AF3-AQA: ASSESSING AND SELECTING ANTIBODY–ANTIGEN STRUCTURES FROM ALPHAFOLD3 ENSEMBLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate antibody–antigen complex structures are important for structure-based antibody discovery and design. Although AlphaFold3 (AF3) can generate diverse candidate structures through stochastic sampling, these candidates often differ substantially in binding pose and structural accuracy, making reliable candidate selection a critical challenge. We study this problem by systematically analyzing scoring signals and developing a learned model to assess candidate quality. To support this study, we introduce **AF3-AQA** (AlphaFold3 Antibody–Antigen Quality Assessment), comprising 430,100 predicted structures for 391 experimentally resolved targets, with independently sampled pools of 100 and 1,000 candidates per target. Under a unified evaluation protocol, we compare AF3 confidence scores and external rankers, and investigate how different factors contribute to candidate selection. Our analyses show that regional restriction has feature-dependent benefits, aggregation can substantially change selection performance, and strong scalar scores can still fail to distinguish different candidate poses. We further introduce **AQA-Score**, which transfers single-chain interface-site and residue-pair compatibility representations into candidate-specific assessment with additional structure and energy features. On AF3-AQA-100, AQA-Score improves DockQ@1/5 from 0.345/0.380 for iPTM to 0.363/0.402. Together, these results characterize the strengths and limitations of current selection signals and establish a learned approach for improving antibody–antigen structure selection.

1 INTRODUCTION

Antibody–antigen interactions are central to adaptive immunity and form the foundation of many biologic therapeutics (Joubbi et al., 2024). Accurate complex structures support the interpretation of molecular recognition, structure-based antibody design, and computational screening (Fromm et al., 2025; Watson et al., 2023; Schneider et al., 2022; Zhang & Tsuda, 2025). Recent advances in biomolecule structure prediction, including AlphaFold3 (Abramson et al., 2024) (AF3), enable the generation of multiple candidate complex structures for the same antibody–antigen pair through stochastic sampling. These candidates can differ substantially in binding orientation, epitope location, and interface accuracy. Making effective use of such predictions therefore requires both generating useful structures and reliably identifying them among alternative candidates.

We study this challenge as an **Antibody–Antigen Quality Assessment (AQA)** task: given an ensemble of AF3-generated structures for a specific antibody–antigen pair, the goal is to rank candidates by their agreement with an experimentally determined reference structure. Existing selectors like AF3 confidence scores, physical energy terms, and interface geometry provide different sources of evidence, but their combination usefulness remains less explored. This raises two related questions: *which properties of these signals make them effective selectors*, and *how can a learned model combine binding-related knowledge from a specific candidate interface?*

Addressing these questions requires a common evaluation setting. Existing scoring methods are evaluated on different candidate sources, datasets, and protocols, making their relative strengths difficult to establish for AF3-generated antibody–antigen structures. We introduce **AF3-AQA**, constructed from experimentally resolved complexes in SAbDab (Dunbar et al., 2014). It contains

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.

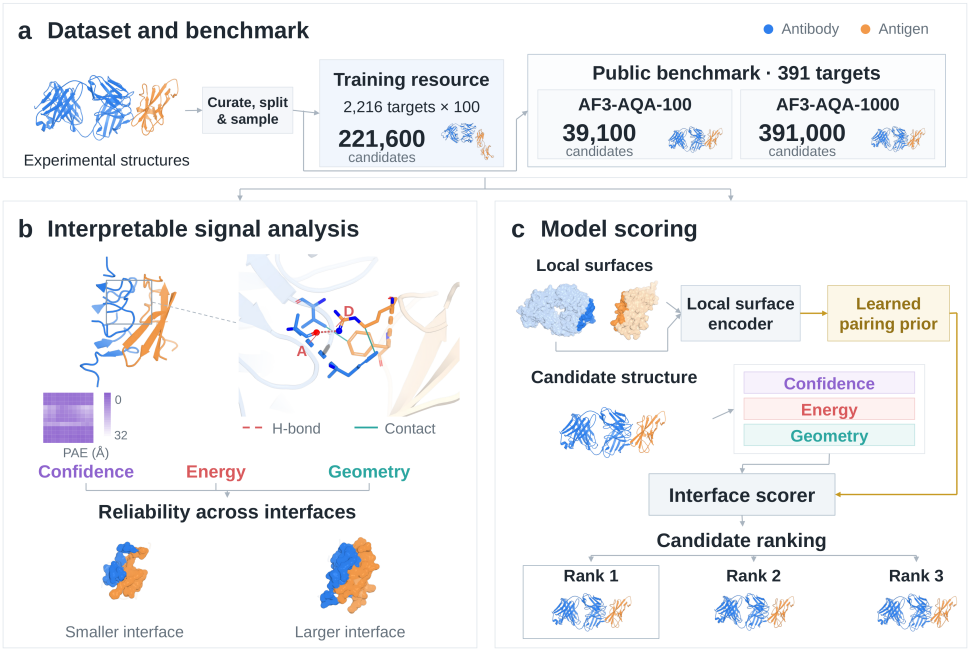


Figure 1: **Overview of AF3-AQA.** (a) Dataset construction from SabDab. (b) Analysis of confidence, energy, and geometric signals for structure selection. (c) AQA-Score combines learned binding priors with candidate-specific interface evidence to rank structures.

430,100 predicted structures for 391 test targets, organized into independently sampled pools of 100 and 1,000 candidates per target. Using AF3-AQA, we systematically investigate confidence, energy, and geometric signals. Starting from individual features, we vary antibody regions and aggregation rules while holding the remaining score definition fixed. The results show that restricting attention to biologically important regions does not uniformly improve selection: the effect depends on the signal being aggregated. Our analyses characterize concrete strengths and limitations of commonly available evidence for structure selection.

We further propose **AQA-Score**, a learned model that brings binding-related representations into candidate-specific quality assessment. AQA-Score transfers single-chain interface-site representations and cross-chain residue-pair compatibility and combines these representations with candidate geometry, AF3 confidence, and local energy features in a sparse interface model. By representing both formed contacts and unformed pair hypotheses, AQA-Score can assess the realized interface alongside potential interactions proposed by the learned prior or AF3 contact probabilities. Figure 1 summarizes the benchmark, signal analysis, and learned scoring framework.

We evaluate AQA-Score against AF3 confidence scores and existing rankers. On AF3-AQA-100, AQA-Score improves DockQ@1/5 from 0.345/0.380 for iPTM to 0.363/0.402, with the clearest gains in pools whose best candidate has medium DockQ quality. Across nested sampling budgets in AF3-AQA-1000, selected quality eventually plateaus while the best available candidate continues to improve. Together, these results identify both progress and substantial remaining opportunities in learning to select AF3 antibody–antigen structures. Our contributions are summarized as follows:

- **Systematic analysis of selection signals.** We characterize how confidence, energy, geometry, regional restriction, aggregation, and sampling support affect candidate selection, including limitations in score resolution and candidate retention.
- **Binding-aware candidate assessment.** We introduce AQA-Score, which combines pretrained site and residue-pair representations with candidate-specific interface evidence, trained with local structural supervision and a pool-wise selection objective.
- **A unified benchmark and empirical evaluation.** We provide AF3-AQA and compare native confidence scores and learned rankers across candidate pools, attainable-quality strata, and sampling budgets, quantifying improvements and unresolved selection gaps.

