten seeds and five samples per seed; assessing whether more seeds and samples change our conclusions remains for future work.

AI USE STATEMENT

Generative AI tools were used to assist with language refinement, literature search and exploration, scientific discussion, and the development of data-processing scripts. All AI-assisted text, code, and computational outputs were carefully reviewed and validated by the authors. The authors made all final methodological, analytical, and interpretive decisions and take full responsibility for the content, results, and conclusions presented in this work.

ETHICS STATEMENT

This study combines published antibody-binding data with new yeast- and phage-display experiments. It does not assess clinical outcomes or use human-participant data. Antibody design and selection methods can have dual-use implications; applications to new targets should be subject to appropriate biosafety and institutional review. Our scores prioritize candidates for further testing and should not be interpreted as evidence of safety, therapeutic efficacy, or binding without experimental confirmation.

REPRODUCIBILITY STATEMENT

Code is available in an anonymous repository. The repository provides implementations of RIC, NAC, and all reported baselines, together with evaluation scripts, input specifications, and tests. The paper also provides detailed descriptions of baseline implementations, evaluation protocols, and structure predictor inference settings to support reproducibility. Prediction ensembles are not included due to their large storage requirements (more than one terabyte), and the newly generated wet-lab data is not released in this stage due to data-sharing policies. Upon acceptance, we will release all the required data including newly generated wet-lab data to enable full reproduction of the reported analyses.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Aureka AI OpenDDE Project. Folding, reasoning, and scaling with open-source drug discovery engine. *arXiv preprint arXiv:2607.03787*, 2026.

Nathaniel R. Bennett, Joseph L. Watson, Robert J. Ragotte, et al. Atomically accurate de novo design of antibodies with RFdiffusion. *Nature*, 649:183–193, 2025. doi: 10.1038/s41586-025-09721-5.

Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein–protein interactions using AlphaFold2. *Nature Communications*, 13(1):1265, 2022. doi: 10.1038/s41467-022-28865-w.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735.

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

Chai Discovery Team, Jacques Boitreaud, Jack Dent, Danny Geisz, Matthew McPartlon, Joshua Meier, Zhuoran Qiao, Alex Rogozhnikov, Nathan Rollins, Paul Wollenhaupt, and Kevin Wu. Zero-shot antibody design in a 24-well plate. *bioRxiv*, 2025. doi: 10.1101/2025.07.05.663018.

Ginger Chao, Wai L. Lau, Benjamin J. Hackel, Stephen L. Sazinsky, Shaun M. Lippow, and K. Dane Wittrup. Isolating and engineering human antibodies using yeast surface display. *Nature Protocols*, 1:755–768, 2006. doi: 10.1038/nprot.2006.94.

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

Roland L. Dunbrack, Jr. Rēs ipSAE loquuntur: What's wrong with AlphaFold's ipTM score and how to fix it. *bioRxiv*, 2025. doi: 10.1101/2025.02.10.637595.

M. Frank Erasmus, Daniel Bedinger, Elizabeth Hopkins, Ginger Ferguson, Justine Strickler, Christilyn P. Graff, Samantha R. Summers, Stacy L. Capehart, Joshua D. Slocum, Crystal Richardson, et al. A blinded, prospective benchmark of in silico antibody discovery anchored to experimental affinity and developability. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03238-6.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Zidek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021. doi: 10.1101/2021.10.04.463034.

Aric Huang, Wei Jin, Ahmed S. Fahad, Brooklyn K. Mussman, Grazia Paola Nicchia, Bharat Madan, Matheus Oliveira de Souza, J. Daniel Griffin, Jeffrey L. Bennett, Antonio Frigeri, et al. Strategies to screen anti-AQP4 antibodies from yeast surface display libraries. *Antibodies*, 11(2):39, 2022. doi: 10.3390/antib11020039.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-end full-atom antibody design. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17409–17429. PMLR, 2023.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Martin Pacesa, Lars Nickel, Christian Schellhaas, Jan Schmidt, Ekaterina Pyatova, Lena Kissling, et al. One-shot design of functional protein binders with BindCraft. *Nature*, 646(8084):483–492, 2025. doi: 10.1038/s41586-025-09429-6.

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

Amir Shanehsazzadeh, Julian Alverio, George Kasun, Simon Levine, Ido Calman, Jibran A. Khan, Chelsea Chung, Nicolas Diaz, Breanna K. Luton, et al. IgDesign: In vitro validated antibody design against multiple therapeutic antigens using inverse folding. *bioRxiv*, 2023. doi: 10.1101/2023.12.08.570889.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, et al. BoltzGen: Toward universal binder design. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689494.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988.

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.

Qilin Yu, Liangyue Guo, Xiayan Qin, Xikun Huang, Baihui Tian, Hongzhun Wang, Yu Liu, Yunzhi Lang, Di Wang, Zhouhanyu Shen, Jie Lin, and Mingchen Chen. High-affinity protein binder design via flow matching and in silico maturation. *bioRxiv*, 2026. doi: 10.64898/2026.01.19.700484.

Odin Zhang, Xujun Zhang, Haitao Lin, Cheng Tan, Qinghan Wang, Yuanle Mo, Qiantai Feng, Gang Du, Yuntao Yu, Zichang Jin, et al. ODesign: A world model for biomolecular interaction design. *arXiv preprint arXiv:2510.22304*, 2025.

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*, 2026. doi: 10.64898/2026.04.10.717613.

Wensi Zhu, Aditi Shenoy, Petras Kundrotas, and Arne Elofsson. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics*, 39(7):btad424, 2023. doi: 10.1093/bioinformatics/btad424.

APPENDIX CONTENTS

A Additional Experimental Results 14
A.1 Matched population and benchmark results 14
A.2 Binder separability and regional confidence 15
A.3 Filtering at lower binder prevalence 16
A.4 Sensitivity to ensemble size 17
A.5 Threshold transfer to unseen antigens 18

B Benchmark Construction 18
B.1 Display-based enrichment benchmark 18
B.2 IgDesign SPR benchmark 20
B.3 Literature and patent affinity benchmark 21
B.4 Assay definitions, constructs, and sequence diversity 21

C Experimental Details 22
C.1 Cofolding inference 22
C.2 Baseline implementation 22
C.3 Campaign-specific evaluation protocol 24
C.4 Threshold selection and fixed-budget evaluation 25

D RIC Implementation Details 26

E Related Work 27

A Additional Experimental Results

A.1 Matched population and benchmark results

Table 4 repeats the global-ranking evaluation on the same candidate population as Table 3. RIC/gRIC ranks ahead of NAC for all three predictors when candidates compete across antigens. This reversal therefore persists without changing the evaluated population. Each metric's global threshold is re-optimized on these candidates, not carried over from Table 2. The matched set excludes the mostly positive literature/patent cohort, which helps explain its lower absolute results.

Table 4: **Global filtering on the candidates used in Table 3.** The 2,400 candidates (944 binders, 1,456 non-binders) from 13 antigens are pooled and ranked with one shared threshold. This holds the candidate population fixed while changing the comparison from within to across antigens. Average ranks cover the seven reported measures. ESMFold2 excludes 107 out-of-memory complexes; *gRIC uses geometric contacts.