2 Related Work

Antibody Structure and Complex Modeling. AlphaFold2 (Jumper et al., 2021) and AlphaFold3 (Abramson et al., 2024) have advanced protein structure prediction. IgFold (Ruffolo et al., 2023) and the ABodyBuilder family (Abanades et al., 2023; Kenlay et al., 2024) address antibody modeling, including variable CDR regions, while FAFE (Wu et al., 2024) studies immune-complex modeling with a geodesic distance loss. Our work focuses on assessing and selecting generated antibody–antigen structures.

Confidence-Based Quality Assessment. Prediction models provide confidence quantities such as pLDDT, pTM, PAE, and iPTM. pDockQ and pDockQ2 (Bryant et al., 2022; Zhu et al., 2023) estimate complex quality from interface and confidence features, while ipSAE (Dunbrack Jr, 2025) uses PAE for interface assessment. We evaluate confidence-based selection on AF3-AQA and examine how region selection and aggregation affect performance in Appendix B.

Learned Interface Scoring and Candidate Ranking. DeepRank-Ab (Xu et al., 2025) learns from antibody–antigen interface graphs of docking decoys. ABAG-Rank (Tadiello et al., 2026) combines quality regression, ranking, and distance-preservation objectives. ARID-sf (Grandguillaume et al., 2026) integrates force-field, structural, and protein language model features through attention. AQAScore transfers pretrained interface-site and residue-pair representations into candidate assessment, representing formed contacts and unformed hypotheses and combining local supervision with pool-wise selection.

3 AF3-AQA: Antibody–Antigen Quality Assessment Benchmark

We introduce **AF3-AQA**, a benchmark for assessing and selecting AF3-generated antibody–antigen structures. Each target comprises a fixed antibody–antigen sequence combination, an experimental reference complex, and a pool of AF3-generated candidates. We describe dataset construction, define the quality labels and ranking metrics, and establish confidence-based reference performance.

3.1 Dataset Construction

Antibody–antigen pair filtering. We construct AF3-AQA from SAbDab (Dunbar et al., 2014), starting from 20,943 antibody–antigen complex entries spanning 10,660 unique PDB entries. We retain structures with resolution ≤ 3.5 Å and R-factor ≤ 0.25. Antigens must be proteins in contact with the antibody, where contact requires at least one cross-chain heavy-atom pair within 5 Å. The retained antigen chains must contain no more than 512 residues in total. Antibodies must contain both heavy and light chains, pass ANARCI IMGT numbering (Dunbar & Deane, 2016), and have all six CDRs complete. These filters retain 3,316 antibody–antigen pairs.

We use MMseqs2 (Steinegger & Söding, 2017) to group related sequences, clustering concatenated antibody CDR sequences at 90% identity and concatenated antigen sequences at 70% identity, both with 80% coverage. Each case receives an antibody-cluster identifier and an antigen-cluster identifier. Their combination defines a *cluster pair*. Figure 2 summarizes the filtering and dataset construction procedures.

Temporal and cluster-pair-disjoint split. We use September 30, 2021 as the temporal cutoff and initially assign cases according to their PDB deposition dates. We then enforce the partition at the cluster-pair level: if any member of a cluster pair predates the cutoff, the entire cluster pair is assigned to the pre-cutoff partition. The remaining cluster pairs form the post-cutoff partition, from which 391 representatives define the public test set.

This oldest-wins rule prevents test targets from sharing a cluster pair with the pre-cutoff partition. The separation applies to cluster pairs: individual antibody or antigen clusters need not be disjoint across partitions. The pre-cutoff candidate resource contains 2,216 targets, of which 2,166 are used for training and 50 are held out for validation, each with 100 AF3 candidates.

AF3 candidate generation and retained features. For each test target, we independently generate two candidate pools. **AF3-AQA-100** contains 100 candidates, generated using 20 random seeds with

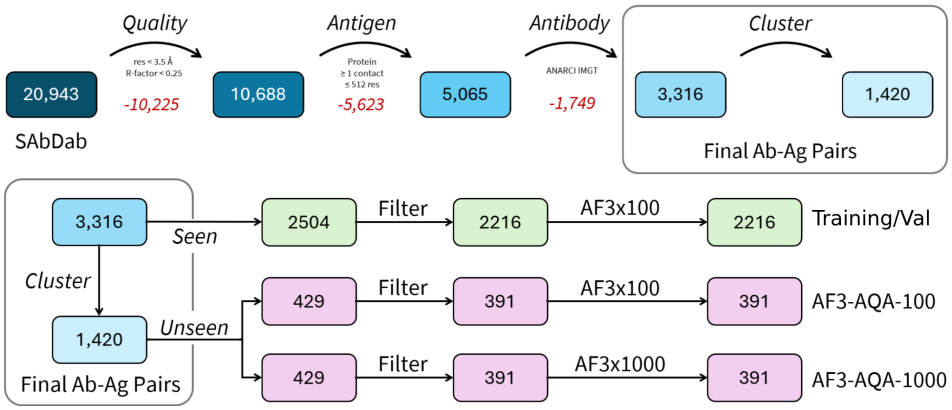


Figure 2: AF3-AQA dataset construction. Top: filtering of SAbDab antibody–antigen complexes. Bottom: sequence clustering, temporal partitioning, and construction of training, validation, and test candidate pools. AF3-AQA-100 and AF3-AQA-1000 contain independently sampled candidates for the same 391 test targets.

five samples per seed. **AF3-AQA-1000** contains 1,000 candidates, generated using 200 random seeds with five samples per seed. The pools use disjoint seeds and contain 39,100 and 391,000 structures, respectively. AF3-AQA-100 is therefore not a subset of AF3-AQA-1000.

AF3 inputs comprise the antibody heavy chain, light chain, and antigen chains, with MSAs prepared through the AF3 input pipeline. For each candidate, we retain the predicted structure, AF3 summary confidence scores, and pairwise confidence tensors, including predicted aligned error (PAE) and contact probabilities (CP). Experimental reference complexes provide quality labels for training and evaluation; they are not used as inputs when scoring test candidates. Additional filtering, input preparation, and data-provenance details appear in Appendix A.1.

3.2 AbAg-DockQ and Ranking Metrics

AbAg-DockQ. We define **AbAg-DockQ** as the primary quality label in AF3-AQA. It retains the standard DockQ scoring form (Basu & Wallner, 2016), combining native-contact recovery, interface RMSD, and ligand RMSD, while fixing the antigen as receptor and the antibody as ligand:

$$\text{AbAg-DockQ} = \frac{1}{3}\left[F_{\text{nat}} + \frac{1}{1+(\text{iRMSD}/1.5)^2} + \frac{1}{1+(\text{Ab-LRMSD}/8.5)^2}\right]. \tag{1}$$

Here, F_{nat} and iRMSD follow the standard DockQ definitions, and RMSDs are measured in Å. Antibody ligand RMSD (Ab-LRMSD) is computed by first aligning the candidate antigen chains to the reference and then measuring the RMSD of the antibody chains. This convention evaluates antibody placement relative to the given antigen target. Unless otherwise specified, all DockQ values in this work refer to AbAg-DockQ.

Top-k ranking metrics. Each method assigns a scalar score to every candidate within a target's pool. Let D_{ci} denote the AbAg-DockQ of candidate i for target c, and let $\text{Top}_k(c)$ contain the k highest-ranked candidates. We evaluate

$$\text{DockQ@}k = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \max_{i\in\text{Top}_k(c)} D_{ci}, \qquad k \in \{1,5\}, \tag{2}$$

where $\mathcal{C}$ is the set of test targets. DockQ@1 measures the quality of the first-ranked candidate. DockQ@5 measures the best quality among the five selected candidates, representing selection with a short candidate list.

We also report success rates at quality threshold t:

$$\text{SR@}t/k = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \mathbf{1}\left[\max_{i\in\text{Top}_k(c)} D_{ci} \geq t\right]. \tag{3}$$

Table 1: Confidence-based baseline performance on AF3-AQA. Both pools contain the same 391 test targets and use independently sampled candidates. Success rates are percentages over all targets. Bold marks the better displayed point estimate between iPTM and ipSAE for each pool and metric, including ties. Random and Oracle are reference rows.

	Top-1				Top-5			
Method	DockQ	SR@.23	SR@.49	SR@.80	DockQ	SR@.23	SR@.49	SR@.80
AF3-AQA-100								
Random	0.2432	34.7	27.0	6.2	0.3539	48.8	38.7	13.3
iPTM	**0.3449**	**49.6**	**40.7**	**11.3**	0.3798	**52.9**	43.2	14.3
ipSAE	0.3302	48.6	39.6	10.0	**0.3846**	52.2	**44.2**	**16.1**
Oracle	0.5015	70.1	53.5	24.6	0.5015	70.1	53.5	24.6
AF3-AQA-1000								
Random	0.2448	34.9	27.4	6.4	0.3587	49.3	39.8	13.6
iPTM	**0.3881**	**56.0**	**46.5**	**13.0**	0.4314	**58.8**	49.9	**18.4**
ipSAE	0.3852	55.8	46.3	12.0	**0.4349**	58.1	**50.9**	**18.4**
Oracle	0.6015	83.1	66.5	35.0	0.6015	83.1	66.5	35.0

We use $t \in \{0.23, 0.49, 0.80\}$ for acceptable, medium, and high quality, respectively, and report success rates as percentages. All metrics weight targets equally. Tie-breaking and uncertainty estimation are described in Appendix A.3.

3.3 Confidence-Based Baselines and Selection Headroom

We first evaluate AF3-native iPTM and the confidence-derived interface score ipSAE (Dunbrack Jr, 2025) as training-free baselines. Two reference selectors characterize the available selection range: **Random** ranks candidates uniformly at random, whereas **Oracle** selects the candidate with the highest reference AbAg-DockQ. The oracle quality for target c is $Q_c^* = \max_i D_{ci}$. All methods rank the same candidates within each pool.

Table 1 shows that iPTM and ipSAE outperform Random but remain below Oracle. From AF3-AQA-100 to AF3-AQA-1000, oracle quality increases from 0.5015 to 0.6015, while confidence-based selection improves more modestly. Larger candidate pools therefore leave substantial quality gains unrealized by these selectors. Additional candidate-quality distributions and selection-regret summaries appear in Appendix A.2. We evaluate learned selectors in Section 5 and analyze individual confidence and energy signals in Appendix B.

4 AQA-Score: Binding-Aware Structure Assessment

AQA-Score combines pretrained residue-pair compatibility with structural evidence from each candidate. The pretrained model describes potential interactions between the antibody and antigen, while the candidate assessment model evaluates their realization in a particular structure. These complementary representations support prediction of the three AbAg-DockQ components and a scalar ranking score (Figure 3).

4.1 Pretrained Binding Representations

Module A encodes each chain through an intra-chain atom graph and atom-to-residue pooling, fused with residue attributes. We group the resulting representations into antibody and antigen partners, H_A and H_B. Module B applies bidirectional cross-attention and a low-rank pair head:

$$(\widetilde{H}_A, \widetilde{H}_B) = \mathrm{CrossAttn}(H_A, H_B), \qquad \ell_{rs} = \mathrm{PairHead}(\widetilde{h}_{A,r}, \widetilde{h}_{B,s}), \tag{4}$$

where ℓ_{rs} is a raw residue-pair compatibility logit. The local states, partner-conditioned states, and pair logits provide binding representations for candidate assessment.

Module A is first pretrained for interface-site prediction, then jointly trained with Module B for residue-pair contact prediction. Both modules are frozen during candidate-quality training. For fixed

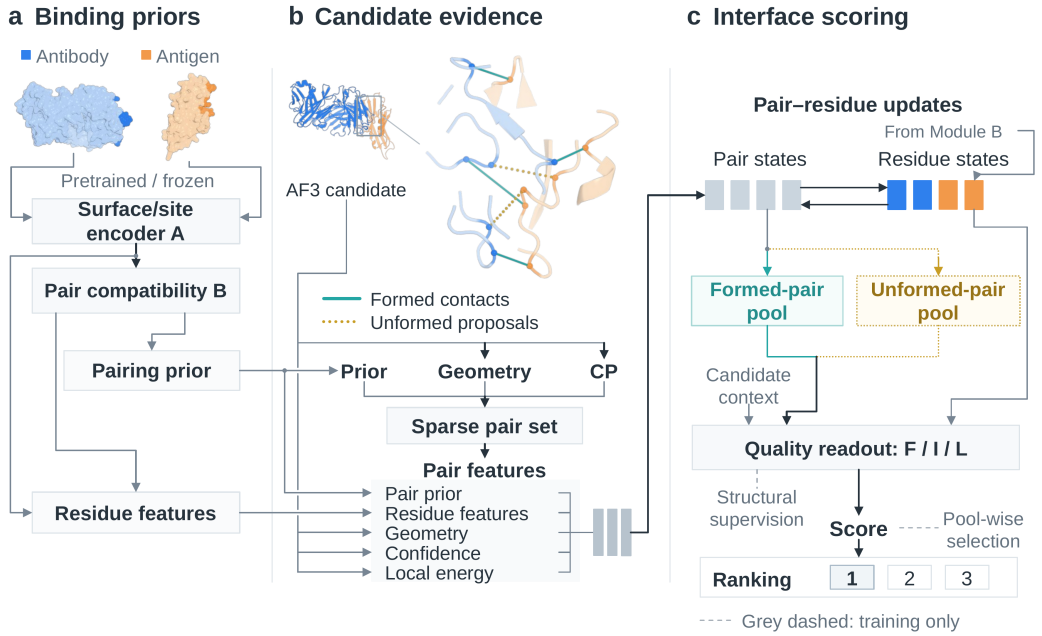


Figure 3: AQA-Score architecture. (a) A single-chain encoder (Module A) and a cross-chain compatibility model (Module B) provide frozen binding representations. (b) Candidate geometry, compatibility priors, and AF3 contact probabilities propose sparse residue pairs. (c) Pair–residue interactions and separate pooling of formed contacts and unformed hypotheses support component-wise quality prediction. Gray dashed connections indicate training-only supervision. Training uses complete candidate pools; inference scores each candidate independently.

chain conformations and atom features, their outputs are independent of relative chain placement. Candidate atom features include pLDDT, while cross-chain geometry, PAE, CP, and energy enter downstream assessment. Architecture and pretraining details appear in Appendix D.1.

4.2 Candidate-Specific Interface Assessment

Assessing a candidate requires connecting potential residue pairings to its current geometry. We construct a sparse cross-chain pair set by combining geometric proximity, pretrained compatibility, and AF3 contact probabilities (CP):

$$\Omega_{ci} = \Omega_{\text{geometry},ci} \cup \Omega_{\text{prior},ci} \cup \Omega_{\text{CP},ci}. \tag{5}$$

Geometry contributes pairs within 10 Å minimum heavy-atom distance; the prior and CP contribute high-scoring partners in both directions. This union retains the observed interface together with proposed pairings that may be spatially separated in the current candidate.

We distinguish *formed contacts*, with minimum heavy-atom distance below 5 Å, from *unformed hypotheses* proposed by the prior or CP outside this contact set. Formed contacts describe interactions realized by the candidate. Unformed hypotheses preserve information about proposed interactions that its current geometry does not realize. They are not assumed to be native contacts. Both sets are constructed entirely from inference-time inputs.