Metric	Recall @ 10% FPR	Recall @ 20% FPR	Recall @ 30% FPR	Recall @ $Q = 5\%$	Recall @ $Q = 20\%$	Utility $U^\star$	Precision @ $U^\star$	Average rank
AlphaFold 3								
Pair-iPTM	2.97	7.20	17.06	2.05	12.29	0.0257	19.27	7.29
Interface PAE ↓	3.50	6.14	12.08	2.97	9.85	0.0243	19.39	8.93
Antibody pLDDT	5.19	8.16	9.43	3.28	9.32	0.0355	23.93	4.93
Ranking score	2.75	6.14	11.12	3.43	10.14	0.0287	21.67	7.57
pDockQ	3.92	5.83	23.94	3.18	14.41	0.0312	28.33	4.29
pDockQ2	3.81	6.25	12.18	2.86	10.38	0.0260	18.64	8.21
ipSAE-d0res	3.81	6.57	11.86	1.80	10.81	0.0257	19.27	8.21
ipSAE-d0dom	4.13	6.36	11.76	1.59	10.70	0.0272	20.79	7.57
ipSAE-d0chn	4.34	7.31	11.65	2.22	10.49	0.0280	21.99	6.29
AntiConf	3.81	6.14	13.45	2.86	11.12	0.0263	20.83	6.71
RIC (Interface)	6.04	7.52	13.45	3.28	10.81	0.0407	29.08	**2.36**
NAC (Antibody)	5.08	8.05	9.32	2.97	9.00	0.0382	24.70	5.64
Protenix-v2								
Pair-iPTM	6.99	8.47	14.62	5.83	11.12	0.0498	40.00	2.93
Pair-gPDE ↓	5.61	6.99	11.23	4.34	10.17	0.0380	24.51	10.64
Interface PAE ↓	6.67	8.37	13.45	4.03	10.38	0.0464	34.07	7.07
Antibody pLDDT	6.99	8.37	10.81	4.45	9.53	0.0476	31.58	7.43
Ranking score	5.19	8.26	14.30	3.81	10.38	0.0346	26.63	9.93
pDockQ	6.25	7.31	10.17	4.24	8.47	0.0431	29.33	10.50
pDockQ2	6.89	8.47	13.24	3.81	9.96	0.0473	32.02	7.64
ipSAE-d0res	5.30	6.78	14.09	4.45	10.38	0.0365	32.45	8.93
ipSAE-d0dom	6.78	8.47	14.19	5.61	10.59	0.0473	33.16	5.07
ipSAE-d0chn	6.99	8.47	15.04	6.25	11.23	0.0505	39.88	2.36
AntiConf	6.46	7.31	13.24	3.50	10.17	0.0436	29.25	9.93
RIC (Interface)	9.00	14.09	14.94	5.72	14.41	0.0710	36.30	**1.86**
NAC (Antibody)	6.14	8.58	16.10	3.28	10.81	0.0458	30.09	6.71
ESMFold2								
Pair-iPTM	1.47	5.08	11.17	1.47	9.26	0.0250	18.26	10.86
Interface PAE ↓	4.40	6.21	11.51	4.06	10.05	0.0290	46.75	6.36
Antibody pLDDT	5.87	8.80	9.71	4.29	9.59	0.0427	26.47	4.86
pDockQ	5.19	10.50	18.06	3.61	13.54	0.0406	24.91	3.43
pDockQ2	4.18	5.87	11.40	4.06	10.05	0.0285	36.36	7.29
ipSAE-d0res	4.51	6.43	12.30	3.39	9.93	0.0292	22.40	7.50
ipSAE-d0dom	4.97	6.55	12.08	2.26	10.05	0.0324	25.00	6.43
ipSAE-d0chn	4.97	6.55	13.66	3.50	10.84	0.0328	25.73	5.14
AntiConf	4.18	5.98	15.58	4.06	9.93	0.0291	50.70	6.07
gRIC*	7.00	10.05	12.98	4.18	11.29	0.0488	28.46	**2.43**
NAC (Antibody)	4.74	8.47	17.49	4.06	9.82	0.0329	22.12	5.64

Table 5 similarly separates the campaign-specific results in Table 3 by benchmark. NAC leads on the display-based targets for all three predictors; the IgDesign results are more mixed, particularly for Protenix-v2.

Table 5: **Campaign-specific results from Table 3, separated by benchmark.** Antigens are evaluated separately, then aggregated within each benchmark. R10 is recall at 10% FPR (in percent), U is filtering utility, and ranks use all seven measures. IgDesign includes six eligible antigens; ESMFold2 excludes 107 out-of-memory display-based candidates.

Predictor	Candidates (P/N)	Best metric	RIC/gRIC rank / R10 / U	NAC rank / R10 / U	Antibody pLDDT rank
Display-based enrichment					
AlphaFold 3	1,400 (700/700)	NAC	5.64 / 18.43 / 0.0898	1.07 / 30.00 / 0.1626	3.07
Protenix-v2	1,400 (700/700)	NAC	6.71 / 15.29 / 0.0722	1.07 / 29.00 / 0.1430	2.21
ESMFold2	1,293 (642/651)	NAC	4.71 / 8.48 / 0.0378	1.00 / 28.53 / 0.1433	2.14
IgDesign SPR					
AlphaFold 3	1,000 (244/756)	NAC	11.57 / 12.68 / 0.0506	1.00 / 37.94 / 0.1957	2.43
Protenix-v2	1,000 (244/756)	Interface PAE	13.00 / 13.28 / 0.0583	9.57 / 44.09 / 0.2537	7.57
ESMFold2	1,000 (244/756)	NAC	11.00 / 15.29 / 0.0753	1.57 / 47.32 / 0.2897	4.14

A.2 Binder separability and regional confidence

Table 6 examines how well the predicted outputs distinguish binders from non-binders within an antigen. Confidence yields higher mean AUC than coordinates alone in all six benchmark–predictor comparisons, although combining the two views does not always improve it further.

Table 6: **Within-antigen binder separability from predicted coordinates and confidence.** Values are unweighted mean AUCs over the seven antigens in each benchmark, using five-fold within-antigen cross-fitting. All three predictors are evaluated on the same complete-ensemble candidates (1,293 display-based and 1,171 IgDesign pairs). The combined view joins coordinates and confidence. These views include non-CDR antibody pLDDT and omit predicted contact probabilities, so they differ from those in Figure 1. The matched literature/patent subset has no antigen with enough examples of both classes for within-antigen AUC.

Benchmark	Predictor	Coordinates	Confidence	Combined
Display-based enrichment	AlphaFold 3	0.811	0.855	0.847
	Protenix-v2	0.808	0.810	0.808
	ESMFold2	0.779	0.822	0.820
IgDesign SPR	AlphaFold 3	0.796	0.822	0.823
	Protenix-v2	0.817	0.828	0.831
	ESMFold2	0.825	0.829	0.838

Regional antibody confidence. The CDRs and binding interface are central to recognition, but their conformations and contacts can also be difficult to predict (Ünsal et al., 2026; Erasmus et al., 2026). Confidence focused on these regions may therefore be more sensitive to local modeling uncertainty than a broader signal from the antibody framework. Whole-antibody pLDDT mixes that signal with CDR confidence, while interface pLDDT also depends on predicted docking. This suggests a trade-off between direct relevance to binding and prediction reliability. To examine this possibility, Table 7 compares pLDDT over the predicted interface, the CDRs, the non-CDR residues, and the whole antibody, using the same candidates and ensemble aggregation throughout.