For each pair, we combine pretrained residue states and compatibility with candidate geometry, directional PAE and CP, and context features including residue confidence, antibody-region identity, and local Rosetta descriptors. Pair–residue updates propagate this evidence between pairs and their incident residues. Separate attention pooling summarizes formed contacts and unformed hypotheses, preserving their distinct roles before combining them with candidate-level context for quality prediction. The Rosetta inputs are local descriptors. Feature definitions and pair truncation are specified in Appendix D.2.

Table 2: Selection performance on AF3-AQA-100. Success rates are percentages over all 391 targets. Bold marks the largest displayed point estimate among practical selectors in each column, including ties. Random and Oracle are reference rows.

	Top-1				Top-5			
Method	DockQ	SR@.23	SR@.49	SR@.80	DockQ	SR@.23	SR@.49	SR@.80
iPTM	0.345	49.6	40.7	**11.3**	0.380	52.9	43.2	14.3
ipSAE	0.330	48.6	39.6	10.0	0.385	52.2	44.2	16.1
ABAG-Rank	0.285	41.7	30.9	8.4	0.354	49.4	38.9	13.3
ARID-sf	0.227	32.7	23.8	5.4	0.312	42.7	33.5	10.5
DeepRank-Ab	0.333	48.8	40.7	9.0	**0.404**	**56.0**	45.6	15.3
AQA-Score	**0.363**	**51.9**	**43.5**	11.0	0.402	54.7	**45.8**	**17.6**
Random	0.245	35.3	27.6	5.6	0.348	48.1	36.3	12.8
Oracle	0.502	70.1	53.5	24.6	0.502	70.1	53.5	24.6

4.3 Quality Prediction and Pool-Wise Learning

We predict the three quality components in Equation 1. A sigmoid head estimates native-contact recovery $\widehat{F}_{ci}$, while residue-level heads predict non-negative structural errors. These errors are aggregated into interface and antibody RMSD estimates using candidate-interface residues (within 10 Å) and all antibody residues, respectively. Applying the corresponding DockQ transforms yields $\widehat{I}_{ci}$ and $\widehat{L}_{ci}$. The final ranking score is

$$s_{ci} = \frac{1}{3}\left(\widehat{F}_{ci} + \widehat{I}_{ci} + \widehat{L}_{ci}\right). \tag{6}$$

We supervise the components separately, with auxiliary losses for local residue errors and formed-contact correctness. Reference structures provide supervision only; pair construction and inference masks remain candidate-defined. Aggregation rules and local objectives appear in Appendix D.3.

To optimize selection within each target's candidate pool, we additionally minimize expected selection regret:

$$p_{ci} = \frac{\exp(s_{ci}/\tau)}{\sum_{j=1}^{N_c}\exp(s_{cj}/\tau)}, \qquad \mathcal{L}_{\text{select},c} = \sum_{i=1}^{N_c} p_{ci}\left(Q_c^* - D_{ci}\right), \tag{7}$$

where N_c is the pool size and τ is the softmax temperature; D_{ci} and Q_c^* follow Section 3. Training uses complete 100-candidate pools and weights targets equally. At inference, each candidate is scored independently and ranked by s_{ci}. The complete objective and optimization settings appear in Appendix D.4.

5 Ranker Evaluation on AF3-AQA

Compared methods. We evaluate confidence-based selectors, external learned rankers, and AQA-Score under the same AF3-AQA protocol. Confidence baselines include AF3-native iPTM and ipSAE (Dunbrack Jr, 2025). External learned rankers include ABAG-Rank (Tadiello et al., 2026), ARID-sf (Grandguillaume et al., 2026), and DeepRank-Ab (Xu et al., 2025). These models retain their original training histories and are not retrained under the AF3-AQA split. We compare them with AQA-Score, introduced in Section 4, and retain Random and Oracle as reference selectors.

AF3-AQA-100 provides the full comparison across methods. On AF3-AQA-1000, we evaluate iPTM, ipSAE, and AQA-Score, together with Random and Oracle, to examine performance at a larger candidate-pool size. Both benchmarks contain the same 391 test targets, with independently sampled candidate pools. All methods evaluated on a given benchmark rank the same candidates.

5.1 Results on AF3-AQA-100

External learned rankers do not uniformly improve over confidence-based selection (Table 2). ABAG-Rank and ARID-sf fall below iPTM in both DockQ@1 and DockQ@5. DeepRank-Ab improves Top-5

Table 3: Selection performance on AF3-AQA-1000. Success rates are percentages over the same 391 test targets. Bold marks the largest displayed point estimate among practical selectors in each column. Random and Oracle are reference rows.

	Top-1				Top-5			
Method	DockQ	SR@.23	SR@.49	SR@.80	DockQ	SR@.23	SR@.49	SR@.80
iPTM	0.3881	56.0	46.5	13.0	0.4314	58.8	49.9	18.4
ipSAE	0.3852	55.8	46.3	12.0	0.4349	58.1	50.9	18.4
AQA-Score	**0.4004**	**57.3**	**47.3**	**14.6**	**0.4518**	**62.1**	**52.4**	**19.7**
Random	0.2448	34.9	27.4	6.4	0.3587	49.3	39.8	13.6
Oracle	0.6015	83.1	66.5	35.0	0.6015	83.1	66.5	35.0

selection but has lower DockQ@1 than iPTM. AQA-Score achieves the highest DockQ@1 point estimate among the evaluated selectors, reaching 0.363 compared with 0.345 for iPTM. DeepRank-Ab achieves the highest DockQ@5 at 0.404, slightly above AQA-Score's 0.402. Performance also varies across success thresholds: iPTM has the highest Top-1 SR@.80, whereas DeepRank-Ab has the highest Top-5 SR@.23. AQA-Score therefore improves average first-choice quality without dominating every retrieval criterion.

Using unrounded results, the paired DockQ@1/5 improvements over iPTM are 0.0183/0.0224. The Top-1 antigen-family bootstrap interval includes zero, whereas the Top-5 improvement remains positive under both target and antigen-family resampling. Uncertainty estimates are reported in Appendix E.2.

5.2 Scale-up Evaluation on AF3-AQA-1000

AQA-Score remains above iPTM and ipSAE on both DockQ metrics and all reported success rates in AF3-AQA-1000 (Table 3). It achieves DockQ@1/5 of 0.4004/0.4518, compared with 0.3881/0.4314 for iPTM and 0.3852/0.4349 for ipSAE. Its selection advantage therefore extends to the larger candidate pools.

Nevertheless, AQA-Score remains below the oracle quality of 0.6015, even with five retained candidates. At the acceptable-quality threshold, its Top-5 success rate is 62.1%, compared with 83.1% candidate availability. Thus, substantial selection headroom remains. Additional sampling-budget experiments are reported in Appendix F.

5.3 Difficulty-Stratified Analysis

We stratify AF3-AQA-100 targets by their oracle AbAg-DockQ using the quality thresholds defined in Section 3.2. These reference-based strata characterize attainable candidate quality and are used only for analysis. Table 4 reports DockQ@1 and DockQ@5 within each stratum.

AQA-Score's largest improvement over iPTM occurs in the Medium stratum, where DockQ@1/5 increases from 0.429/0.484 to 0.485/0.539. The gain therefore extends to both first-choice quality and five-candidate shortlists. These 113 targets account for 88.6% of the summed Top-1 gain over iPTM, with positive Medium-stratum improvements under both target and antigen-family resampling.