In five of six predictor–benchmark comparisons, mean AUC follows the same order: non-CDR, whole-antibody, CDR, then interface pLDDT. The exception is Protenix-v2 on IgDesign, where interface pLDDT ranks second and nearly matches NAC (0.785 versus 0.787). NAC has the highest mean AUC in all six comparisons; across individual antigens, it exceeds whole-antibody pLDDT in 28 of 42 comparisons and CDR pLDDT in 30 of 42. This ordering fits the proposed trade-off: CDRs and interface contacts matter most directly to binding, but their conformational variability and docking uncertainty make them harder to predict accurately. **Our observation.** For within-antigen filtering, confidence in the less directly involved but more reliably modeled non-CDR region provides a stronger signal than confidence in the harder-to-predict CDRs or predicted interface.

Table 7: **Within-antigen binder separability of regional pLDDT.** The interface score averages pLDDT over unique antibody and antigen residues in predicted 8-Å Cβ contacts (Cα when Cβ is unavailable); a sample without contacts scores zero. The other scores average pLDDT over the indicated antibody residues. Sample scores are aggregated by the median within a run and then over runs. Values are unweighted mean AUCs across seven antigens per benchmark, computed directly from the scores without fitting a classifier. All predictors use the same 2,464 complete-ensemble candidates (1,293 display-based and 1,171 IgDesign pairs). Best values in each row are bold and second-best values are underlined.

Benchmark	Predictor	Interface	CDR	Non-CDR	Whole antibody
Display-based enrichment	AlphaFold 3	0.569	0.600	**0.677**	<u>0.638</u>
	Protenix-v2	0.581	0.582	**0.669**	<u>0.627</u>
	ESMFold2	0.530	0.577	**0.630**	<u>0.601</u>
IgDesign SPR	AlphaFold 3	0.667	0.703	**0.757**	<u>0.735</u>
	Protenix-v2	<u>0.785</u>	0.772	**0.787**	0.776
	ESMFold2	0.703	0.706	**0.794**	<u>0.746</u>

A.3 Filtering at lower binder prevalence

New AI-design candidate pools may contain a smaller fraction of binders than the benchmark. To examine this setting, we resampled candidates within each antigen to approximately 5%, 10%, 15%, and 20% binder prevalence, then repeated fixed-budget filtering. The analysis uses the 1,935 candidates from six display-based and four IgDesign antigens with complete predictions from all three predictors and enough candidates to retain at least five binders per antigen at 5% prevalence. We repeated each resampling 500 times and used the same sampled candidates for every score. Table 8 compares NAC with the strong antibody pLDDT baseline from Table 3.

Table 8: **Within-antigen filtering at lower binder prevalence.** Each predictor column reports mean Precision / Recall (%) at $Q = 20\%$ over 500 resamples; random selection is evaluated with the same per-antigen budget. Binders or non-binders are subsampled within each antigen to reach the stated prevalence. The 10 antigen weights are fixed from the matched population before subsampling, with equal total weight for the two benchmarks.

Binder rate	Score	Random	AlphaFold 3	Protenix-v2	ESMFold2
5%	NAC	4.9 / 19.8	11.3 / 45.6	12.7 / 51.2	13.6 / 54.2
	Antibody pLDDT		11.2 / 44.8	11.6 / 46.6	12.8 / 50.8
10%	NAC	10.0 / 19.8	22.6 / 44.9	24.6 / 48.9	26.9 / 53.4
	Antibody pLDDT		21.4 / 42.6	23.3 / 46.3	25.6 / 50.8
15%	NAC	15.1 / 19.6	33.3 / 43.5	35.4 / 46.3	38.8 / 50.9
	Antibody pLDDT		30.9 / 40.5	34.6 / 45.3	37.3 / 48.9
20%	NAC	20.0 / 20.0	42.7 / 42.7	44.3 / 44.3	49.1 / 49.1
	Antibody pLDDT		40.0 / 40.0	42.6 / 42.6	48.0 / 48.0

At 5% binder prevalence, NAC recovers 45.6–54.2% of binders while testing 20% of candidates, compared with 19.8% recall under random selection. Its precision is 11.3–13.6%, more than twice the random expectation of 4.9%. NAC also exceeds whole-antibody pLDDT in mean recall at every predictor–prevalence combination shown. These subsets change class balance among observed candidates; they do not reproduce the sequence distribution of a new AI-design campaign.

A.4 SENSITIVITY TO ENSEMBLE SIZE

We tested whether smaller prediction ensembles preserve fixed-budget filtering performance. On the 2,400 campaign candidates with complete AlphaFold 3 and Protenix-v2 predictions, we compared two, six, and ten runs (seeds), retaining five samples per run. We also fixed ten runs and compared one, three, and five samples per run. RIC, NAC, antibody pLDDT, pair-iPTM, and ipSAE-d0chn were recomputed from the same prediction subsets before evaluating the same candidates and labels. The full ten-run, five-sample scores reproduce Table 3.

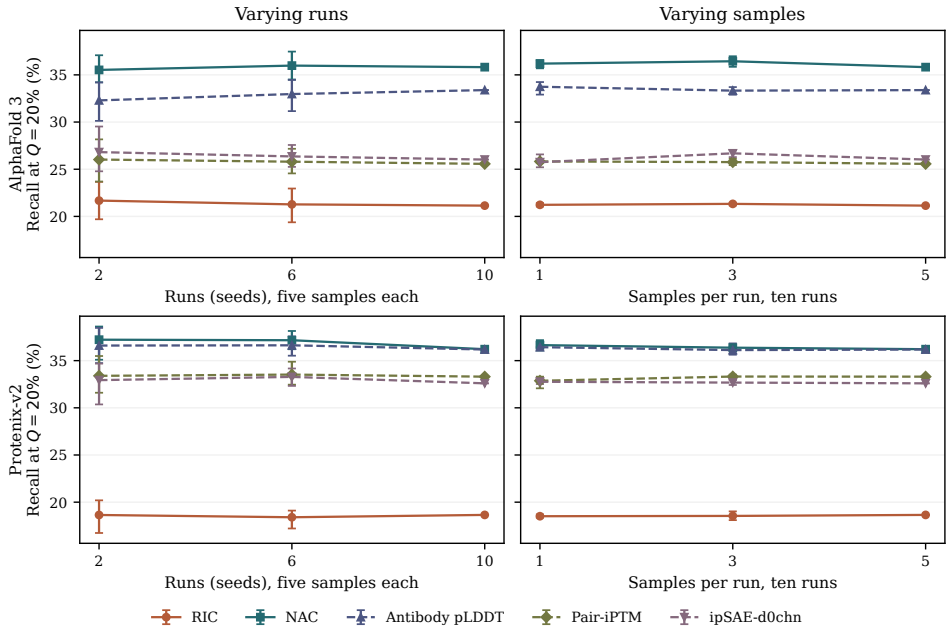


Figure 2: **Fixed-budget recall with smaller prediction ensembles.** Within-antigen Recall at $Q = 20\%$ for AlphaFold 3 (top) and Protenix-v2 (bottom). Points are means and bars span the selected subsets. The run analysis uses all 45 two-run pairs and 50 prespecified six-run subsets; the sample analysis uses three prespecified subsets at each reduced count. The full ensemble has ten runs and five samples per run. All curves use the same 2,400 campaign candidates; ESMFold2 is not included.

Figure 2 shows that NAC retains the highest mean Recall at $Q = 20\%$ among these scores at every tested budget. With two versus ten runs, its mean recall is 35.5% versus 35.8% for AlphaFold 3 and 37.2% versus 36.2% for Protenix-v2. Mean recall also changes little when the number of samples per run varies. Individual run choices matter, however: with two runs, RIC recall ranges from 19.7% to 23.7% for AlphaFold 3 and from 16.7% to 20.2% for Protenix-v2. Thus, smaller ensembles preserve much of the average fixed-budget signal on these candidates, but their rankings and threshold-dependent performance need not be identical.

A.5 THRESHOLD TRANSFER TO UNSEEN ANTIGENS

Table 9 tests whether a threshold chosen on labeled antigens can be applied to a new antigen. We hold out each of the 13 campaign antigens in turn. For each predictor and score, we choose one absolute cutoff that maximizes the weighted mean utility on the other 12 antigens, then apply it unchanged to the held-out antigen. We also calibrate using only the other antigens from the same assay benchmark. The held-out labels are used only to evaluate the selected cutoff. Both calibration strategies use the score directions, threshold tie rules, and utility definition of Table 3.

Table 9: **Leave-one-antigen-out threshold transfer.** Weighted mean held-out utility when the cutoff is selected on all other antigens or only on other antigens from the same assay benchmark. The oracle column uses each held-out antigen's own labels to select its best cutoff and repeats $U^\star$ from Table 3; it is not a prospective estimate. The final column counts held-out antigens on which the all-other-antigen cutoff selects at least one candidate, out of 13. AlphaFold 3 and Protenix-v2 use 2,400 candidates each; ESMFold2 uses its 2,293 successfully predicted candidates.

Predictor	Score	All other	Same assay	Oracle	Selected / 13
AlphaFold 3	RIC	0.0490	0.0503	0.0702	12
	NAC	0.0000	0.0015	0.1791	1
	Antibody pLDDT	0.0138	0.0246	0.1367	11
	Pair-iPTM	0.0548	0.0427	0.0807	12
	ipSAE-d0chn	0.0122	0.0137	0.0539	10
Protenix-v2	RIC	0.0322	0.0300	0.0652	13
	NAC	0.0085	0.0423	0.1983	2
	Antibody pLDDT	0.0185	0.0412	0.1837	2
	Pair-iPTM	0.0920	0.1197	0.1674	4
	ipSAE-d0chn	0.0837	0.0993	0.1682	4
ESMFold2	gRIC	0.0080	0.0069	0.0566	9
	NAC	0.0047	0.0086	0.2165	2
	Antibody pLDDT	0.0133	0.0475	0.1889	2
	Pair-iPTM	0.0000	0.0029	0.1385	2
	ipSAE-d0chn	0.0012	0.0043	0.1299	5

The strong within-antigen NAC results do not imply that one absolute NAC cutoff transfers across antigens. Calibration on all other antigens selects candidates on only one or two held-out antigens, depending on the predictor, and its held-out utility is well below the per-antigen oracle. Reusing an absolute cutoff tests whether score scales are comparable across antigens, much like across-antigen filtering. RIC/gRIC retains a larger share of its per-antigen oracle utility than NAC for all three predictors, consistent with its relative strength across antigens. This does not diminish NAC's within-antigen ranking result: **within-antigen candidate ranking can remain informative even when absolute NAC scores shift between targets**. Limiting calibration to the same assay benchmark improves several scores, including Protenix-v2 NAC, but leaves a substantial gap. **Our observation:** The best absolute threshold depends on the antigen for every metric tested; cutoffs calibrated on other antigens fall short of per-antigen optimization. This may reflect antigen-dependent score biases, potentially shaped by the predictors' training data. Fixed-budget top-Q selection avoids an absolute cutoff while retaining NAC's strong within-antigen ranking signal: at $Q = 20\%$, it recovers 35.82%, 36.21%, and 39.22% of binders with AlphaFold 3, Protenix-v2, and ESMFold2, respectively (Table 3).

B BENCHMARK CONSTRUCTION

B.1 DISPLAY-BASED ENRICHMENT BENCHMARK

Experimental readout. The benchmark combines two display platforms. Phage display was used for CD98hc, EGFR, PLXNA1 ectodomain, and TNFRSF23. Antibody fragments displayed on phage were incubated with antigen-coated magnetic beads; retained phage were recovered, amplified, and

carried into the next selection round. Yeast display was used for TfR, SARS-CoV-2 spike S2, and GPC3. Antibody fragments displayed on yeast were selected with labeled antigen using magnetic and fluorescence-activated sorting (Chao et al., 2006; Huang et al., 2022). Both platforms were sequenced across rounds to track clone enrichment.

For the four phage-display targets, 5, 4, 2, and 1 μg of antigen were used in successive panning rounds. For the three yeast-display targets, the final sorting round used antigen at 1 nM.

Filtering, quality control, and label construction. For each antigen, reads from all available rounds are translated, sequence quality filtered, exact-deduplicated, and grouped by antibody sequence. For clone a in round r, its normalized abundance is $f_{a,r} = n_{a,r} / \sum_b n_{b,r}$, where $n_{a,r}$ is the number of retained reads for clone a. We calculate an enrichment signal from the normalized frequency change between the early and later selection stages, $E_a = \log_2[(f_{a,\text{late}} + \varepsilon)/(f_{a,\text{early}} + \varepsilon)]$, where ε is a fixed read-level pseudocount. Candidates must pass sequence and read-quality filters and have a frequency trajectory consistent with the sign of E_a. For each antigen, the up to 100 late-enriched candidates with the largest E_a form the positive set. Negative controls are drawn from the baseline or earliest-round pool and show depletion or no enrichment in later rounds; early low-enrichment candidates are used only when needed. Later rounds use lower antigen inputs, so E_a reflects survival under increasing selection stringency.

Composition. The display-based enrichment benchmark contains 1,400 antibodies across seven antigens, with 100 positives and 100 negatives for each antigen. The included antigen systems are CD98hc, EGFR, GPC3, PLXNA1 ectodomain, SARS-CoV-2 spike S2, transferrin receptor, and TNFRSF23.

Library depth and sequence diversity. Table 10 reports the scale of the source libraries before benchmark subsampling. For the six experiments supplied as per-sequence count tables, an input record is one sequence entry carried into AIRR/full-VH processing. "QC-retained read count" is the sum of duplicate-aware read counts after productive, in-frame IGH and amino-acid quality control. "Unique full VH" is the number of exact full-length VH amino-acid sequences retained within that round. Consequently, the antigen-level total in the final column counts sequence–round observations rather than deduplicating a clone observed in multiple rounds. Across the seven antigens, individual rounds contain 56,284–1,878,482 unique full-VH sequences and approximately 2.34×10^5–4.57×10^6 retained reads, substantially exceeding the 200 candidates per antigen included in the benchmark. In aggregate, the source libraries comprise 26,540,032 input sequence records, a QC-retained read count of 36,235,220, and 15,226,144 unique sequence–round observations.

Table 10: **Source-library depth for the display-based enrichment benchmark.** Input sequence records are records entering full-VH quality control. QC-retained read count sums duplicate-aware counts after quality control; unique full VH denotes exact amino-acid sequences within a round. Antigen-level totals sum round-level values, so the final-column total is a count of sequence–round observations, not an across-round deduplicated clone count. A dash marks the pre-selection baseline, a total, or a yeast-display round for which the antigen concentration was not provided. Phage-display inputs are antigen masses; yeast-display inputs are antigen concentrations.