Improvements in the other strata are smaller. In the High stratum, AQA-Score and iPTM have nearly identical DockQ@1, while AQA-Score improves DockQ@5 from 0.815 to 0.827. Below the acceptable threshold, no method can retrieve an acceptable structure because none is available. Overall, the benefit of AQA-Score is concentrated in the Medium stratum rather than distributed uniformly across attainable-quality levels. Additional stratified results are provided in Appendix E.

5.4 Input and Loss Ablations

We evaluate two input ablations and two loss ablations on AF3-AQA-100. The input variants remove AF3 confidence or local Rosetta energy features. The loss variants replace separate component regression with regression on the mean predicted score, or remove the pool-wise selection loss $\mathcal{L}_{\text{select}}$.

Table 4: Selection performance across oracle-quality strata on AF3-AQA-100. Strata are defined by the best available AbAg-DockQ Q^*: below acceptable ($Q^* < 0.23$), acceptable ($0.23 \leq Q^* < 0.49$), medium ($0.49 \leq Q^* < 0.80$), and high ($Q^* \geq 0.80$). D@1/5 denotes DockQ@1/5. Bold marks the largest displayed point estimate among practical selectors in each column, including ties. Random and Oracle are reference rows.

	Below acceptable $n = 117$		Acceptable $n = 65$		Medium $n = 113$		High $n = 96$	
Method	D@1	D@5	D@1	D@5	D@1	D@5	D@1	D@5
iPTM	0.030	0.042	0.131	0.164	0.429	0.484	0.774	0.815
ipSAE	0.024	0.046	0.100	0.154	0.417	0.497	0.757	0.821
ABAG-Rank	0.029	0.048	0.105	0.169	0.353	0.460	0.639	0.727
ARID-sf	**0.031**	**0.051**	0.102	0.145	0.291	0.405	0.475	0.634
DeepRank-Ab	**0.031**	0.048	0.100	**0.180**	0.446	**0.550**	0.726	0.820
AQA-Score	0.030	0.045	**0.142**	**0.180**	**0.485**	0.539	**0.775**	**0.827**
Random	0.025	0.049	0.099	0.168	0.273	0.431	0.576	0.735
Oracle	0.104	0.104	0.363	0.363	0.673	0.673	0.877	0.877

Table 5: Input and loss ablations on AF3-AQA-100. Results are single-run point estimates over the same 391 test targets. Aggregate-score supervision replaces separate component regression with regression on their mean score.

Variant	DockQ@1	DockQ@5
Full AQA-Score	**0.3631**	**0.4023**
Input ablations		
Without AF3 confidence	0.3426	0.3935
Without local Rosetta energy	0.3518	0.3953
Loss ablations		
Aggregate-score supervision	0.3509	0.3978
Without $\mathcal{L}_{\text{select}}$	0.3475	0.3924

Removing either input channel lowers both metrics, with the larger Top-1 decrease occurring without AF3 confidence (Table 5). Both loss variants also reduce performance. Removing $\mathcal{L}_{\text{select}}$ produces a larger decrease than replacing component regression with aggregate-score supervision.

These single-run results support the usefulness of both input channels and the evaluated training objectives. They do not quantify training-seed variation or separately isolate the contributions of pretraining and unformed hypotheses. Additional ablation details appear in Appendix E.3.

Complementary analyses of standalone confidence and energy scores, including the effects of region selection and aggregation, are provided in Appendix B.

6 Conclusion

We introduced AF3-AQA, a benchmark for selecting antibody–antigen structures from independently sampled AF3 candidate pools, and examined confidence, energy, and geometric signals under a common evaluation protocol. Our analyses show that selection depends on both the region and aggregation used to construct a score, while competitive scalar scores can still leave distinct candidate poses unresolved. AQA-Score combines pretrained binding representations with candidate-specific interface evidence to improve first-choice quality over the evaluated baselines, although DeepRank-Ab retains a slightly higher Top-5 point estimate. Gains are concentrated in medium-quality pools, and selected quality plateaus as sampling increases despite continued improvement in the best available candidate. The overall Top-1 gain remains uncertain under antigen-family resampling, and the available input ablations use single training runs. Closing the remaining selection gap and establishing broader generalization remain important directions for future work.

REFERENCES

Brennan Abanades, Wing Ki Wong, Fergus Boyles, Guy Georges, Alexander Bujotzek, and Charlotte M Deane. Immunebuilder: Deep-learning models for predicting the structures of immune proteins. *Communications Biology*, 6(1):575, 2023.

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(8016):493–500, 2024.

Sankar Basu and Björn Wallner. Dockq: a quality measure for protein-protein docking models. *PloS one*, 11(8):e0161879, 2016.

Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using alphafold2. *Nature communications*, 13(1):1265, 2022.

James Dunbar and Charlotte M Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42 (D1):D1140–D1146, 2014.

Roland L Dunbrack Jr. Rēs ipsae loquunt: What's wrong with alphafold's iptm score and how to fix it. *bioRxiv*, 2025.

Samuel Fromm, Marko Ludaic, and Arne Elofsson. Evaluating deep learning based structure prediction methods on antibody-antigen complexes. *bioRxiv*, pp. 2025–07, 2025.

Ilyas Grandguillaume, Catherine Etchebest, and Fernando Luis Barroso da Silva. Arid-sf: A physics-informed deep learning scoring function to improve antibody-antigen docking model ranking. *bioRxiv*, pp. 2026–01, 2026.

Sara Joubbi, Alessio Micheli, Paolo Milazzo, Giuseppe Maccari, Giorgio Ciano, Dario Cardamone, and Duccio Medini. Antibody design using deep learning: from sequence and structure design to affinity maturation. *Briefings in Bioinformatics*, 25(4):bbae307, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Henry Kenlay, Frédéric A Dreyer, Daniel Cutting, Daniel Nissley, and Charlotte M Deane. Abodybuilder3: improved and scalable antibody structure predictions. *Bioinformatics*, 40(10):btae576, 2024.

Jeffrey A Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature communications*, 14(1):2389, 2023.

Constantin Schneider, Andrew Buchanan, Bruck Taddese, and Charlotte M Deane. Dlab: deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics*, 38(2):377–383, 2022.

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.

Matteo Tadiello, Marko Ludaic, Vsevolod Viliuga, and Arne Elofsson. Abag-rank: Improving model selection of alphafold antibody-antigen complexes by learning to rank. *bioRxiv*, pp. 2026–03, 2026.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Ruidong Wu, Ruihan Guo, Rui Wang, Shitong Luo, Yue Xu, Jiahan Li, Jianzhu Ma, Qiang Liu, Yunan Luo, and Jian Peng. Fafe: Immune complex modeling with geodesic distance loss on noisy group frames. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Xiaotong Xu, Ilaria Coratella, Victor Reys, and Alexandre MJJ Bonvin. Deeprank-ab: a scoring function for antibody-antigen complexes based on geometric deep learning. *bioRxiv*, pp. 2025–12, 2025.

Yiming Zhang and Koji Tsuda. Ab-vs: Evaluating large language models for virtual antibody screening via antibody-antigen interaction prediction. *bioRxiv*, pp. 2025–07, 2025.

Wensi Zhu, Aditi Shenoy, Petras Kundrotas, and Arne Elofsson. Evaluation of alphafold-multimer prediction on multi-chain protein complexes. *Bioinformatics*, 39(7):btad424, 2023.

Reproducibility statement

Benchmark construction, evaluation protocols, model inputs, and training procedures are described in the main text and Appendix D. Candidate-level results and figure-generation code accompany the submission. Individual candidate structures and repeated subsamples are not treated as independent biological targets.

Ethics statement

Structural predictions and rankings are hypotheses, not evidence of binding affinity, therapeutic efficacy, or safety. Experimental validation remains necessary before biological or clinical use.

AI assistance disclosure

We acknowledge the use of large language models (LLMs) for writing assistance during the preparation of this manuscript. Specifically, LLMs were used to refine language, improve clarity and readability, and enhance the coherence of the presentation. All core research contributions, including the methodology, experimental design, theoretical analysis, and scientific findings, were developed by the authors. The use of LLMs was limited to language polishing and did not extend to research conception, experimental validation, or interpretation of results. The authors reviewed all AI-assisted revisions and take full responsibility for the accuracy and integrity of the manuscript.

A Dataset and Evaluation Details

A.1 Dataset Provenance

AF3-AQA is curated from 20,943 SAbDab antibody–antigen entries spanning 10,660 unique PDB entries. We retain structures with resolution ≤ 3.5 Å and R-factor ≤ 0.25. Antigens must be proteins with at least one antibody–antigen heavy-atom pair within 5 Å, and retained antigen chains must contain at most 512 residues in total. Antibodies must contain paired heavy and light chains, pass ANARCI IMGT numbering (Dunbar & Deane, 2016), and have all six CDRs complete. These filters retain 3,316 antibody–antigen pairs.

MMseqs2 (Steinegger & Söding, 2017) clusters concatenated antibody CDR sequences at 90% identity and concatenated antigen sequences at 70% identity, both with 80% coverage. The combination of cluster identifiers defines the partitioning unit. A cluster pair belongs to the pre-cutoff partition if any member has a PDB deposition date on or before September 30, 2021. Individual antibody or antigen clusters need not be disjoint across partitions.

The pre-cutoff candidate resource contains 2,216 targets with 100 candidates each. We hold out 50 targets for validation and use the remaining 2,166 for training. The public test set contains 391 representatives from the post-cutoff cluster pairs. AF3 inputs include the antibody heavy and light

chains and retained antigen chains, with MSAs generated through the AF3 input pipeline. We retain candidate structures, native summary confidence scores, and predicted aligned error (PAE) and contact probability (CP) tensors.

Modules A and B are pretrained on a processed corpus of general protein complexes. Records must contain at most 1,024 protein residues, at least two chains, and at least 10 valid residues per chain. An eligible chain pair must have at least 30 positive residue contacts and at least 10 contacting residues on each chain. Of 170,629 source training records, 46,496 yield 68,126 eligible chain pairs. Of 15,545 records in the upstream sequence-cluster validation split, 3,956 yield 5,364 pairs. The source metadata specifies a training cutoff of September 30, 2021, but this cutoff was not independently revalidated during the present conversion. A project-specific homology map against the AF3-AQA test set is also unavailable, so pretraining–test homology exclusion has not been fully audited.

The public evaluation pools contain 39,100 valid candidates in AF3-AQA-100 and 391,000 in AF3-AQA-1000. A candidate is identified by its pool, target, AF3 seed, and sample identifiers. All candidates have valid corrected DockQ labels, and no candidates are removed from the reported comparison intersections. The two pools use disjoint seeds and independent sampling campaigns for the same 391 targets. AF3-AQA-100 is therefore not a subset of AF3-AQA-1000.

A.2 Complete Candidate-Quality Profiles

For target c and candidate i, standard DockQ is

$$D_{ci} = \frac{1}{3}\left[F_{\mathrm{nat},ci} + \frac{1}{1+(\mathrm{iRMSD}_{ci}/1.5)^2} + \frac{1}{1+(\mathrm{LRMSD}_{ci}/8.5)^2}\right]. \tag{8}$$

Here F_{nat} is the fraction of native interface contacts, and iRMSD and LRMSD measure interface and ligand RMSD in angstroms. For LRMSD, we align antigen chains and measure antibody displacement.

For candidate pool P_c, its oracle quality is

$$Q_c^*(P_c) = \max_{i \in P_c} D_{ci}. \tag{9}$$

We define four mutually exclusive oracle strata: Below acceptable ($Q_c^* < 0.23$), Acceptable ($0.23 \le Q_c^* < 0.49$), Medium ($0.49 \le Q_c^* < 0.80$), and High ($Q_c^* \ge 0.80$). Strata are computed separately for each independently sampled pool. They use reference quality only for analysis and are unavailable during candidate selection.

Writing $q_c^{(k)} = \max_{i \in \mathrm{Top}_k(c)} D_{ci}$, the success rate at threshold τ is

$$\mathrm{SR@}\tau/k = \frac{1}{|\mathcal{C}|}\sum_{c \in \mathcal{C}} \mathbf{1}\left[q_c^{(k)} \ge \tau\right], \qquad \tau \in \{0.23, 0.49, 0.80\}. \tag{10}$$

We additionally measure candidate prevalence, $f_c(\tau) = |P_c|^{-1}\sum_{i \in P_c}\mathbf{1}[D_{ci} \ge \tau]$, and selection regret, $r_c^{(k)} = Q_c^*(P_c) - q_c^{(k)}$.

Figure 4 compares attainable and selected quality. On AF3-AQA-100, random ordering achieves DockQ@1 of 0.2447, iPTM reaches 0.3449, and the oracle reaches 0.5015. The independently sampled AF3-AQA-1000 pool has a mean oracle of 0.6015. Medium pools have the largest mean iPTM Top-1 regret in both pools, at 0.244 and 0.312, respectively. Acceptable pools also retain substantial regret, at 0.232 and 0.273.

Table 6 summarizes candidate-quality distributions within the oracle strata defined above. For each target, we compute the candidate mean, median, 90th percentile, and fraction above each quality threshold, then average these summaries across targets in the stratum. Every target receives equal weight. The median and 90th-percentile columns are therefore averages of within-target statistics, rather than quantiles of a pooled candidate distribution.

Table 7 compares stratum assignments across the two independent pools. Of the 391 targets, 121 change stratum. Because these pools are independently sampled, a larger pool need not have a higher oracle for every individual target. This comparison differs from the nested-prefix experiment in Appendix F, where the oracle is non-decreasing along each sampling sequence.

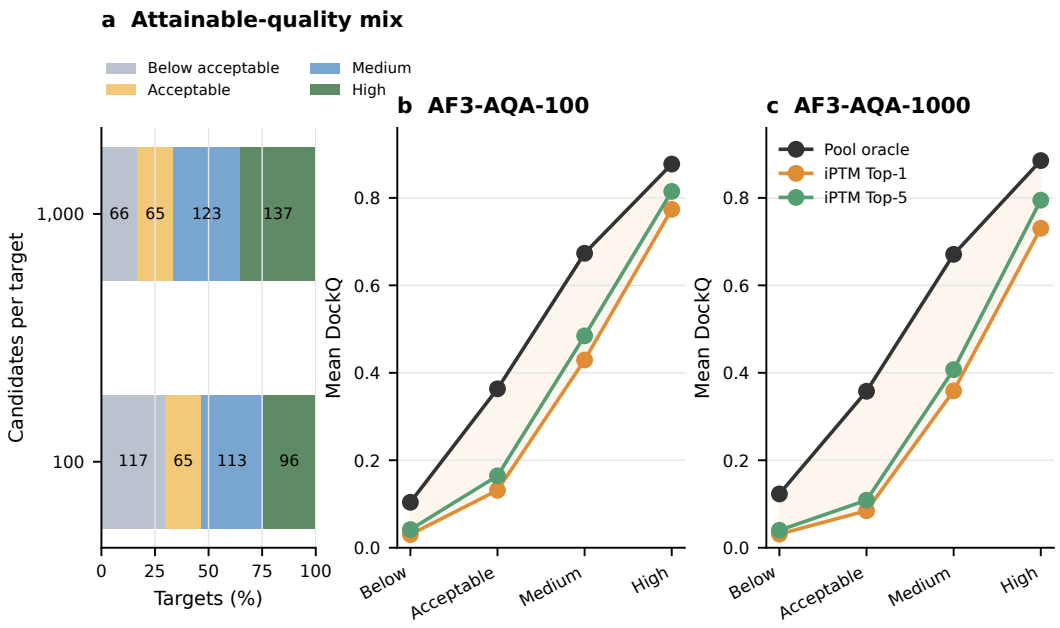


Figure 4: Candidate quality and selection difficulty in AF3-AQA. (a) Target counts in the four pool-specific oracle strata. (b–c) Mean oracle and quality retrieved by iPTM at Top-1 and Top-5 within each stratum. Shading indicates the remaining Top-1 gap to the oracle.