Antigen	Round	Antigen input	Input sequence records	QC-retained read count	Unique full VH
CD98hc	R1	5 μg	3,019,877	2,237,462	1,878,482
	R2	4 μg	288,646	247,728	162,207
	R3	2 μg	2,085,258	2,095,645	1,666,243
	R4	1 μg	417,666	1,845,768	126,526
	Total	–	5,811,447	6,426,603	3,833,458
EGFR	R1	5 μg	1,169,527	924,008	689,821

Continued on next page

Table 10 continued

Antigen	Round	Antigen input	Input sequence records	QC-retained read count	Unique full VH
	R2	4 μg	1,243,577	1,576,915	774,560
	R3	2 μg	2,503,238	4,570,804	1,296,785
	R4	1 μg	1,276,218	2,472,962	605,591
	Total	–	6,192,560	9,544,689	3,366,757
GPC3	R1	–	705,045	542,241	324,749
	R2	–	800,666	903,710	436,174
	R3	–	166,630	389,192	56,284
	R4	1 nM	951,617	632,421	533,960
	Total	–	2,623,958	2,467,564	1,351,167
PLXNA1 ECD	R1	5 μg	1,315,443	802,537	676,591
	R2	4 μg	954,430	1,092,403	613,320
	R3	2 μg	863,064	1,153,048	500,881
	R4	1 μg	813,671	1,247,029	414,013
	Total	–	3,946,608	4,295,017	2,204,805
Spike S2	R1	–	950,482	1,816,251	499,915
	R2	–	255,387	905,852	109,195
	R3	–	562,782	233,744	132,920
	R4	1 nM	366,993	1,541,237	134,778
	Total	–	2,135,644	4,497,084	876,808
TfR	R0	–	127,668	445,113	100,009
	R1	–	651,716	926,675	500,796
	R2	–	550,195	1,272,308	343,741
	R3	1 nM	577,530	1,287,790	346,152
	Total	–	1,907,109	3,931,886	1,290,698
TNFRSF23	R1	5 μg	966,612	742,951	512,726
	R2	4 μg	1,344,649	1,709,197	912,832
	R3	2 μg	925,774	1,470,200	534,134
	R4	1 μg	685,671	1,150,029	342,759
	Total	–	3,922,706	5,072,377	2,302,451
All antigens	Total	–	26,540,032	36,235,220	15,226,144

B.2 IGDESIGN SPR BENCHMARK

Source experiment. IgDesign conditions antibody inverse folding on a native antibody–antigen complex backbone and framework sequence, then redesigns either HCDR3 or all three heavy-chain CDRs. The original study generated 100 HCDR3 and 100 HCDR123 designs for each of eight therapeutic antigens and screened them by SPR (Shanehsazzadeh et al., 2023). Seven unblinded antigen systems have sufficient released sequence and replicate-level binding data for this study.

Replicate-consistent labels and sequence reconstruction. We retain a designed antibody as positive only when every available replicate reports binding and as negative only when no replicate reports binding; both classes require at least two replicates. Positive-control antibodies and rows with discordant or incomplete replicate calls are excluded. The released K_D value is the mean of quantifiable replicate measurements; it is not the threshold used for the binary binder label. Full variable domains are reconstructed by grafting the released designed HCDRs into the corresponding reference scaffold, while preserving the matched light chain and antigen construct. The resulting benchmark contains 249 binders and 922 non-binders.

B.3 LITERATURE AND PATENT AFFINITY BENCHMARK

Source collection and affinity labels. We extract antibody sequences, antigen identities, experimental constructs, and direct equilibrium affinity measurements from primary papers and patents. Positive labels require a monovalent SPR, biolayer interferometry (BLI), or isothermal titration calorimetry (ITC) measurement with $K_D \leq 10$ nM; negative labels require $K_D \geq 1{,}000$ nM. EC_{50}- or IC_{50}-only measurements, cell-based readouts, competition-only experiments, and ambiguous or conflicting records are excluded.

Construct matching. Each retained record must support the antibody variable-domain sequence and an antigen construct suitable for cofolding. We prefer the exact experimental antigen sequence. When the source specifies a standard extracellular domain, receptor-binding domain, soluble construct, or explicit residue range without printing its full sequence, we map that construct to the corresponding reference sequence. Records lacking a defensible construct assignment are discarded. The final benchmark contains 1,613 pairs: 1,510 binders and 103 non-binders across 165 antigen identities. The literature/patent benchmark contributes to portfolio selection but is excluded from campaign selection because its antigen-level negative support is too sparse for the reported low-FPR endpoints.

Per-antigen composition. Table 11 reports the 13 antigens in the campaign-selection population. These comprise all seven display-based enrichment antigens and the six IgDesign antigens that satisfy the minimum positive and negative support requirements.

Table 11: Target-level composition of the display-based enrichment and IgDesign benchmarks used for campaign selection.

Benchmark	Antigen	Positive	Negative
Display-based enrichment	CD98hc	100	100
	EGFR	100	100
	GPC3	100	100
	PLXNA1 ectodomain	100	100
	SARS-CoV-2 spike S2	100	100
	Transferrin receptor	100	100
	TNFRSF23	100	100
IgDesign SPR	ACVR2B	13	188
	C5	29	164
	FXI	33	69
	IL-36R	28	137
	TNFRSF9	27	149
	TSLP	114	49

B.4 ASSAY DEFINITIONS, CONSTRUCT STATISTICS, AND SEQUENCE DIVERSITY

Assay and label definitions. The three benchmarks deliberately provide different kinds of experimental evidence (Table 12). The display-based enrichment labels capture selection-associated enrichment; the IgDesign labels retain only unambiguous replicate SPR calls; and the literature/patent collection uses widely separated direct-affinity regimes. This diversity is useful for evaluating whether a structural filtering metric is tied to one experimental setting.

Construct statistics. Figure 3 summarizes the length profiles of the exact cofolding input sequences using their median, interquartile range, and observed range. Heavy-chain CDR3s are assigned by IMGT numbering with ANARCI (Dunbar & Deane, 2016). The literature/patent collection spans a substantially broader range of antigen sequences and contains both paired Fv and heavy-chain-only antibodies; the display-based benchmark is heavy-chain-only and IgDesign is paired Fv by design.

Table 12: **Assay and label definitions.** Display-based enrichment labels are based on round-wise enrichment under progressively lower antigen inputs.

Benchmark and evidence	Positive and negative definition	Exclusions
Display-based enrichment Yeast or phage display with successive selection rounds and a sequencing readout.	**Positive:** up to 100 late-enriched candidates with the largest normalized-frequency log-fold change. **Negative:** baseline or earliest-round sequences depleted or not enriched later, supplemented when necessary by early low-enrichment candidates.	Invalid translations, low-quality reads, exact duplicates, and over-represented near-identical families.
IgDesign SPR Replicate SPR measurements released with the original inverse-folding study.	**Positive:** binding in every available replicate. **Negative:** no binding in any replicate. Both classes require at least two measurements.	Positive controls, discordant replicate calls, and incomplete replicate data.
Literature/patent affinity Direct K_D measurements by SPR, BLI, or ITC reported in papers and patents.	**Positive:** $K_D \leq 10$ nM. **Negative:** $K_D \geq 1{,}000$ nM.	EC_{50}/IC_{50}-only, cell-based, competition-only, ambiguous, conflicting, or construct-unresolved records.

Diversity and redundancy. Table 13 reports both exact sequence uniqueness and sequence-identity clusters. Because the display-based enrichment benchmark contains only heavy chains, heavy-chain clustering is the common antibody representation across all datasets. We cluster heavy chains independently within each dataset at 90% sequence identity and 90% coverage using MMseqs2 (Steinegger & Söding, 2017); antigen constructs are clustered at 50% identity and 80% coverage after exact deduplication. No exact heavy-chain sequence is shared between any pair of datasets. Figure 4 complements these summary statistics by showing the full heavy-chain cluster-size spectrum and the positive fraction within each cluster-size bin.