Table 6: Candidate-quality distributions within pool-specific oracle strata. n is the number of targets; $f(\tau)$ is the mean within-target fraction with DockQ $\geq \tau$. Below denotes below acceptable.

Pool	Stratum	n	Mean	Median	90th pct.	$f(.23)$	$f(.49)$	$f(.80)$
100	Below	117	0.028	0.022	0.050	0.000	0.000	0.000
	Acceptable	65	0.088	0.072	0.156	0.134	0.000	0.000
	Medium	113	0.283	0.253	0.449	0.476	0.330	0.000
	High	96	0.564	0.582	0.762	0.761	0.710	0.254
1,000	Below	66	0.025	0.020	0.042	0.000	0.000	0.000
	Acceptable	65	0.053	0.041	0.104	0.041	0.000	0.000
	Medium	123	0.213	0.191	0.350	0.371	0.211	0.000
	High	137	0.470	0.468	0.667	0.643	0.591	0.183

A.3 Ranking, Retrieval, and Uncertainty

Scores are oriented so that larger values indicate better candidates, then sorted in descending score order, with ties resolved by ascending AF3 seed and sample identifiers. DockQ@5 is the largest DockQ among the first five candidates after this ordering. Metrics are first computed within each target and then averaged across targets.

Overall success rates use all evaluated targets as the denominator. For retrieval conditional on candidate availability, the denominator is instead the number of targets whose pool contains at least one candidate meeting the threshold. These conditional counts distinguish a failure to retrieve a qualifying candidate from its absence in the sampled pool.

Paired comparisons use 10,000 target-bootstrap resamples with seed 20260926. The same resampled targets are used for both methods in each comparison. As a robustness analysis, antigen-family intervals resample the 194 antigen clusters as blocks. Reported intervals are 95% confidence intervals. Neither resampling scheme estimates variation across model-training seeds.

Table 7: Cross-pool oracle-stratum transitions. Rows correspond to AF3-AQA-100 and columns to AF3-AQA-1000.

	Below	Acceptable	Medium	High
Below	64	30	14	9
Acceptable	1	34	29	1
Medium	1	1	78	33
High	0	0	2	94

B Analysis of Structure-Selection Signals

We investigate three questions about fixed scoring rules: which energy scores are effective selectors, whether restricting scores to antibody CDRs improves selection, and how aggregation affects confidence signals. All comparisons use AF3-AQA-1000, with the same 391 targets and 391,000 candidates. Scores are constructed from candidate-visible information before joining reference quality labels. These exploratory analyses characterize standalone scoring rules rather than their contributions within AQA-Score.

Which energy scores are effective selectors? The ref2015-weighted interface total achieves DockQ@1/5 of 0.374/0.440, outperforming each tested individual energy score (Table 9). Hydrogen bonding is the strongest individual interface term at 0.333/0.408. For comparison, iPTM achieves 0.388/0.431: the weighted energy total has lower Top-1 quality but a higher Top-5 point estimate. These comparisons do not isolate the marginal contribution of each term within the total.

Does restricting scores to CDRs help? The effect depends on the signal. Restricting PAE mean from all antibody residues to all CDRs improves DockQ@1/5 from 0.386/0.429 to 0.394/0.434. The same restriction lowers hydrogen-bond energy performance from 0.333/0.408 to 0.320/0.395. Region choice therefore needs to be evaluated jointly with the scoring signal.

How much does aggregation matter? For the fixed H3 region, replacing CP mean with maximum aggregation improves DockQ@1/5 from 0.223/0.265 to 0.394/0.435 (Table 8). The paired Top-1 improvement is 0.171 (95% target-bootstrap CI $[0.141, 0.200]$). PAE mean and minimum differ less, with their relative performance depending on region and retrieval depth. Regional energy sums and means induce identical rankings under the fixed residue masks defined below. Aggregation effects are therefore signal-dependent.

B.1 Confidence Regions and Aggregation

We use six antibody-side regions: all antibody residues, the complete heavy chain, the complete light chain, all six CDRs, the three heavy-chain CDRs (H1–H3), and H3 alone. All regions are nonempty for the evaluated targets. CDR indices in this analysis follow the recorded Chothia/ANARCI mapping; dataset inclusion in Section 3.1 uses IMGT numbering. The historical cache shorthand `h` denotes H1–H3 rather than the complete heavy chain. The antigen side contains all antigen tokens, and no native-interface mask is used.

For antibody token a and antigen token g, the controlled PAE grid first averages both directions:

$$\overline{\mathrm{PAE}}_{ag} = \frac{\mathrm{PAE}_{a\to g} + \mathrm{PAE}_{g\to a}}{2}. \tag{11}$$

We then take the mean or minimum over the selected region–antigen pairs. CP uses the cached antibody-to-antigen matrix, with mean or maximum aggregation. Lower PAE and higher CP rank first. Table 8 reports the complete comparison.

The historical PAE-H mean selector reported in supplementary comparisons uses the unidirectional H1–H3-to-antigen mean. It is distinct from the bidirectional definition used in this grid.

The available confidence-matrix cache does not provide a complete pLDDT score table for AF3-AQA-1000. We therefore omit pLDDT from this scalar comparison. This concerns availability for the diagnostic table, not the presence of residue-confidence inputs in AQA-Score.

Table 8: Complete confidence grid on AF3-AQA-1000. D@1/5 denotes DockQ@1/5; success rates are percentages over all 391 targets. PAE uses bidirectional pair averaging, whereas CP uses antibody-to-antigen probabilities.

Feature	Region	Aggregation	D@1	D@5	SR@.23/1	SR@.23/5
PAE	All antibody	mean	0.386	0.429	55.8	58.3
PAE	Heavy chain	mean	0.391	0.431	56.3	58.6
PAE	Light chain	mean	0.387	0.427	55.8	58.3
PAE	All CDRs	mean	0.394	0.434	56.3	58.3
PAE	H1–H3	mean	0.396	0.437	56.5	58.8
PAE	H3	mean	0.392	0.432	56.0	57.5
PAE	All antibody	min	0.390	0.434	56.0	58.3
PAE	Heavy chain	min	0.389	0.436	56.3	58.8
PAE	Light chain	min	0.387	0.433	56.0	58.1
PAE	All CDRs	min	0.391	0.434	56.0	58.1
PAE	H1–H3	min	0.388	0.437	56.0	58.6
PAE	H3	min	0.389	0.439	56.8	59.1
CP	All antibody	mean	0.210	0.267	29.2	36.6
CP	Heavy chain	mean	0.217	0.272	31.2	37.3
CP	Light chain	mean	0.187	0.250	27.1	34.5
CP	All CDRs	mean	0.204	0.256	28.9	34.5
CP	H1–H3	mean	0.223	0.274	31.2	37.9
CP	H3	mean	0.223	0.265	30.2	35.0
CP	All antibody	max	0.389	0.429	55.0	57.0
CP	Heavy chain	max	0.391	0.431	55.5	57.3
CP	Light chain	max	0.372	0.417	52.7	55.5
CP	All CDRs	max	0.389	0.429	55.0	57.3
CP	H1–H3	max	0.393	0.430	56.0	57.5
CP	H3	max	0.394	0.435	56.0	58.6

B.2 Energy Definitions and Regional Aggregation

The interface atom-pair cache provides raw `fa_atr`, `fa_rep`, and a combined hydrogen-bond channel assigned to antibody residues. The interface residue cache provides `fa_elec`. These quantities use candidate-interface neighborhoods and the source pipeline's atom/residue accounting. They are local energy descriptors rather than monomer-subtracted binding free energies. For each component, we sum contributions assigned to the selected antibody region and rank lower values first.