Table 13: **Diversity and redundancy of cofolding constructs.** Antigen clustering is applied to unique construct sequences; the largest antigen cluster contains two constructs in the literature/patent dataset and one in each other dataset. Exact VH–VL pairs number 1,361, 1,171, and 1,608, respectively.

Dataset	Unique VH	VH clusters (90%/90%)	Largest VH cluster	Antigen constructs / clusters (50%/80%)
Display-based enrichment	1,361	632	52 (3.7%)	7 / 7
IgDesign SPR	1,171	22	201 (17.2%)	7 / 7
Literature/patent affinity	1,453	723	38 (2.4%)	165 / 160

C Experimental Details

C.1 Cofolding inference

Every antibody–antigen pair is evaluated with the same nominal inference budget for each co-folding predictor. MSAs are enabled, structural templates are disabled, and the MSA and sequence inputs remain fixed across seeds and samples. We run 10 seeds and draw 5 samples from each seed, producing at most 50 predictions per pair and predictor. Table 14 reports both the shared prediction budget and the predictor-specific parameters.

C.2 Baseline implementation

Predictor-native baselines. We read ranking confidence and pair-iPTM directly from the predictor output. For complexes with multiple antibody or antigen chains, pair-iPTM is the largest antibody–antigen entry in the reported chain-pair matrix. Antibody pLDDT is averaged over antibody residues. Interface PAE is the mean of both PAE directions over predicted antibody–antigen Cβ contacts within

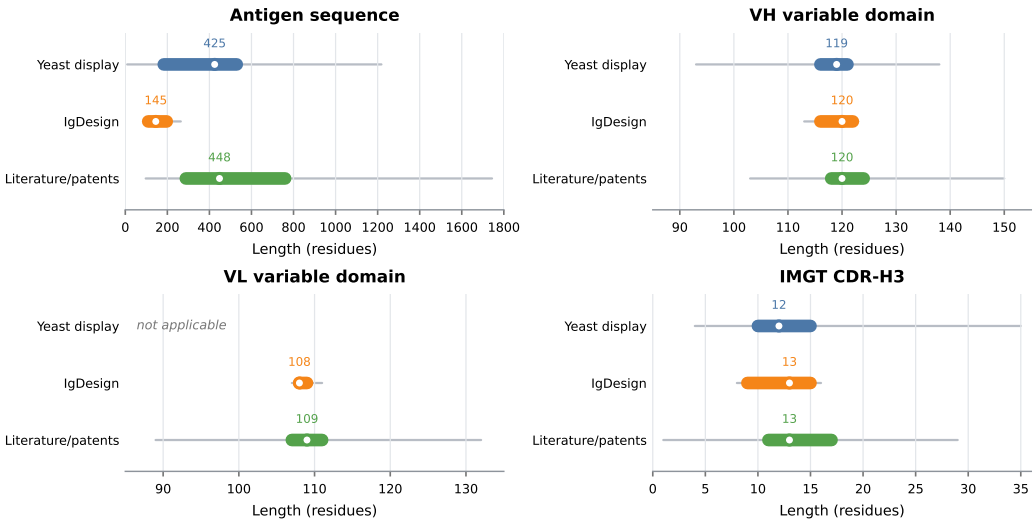


Figure 3: Length profiles of the exact cofolding inputs. Thick colored segments show the interquartile range, central dots the median, and thin lines the observed range. Antigen, heavy-chain variable-domain (VH), light-chain variable-domain (VL), and IMGT CDR-H3 lengths are measured in residues. VL is absent from the heavy-chain-only display-based enrichment benchmark; IgDesign is paired Fv, whereas the literature/patent collection contains both paired Fv and heavy-chain-only antibodies.

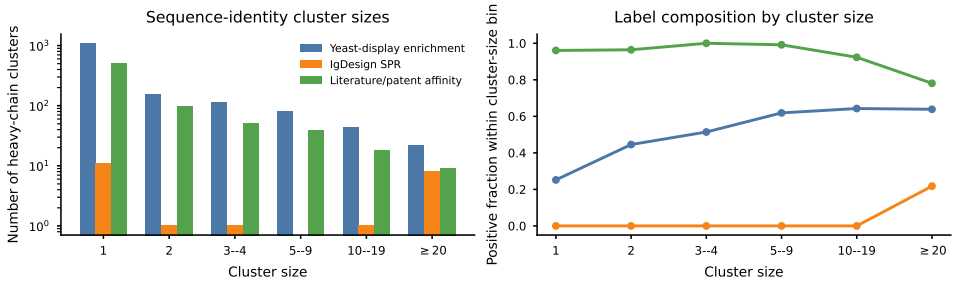


Figure 4: Heavy-chain sequence-identity clusters at 90% identity and 90% coverage. Left: number of clusters in each size bin. Right: candidate-level positive fraction within clusters in each size bin.

Table 14: **Cofolding inference configuration for each antibody–antigen pair.** The first five rows are shared across predictors. Remaining rows report predictor-specific settings used at inference, including official defaults when the runner did not override them. MSA limits denote model-input caps after MSA preprocessing; blank cells indicate that no corresponding control applies.

Setting	AlphaFold 3	Protenix-v2	ESMFold2
Unpaired MSA	enabled	enabled	enabled
Structural templates	disabled	disabled	disabled
Random seeds	10	10	10
Diffusion samples per seed	5	5	5
Maximum predictions per pair	50	50	50
Predictor-specific inference settings			
Recycling cycles	10	10	
Diffusion or sampling steps	200	200	100
Structure-update loops			20
Attention implementation	XLA		
Numerical precision	mixed bfloat16; outputs fp32	mixed bfloat16; confidence fp32	mixed: ESMC bfloat16; folding fp32
Language-model dropout			0.30
MSA rows retained for inference	total $\leq$ 16,384; paired $\leq$ 600/species	total $\leq$ 16,384; chunk size 2,048	depth $\leq$ 1,024; sequences $\leq$ 16,384
MSA column-mask rate			0.10

8 Å (Cα for glycine). A prediction without such a contact receives the largest cross-partner PAE. PDE/gPDE is read from the corresponding predictor output. Lower-is-better scores are sign-reversed only during threshold evaluation.

Published interface scores. The pDockQ, pDockQ2, ipSAE, and AntiConf calculations follow their authors' reference implementations. pDockQ uses the number of 8 Å Cβ contacts and mean interface pLDDT. pDockQ2 applies its PAE-weighted interface transform in both partner directions and retains the larger score. Each ipSAE variant is computed independently for every antibody–antigen chain pair, retaining the maximum over both PAE directions and eligible chain pairs. ipSAE-d0res adapts d_0 for each aligned residue, ipSAE-d0dom uses the number of residues in the low-PAE interface footprint, and ipSAE-d0chn uses the complete chain-pair length; all three use a PAE cutoff of 10 Å. AntiConf is $0.3\,\mathrm{pDockQ2}_{\mathrm{Ag}} + 0.7\,\mathrm{pTM}$, where the first term uses antigen as the aligned partner.

Ensemble aggregation. Predictor-specific chain indices are converted to a common antibody and antigen chain-role representation before scoring. Unless a method defines its own ensemble reducer, its per-prediction values are summarized first within each run and then across runs:

$$M = \underset{s=1}{\overset{10}{\mathrm{median}}} \left[\underset{k=1}{\overset{5}{\mathrm{median}}}\, r_{s,k} \right]. \tag{7}$$

This gives every baseline the same 10-seed, 5-sample evidence budget while preventing one unusually favorable sample from dominating a candidate score.

C.3 Campaign-specific evaluation protocol and quality controls

Table 3 evaluates the decision faced within a single-antigen design campaign: candidates are compared only with other candidates designed against the same antigen. We define a stratum as a canonical (dataset, antigen) pair and require every included stratum to contain at least 10 positives and 40 negatives. The negative-support requirement is particularly important for the low-FPR endpoints: with fewer than 10 negatives, a single false positive changes empirical FPR by more than 10 percentage points, while 40 negatives provide four negative observations within the R10 operating range.

The primary campaign-specific population contains the display-based enrichment and IgDesign benchmarks. All seven display-based enrichment strata are eligible, contributing 1,400 candidates

(700 positive and 700 negative). Six IgDesign strata are eligible, contributing 1,000 candidates (244 positive and 756 negative); IL17A is excluded because it contains only five positives. The final population therefore contains 13 strata and 2,400 candidates (944 positive and 1,456 negative). The literature/patent affinity domain is excluded from this analysis because its eight mixed-label strata contain only 4, 1, 1, 6, 1, 6, 2, and 1 negatives, respectively. These strata cannot resolve the reported 10–30% FPR operating points.

To prevent either assay domain from dominating the summary, display-based enrichment and IgDesign each receive one half of the total weight. Within each domain, strata receive square-root information weights. For stratum t, with P_t positives and N_t negatives, define

$$I_t = \frac{2P_tN_t}{P_t + N_t}, \qquad w_t = \frac{1}{2}\,\frac{\sqrt{I_t}}{\sum_{u:\, d(u)=d(t)} \sqrt{I_u}}, \tag{8}$$

where $d(t) \in \{\text{display-based enrichment}, \text{IgDesign}\}$. Thus $\sum_t w_t = 1$. This sublinear weighting gives better-supported campaigns more influence without allowing the largest campaign to dominate the result.

Every endpoint is computed independently within each stratum and then aggregated as $\sum_t w_t E_t$. In particular, Recall@$U^\star$ and Precision@$U^\star$ are weighted means of the values attained at each stratum's own utility-maximizing threshold; they are not computed by pooling selected counts across strata. A stratum whose utility optimum selects no candidate contributes zero recall and precision. Average rank is computed with ties over the eight reported endpoints after aggregation. All metrics use identical eligible candidates, weights, score directions, threshold tie rules, and finite-score coverage requirements.

C.4 Threshold selection and fixed-budget evaluation

Threshold search space. After ensemble aggregation, metric m assigns one finite score M_i to each candidate in $\mathcal{D}$, with larger values indicating stronger evidence. Because a candidate is selected when $M_i \geq \sigma$, the selected set changes only when σ crosses an observed score. It is therefore sufficient to search

$$\Sigma(m, \mathcal{D}) = \{+\infty\} \cup \{M_i : i \in \mathcal{D}\}. \tag{9}$$

This search space enumerates every distinct selection obtainable by thresholding. The value $+\infty$ gives the empty selection; lowering σ through the distinct observed scores produces a nested sequence of increasingly inclusive selections, ending with the full candidate set at the smallest score. Thresholds strictly between adjacent score levels produce the same selection and need not be evaluated, while candidates with equal scores necessarily enter together. At each threshold we compute TP, FP, TN, and FN, followed by the Wilson bounds and utility in Definitions 1–2. We choose the threshold with maximum utility; exact utility ties favor lower FPR, then higher recall, and finally the stricter threshold.

Utility-maximizing thresholds. Let $\mathcal{D}_t$ denote the candidates in campaign stratum t, and let $\mathcal{T}$ contain the class-supported campaign strata defined in Appendix C.3. The portfolio-wide and campaign-specific utilities are

$$\begin{aligned} U^\star_{\text{portfolio}}(m) &= \max_{\sigma \in \Sigma(m,\mathcal{D})} U(m, \sigma; \mathcal{D}), \\ U^\star_{\text{campaign}}(m) &= \sum_{t \in \mathcal{T}} w_t \max_{\sigma_t \in \Sigma(m,\mathcal{D}_t)} U(m, \sigma_t; \mathcal{D}_t). \end{aligned} \tag{10}$$

Portfolio-wide filtering pools all candidates and uses one shared threshold, testing cross-antigen score calibration. Campaign-specific filtering repeats the threshold search within each stratum, testing how well a metric ranks antibodies when the antigen is fixed. Portfolio-wide precision is evaluated at the maximizing shared threshold $\widehat{\sigma}$. If $\widehat{\sigma}_t$ maximizes utility in campaign t, campaign-specific precision is

$$P_{\text{campaign}}@U^\star = \sum_{t \in \mathcal{T}} w_t P_t(\widehat{\sigma}_t), \tag{11}$$

where $P_t(\widehat{\sigma}_t)$ is set to zero when the utility-maximizing threshold selects no candidate.

Sensitivity to the FPR penalty. We repeated the threshold searches with $f_{\text{safe}} = 0.10$ fixed and f_{severe} increased from 0.30 to 0.40 and 0.50. RIC (gRIC for ESMFold2) retained the highest utility in portfolio-wide filtering, and NAC retained the highest utility in campaign-specific filtering, for all three predictors at both alternative settings. Relaxing the penalty generally increased recall but also admitted more false positives. For example, NAC's weighted FPR at its utility-maximizing thresholds rose from approximately 8% to 15% when f_{severe} increased from 0.30 to 0.50. The main analysis retains $f_{\text{severe}} = 0.30$ to emphasize low-FPR filtering.

Recall at a fixed false-positive rate. This evaluation asks how many binders can be retained while allowing at most a specified fraction α of non-binders through the filter. Among all attainable thresholds satisfying that empirical FPR limit, we report the largest recall. For portfolio-wide filtering this is

$$\text{Recall}_{\text{portfolio}} \,@\, \text{FPR}\, \alpha = \max_{\substack{\sigma \in \Sigma(m,\mathcal{D}) \\ F(\sigma;\mathcal{D}) \le \alpha}} R(\sigma;\mathcal{D}). \tag{12}$$

Campaign-specific filtering performs the same constrained search separately within each stratum and then applies the fixed weights in Equation 8:

$$\text{Recall}_{\text{campaign}} \,@\, \text{FPR}\, \alpha = \sum_{t \in \mathcal{T}} w_t \max_{\substack{\sigma_t \in \Sigma(m,\mathcal{D}_t) \\ F_t(\sigma_t;\mathcal{D}_t) \le \alpha}} R_t(\sigma_t;\mathcal{D}_t). \tag{13}$$

We use $\alpha \in \{0.10, 0.20, 0.30\}$. Unlike utility, which uses Wilson bounds to account for finite-sample uncertainty, this operating-point analysis constrains the observed FPR directly. The empty selection has $F = R = 0$ and is always feasible. If every nonempty selection exceeds the FPR limit, the reported recall is therefore zero; in campaign-specific evaluation, that stratum contributes zero with its prespecified weight.

Recall at a fixed testing budget. Fixed-FPR evaluation constrains screening errors, whereas fixed-budget evaluation constrains the number of candidates that can be tested. For a testing fraction q, let $K_q = \lfloor qN \rfloor$ and retain the K_q highest-ranked candidates. This is a top-K_q decision rather than a search for the utility-maximizing threshold. If the boundary cuts through a tied score group, let $\mathcal{H}_q$ contain candidates strictly above the boundary and $\mathcal{E}_q$ those tied at the boundary, and define

$$\rho_q = \frac{K_q - |\mathcal{H}_q|}{|\mathcal{E}_q|}. \tag{14}$$

The expected number of retained positives under label-independent random tie-breaking is

$$\text{TP}_q = \sum_{i \in \mathcal{H}_q} Y_i + \rho_q \sum_{i \in \mathcal{E}_q} Y_i, \qquad \text{Recall}\,@Q = q = \frac{\text{TP}_q}{\sum_i Y_i}. \tag{15}$$

We report $q \in \{0.05, 0.20\}$. In portfolio-wide filtering, all candidates compete for a single testing budget. In campaign-specific filtering, the same fraction is retained independently within each stratum and recall is aggregated with the fixed campaign weights. When tied scores cross the budget boundary, the expression above is the expected recall under label-independent random tie-breaking, preventing file order from determining the result. Recall is defined as zero when $K_q = 0$.

D RIC Implementation Details

Contact forecast and PAE agreement. For each reference run r, we average both directions of predicted contact probability across its samples:

$$\bar{C}_r(a,b) = \frac{1}{|\mathcal{K}_r|} \sum_{j \in \mathcal{K}_r} \frac{C_{rj}^{\text{Ab}\to\text{Ag}}(a,b) + C_{rj}^{\text{Ag}\to\text{Ab}}(b,a)}{2}, \qquad h_r(a) = \max_{b \in \mathcal{G}} \bar{C}_r(a,b). \tag{16}$$

For sample j in run r and sample k in another run t, their bidirectional PAE agreement is

$$A_{rtjk}(a,b) = \frac{1}{2}\left[\frac{\tau}{\tau + P_{rj}^{\text{Ab}\to\text{Ag}}(a,b) + P_{tk}^{\text{Ab}\to\text{Ag}}(a,b)} + \frac{\tau}{\tau + P_{rj}^{\text{Ag}\to\text{Ab}}(b,a) + P_{tk}^{\text{Ag}\to\text{Ab}}(b,a)} \right], \qquad \tau = 16 \text{ Å}. \tag{17}$$

An exploratory sensitivity analysis compared PAE scales of 4, 8, and 16 Å. The 16 Å scale performed best overall, although 8 Å led on some subset-level measures; the differences were generally modest. We therefore used 16 Å for all predictors.

Tie-aware rank weights. For a reference run r and CDR residue a, antigen residues are ranked by $\bar{C}_r(a, b)$. With rank budget $m = \min(8, \lfloor |\mathcal{G}|/2 \rfloor)$, the high-contact weights w^H_{ra} allocate unit mass to ranks $1, \ldots, m$, and the comparison weights w^L_{ra} allocate unit mass to ranks $m + 1, \ldots, 2m$. If equal contact values cross either boundary, the total mass assigned to the affected ranks is divided equally among all residues in the tied block. The resulting weights therefore do not depend on the arbitrary order of tied residues.

Short antigens. When $2 \leq |\mathcal{G}| < 16$, the adaptive definition of m contracts both rank regions so that they fit within the available antigen residues; if the antigen length is odd, the final unpaired residue is unused. For the limiting case $|\mathcal{G}| = 1$, the sole residue receives unit weight in both regions, and the local contrast $\Delta_{rtjk}(a)$ in Equation 5 is exactly zero.

Zero contact support. If $h_r(a) = 0$ for every CDR residue in a reference run, the normalized weights $h_r(a)/\sum_{a'} h_r(a')$ are not evaluated and all contributions from that reference run are defined as zero. This remains a valid zero-valued observation in the median hierarchy rather than being treated as missing data.

Geometric-contact approximation. For ESMFold2, which does not provide contact probabilities, gRIC replaces both contact directions in Equation 16 with the same binary contact map: antibody and antigen residue representative atoms are in contact when their predicted distance is less than 8 Å. The PAE agreement, rank weights, contact-support weighting, and median aggregation remain those of RIC.

Chain mapping and ensemble completeness. The antibody and antigen scopes are defined by the chain roles supplied for each candidate. A single-chain antibody is scored using that chain alone; no placeholder light-chain residues are introduced. Every declared chain must map one-to-one to the predictor tokens and input sequence. Across the ensemble, all predictions must preserve the same ordered antibody residues, antigen residues, and CDR annotation, and all contact and PAE tensors must have matching finite dimensions. The experiments in this paper require the full 10-run, 5-sample ensemble. A missing prediction, a missing declared chain, or an inconsistent residue mapping invalidates the candidate score. We do not aggregate predictions from an incomplete ensemble.

E Related Work

Computational antibody and binder design. Modern antibody and protein-binder design spans backbone generation, joint sequence–structure generation, inverse folding, and predictor-guided sequence optimization. RFantibody adapts diffusion-based structure generation to antibody scaffolds and CDR loops, whereas BindCraft optimizes binder sequences through a structure-prediction network (Bennett et al., 2025; Pacesa et al., 2025). Chai-2, BoltzGen, PPIFlow, and ODesign broaden this space toward zero-shot antibody design, multimodal binder generation, and joint modeling of sequence, structure, and molecular interactions (Chai Discovery Team et al., 2025; Stark et al., 2025; Yu et al., 2026; Zhang et al., 2025). Despite their architectural differences, these methods share a practical dependence on predicted complex structures and post-generation filtering to select candidates for experiment.

Co-folding models for antibody–antigen complexes. Modern co-folding methods have progressively expanded from protein-complex prediction to general all-atom modeling of biomolecular interactions. AlphaFold-Multimer established joint prediction of multichain protein assemblies and introduced interface-aware confidence through ipTM (Evans et al., 2021). More recent all-atom systems, including AlphaFold 3, Protenix-v2, Chai-1, Boltz-2, and OpenDDE, extend this paradigm to heterogeneous biomolecular complexes and provide increasingly rich structural, pairwise, and interface-level outputs (Abramson et al., 2024; Zhang et al., 2026; Chai Discovery, 2024; Passaro et al., 2025; Aureka AI OpenDDE Project, 2026). Within this family, Boltz-2 further couples structure

prediction with binding-affinity prediction (Passaro et al., 2025). In parallel, ESMFold2 follows a protein-language-model-centered route, using learned evolutionary representations for efficient prediction of protein structures and complexes (Candido et al., 2026). Together, these models provide increasingly diverse predictions of antibody–antigen geometry and confidence, creating a rich source of structural evidence for downstream candidate prioritization.

Structural confidence metrics. The native ipTM score summarizes confidence in the relative placement of chains, but its whole-chain normalization can dilute a localized interface. pDockQ combines interface pLDDT with the number of predicted contacts to estimate DockQ, and pDockQ2 incorporates interchain PAE to score individual interfaces in multimeric complexes (Bryant et al., 2022; Zhu et al., 2023). ipSAE restricts scoring to low-PAE interchain residue pairs and varies the TM normalization scale with the supported interface or chain size (Dunbrack, 2025). AntiConf further combines confidence measures for antibody–antigen complex assessment (Ünsal et al., 2026). These advances improve estimates of predicted interface quality; our work asks the complementary decision question of how structural confidence should be aggregated to retain true binders while controlling false positives in in silico antibody filtering.