All candidates of a target use the same annotated residue set, including zero contributions from nonparticipating residues. For component u and fixed region R_c,

$$E_{ci,u}^{\text{mean}}(R_c) = \frac{E_{ci,u}^{\text{sum}}(R_c)}{|R_c|}. \tag{12}$$

Because $|R_c|$ is a positive constant within a target, sum and mean induce identical rankings. Table 9 therefore reports only sums for regional terms. Missing values remain unavailable, whereas zero interface contributions remain valid zeros.

The regional cache does not contain interface-local `fa_sol`. The individual solvation score therefore uses the available whole-complex total and differs in scope from the regional interface terms.

The weighted interface selector, $e_{\text{interface_total}}$, uses cached ref2015 weights of 1.0, 0.55, 1.0, 1.0, and 1.0 for attraction, repulsion, solvation, electrostatics, and the combined hydrogen-bond channel, respectively. The alternative energy_atom_total restores these five per-atom channels to their raw scales and sums them with equal weights over interface atoms. An atom belongs to this interface when its residue participates in a candidate interface pair. The two totals differ in weighting and accounting scope; neither scalar total is directly supplied to AQA-Score. Their comparison is reported in Appendix C.3.

Table 9: Energy-score comparison on AF3-AQA-1000. Regional terms use summed interface-local contributions; means yield identical rankings. Solvation uses a whole-complex total, and weighted interface energy uses ref2015 weights. All rows cover 391 targets. Success rates are percentages; dashes indicate values not included in this summary.

Feature	Region / scope	D@1	D@5	SR@.23/1	SR@.23/5
`fa_atr`	All antibody	0.163	0.254	21.5	33.5
`fa_atr`	Heavy chain	0.160	0.248	21.7	32.5
`fa_atr`	Light chain	0.120	0.217	15.9	27.6
`fa_atr`	All CDRs	0.164	0.250	22.5	34.0
`fa_atr`	H1–H3	0.151	0.229	19.9	30.4
`fa_atr`	H3	0.143	0.245	20.2	33.5
`fa_rep`	All antibody	0.262	0.390	37.1	51.4
`fa_rep`	Heavy chain	0.191	0.328	25.8	41.4
`fa_rep`	Light chain	0.131	0.254	18.2	33.0
`fa_rep`	All CDRs	0.238	0.369	33.5	49.4
`fa_rep`	H1–H3	0.165	0.295	22.8	38.4
`fa_rep`	H3	0.147	0.257	20.2	33.2
`fa_elec`	All antibody	0.269	0.348	37.6	47.1
`fa_elec`	Heavy chain	0.237	0.343	32.7	47.3
`fa_elec`	Light chain	0.182	0.277	26.3	37.1
`fa_elec`	All CDRs	0.247	0.337	35.3	45.8
`fa_elec`	H1–H3	0.226	0.328	32.5	44.8
`fa_elec`	H3	0.201	0.292	28.9	39.4
H-bond	All antibody	0.333	0.408	47.3	55.8
H-bond	Heavy chain	0.306	0.396	43.5	54.2
H-bond	Light chain	0.239	0.344	34.3	47.3
H-bond	All CDRs	0.320	0.395	45.8	53.2
H-bond	H1–H3	0.285	0.367	40.9	50.1
H-bond	H3	0.253	0.347	36.1	47.1
Solvation	Whole complex	0.202	0.297	–	–
Weighted interface energy	Interface total	0.374	0.440	–	–

Additional fixed selectors achieve DockQ@1/5 of 0.389/0.432 for the AF3 ranking score and 0.182/0.265 for the number of cross-chain atom pairs in the distance interval $[3, 4)$ Å. The latter result concerns this particular geometric statistic rather than geometric representations in general.

Further diagnostics of CP-score ties, residue masking, energy perturbations, and pose-cluster filtering are reported in Appendix C.

C Additional Signal Diagnostics

The following analyses supplement the fixed-score comparisons in Appendix B. They evaluate concrete scoring and filtering rules on AF3-AQA-1000 and are exploratory rather than preregistered tests. Unless stated otherwise, all 391 targets are included and intervals use paired target resampling.

C.1 Ties in Contact-Probability Scores

Exact ties are defined on stored scalar values under the fixed H3 mask. All five samples from every AF3 seed have identical CP-mean and CP-max summaries. CP-H3 mean has five top-scoring candidates and one top-scoring seed on average. CP-H3 maximum ties across 28.35 candidates and 5.67 seeds, spanning 3.71 pose clusters under the existing 5 Å clustering, with a mean tied DockQ range of 0.129. Top-score ties occur in all 391 targets.

Keeping one sample per seed leaves the same mean tie across 5.67 seeds for CP-H3 maximum. This separates the within-seed shared-score contribution from ties across different seeds. The audit establishes equality of these fixed-mask summaries; it does not establish that all underlying matrices are identical or that greater floating-point precision would resolve the ties.

C.2 Masking High-Confidence Residues

The masking analysis removes 1%, 5%, or 10% of residues selected by CP or PAE. Each condition is compared with 20 unique random residue sets matched for residue count and chosen to minimize mismatch in the remaining pair coverage. The strict sensitivity analysis requires mean absolute coverage error at most 0.01.

For CP-mean selection, removing 10% of CP-directed residues lowers DockQ@1/5 relative to matched random removal by $0.0820/0.0766$ across all 391 targets. The strict coverage-matched subset contains 352 targets. Directed-minus-random differences on this subset are -0.0869 (95% CI $[-0.1109, -0.0651]$) and -0.0809 ($[-0.1048, -0.0582]$). These comparisons support useful selection information in the removed positions under the tested CP-mean rule; they do not establish the contribution of those positions to AQA-Score.

C.3 Sensitivity to Energy Definition

The weighted interface total and equal-weight raw-scale atom total have mean within-target Spearman correlation 0.951 (95% CI $[0.945, 0.956]$). Despite this high correlation, atom-total minus interface-total DockQ@1/5 is -0.0266 ($[-0.0411, -0.0129]$) and -0.0219 ($[-0.0333, -0.0116]$). Thus, high rank correlation over the full candidate pool does not make these energy definitions interchangeable for Top-1 or Top-5 selection.

C.4 Deleting Extreme Atom Contributions

This analysis modifies the atom-total score by deleting either the atom with the largest repulsive contribution or the atom with the largest absolute total contribution. Each deletion is compared with 20 equal-count random interface-atom deletions per target. Targeted-minus-random DockQ@1 is -0.0018 (95% CI $[-0.0110, 0.0069]$) for the repulsive criterion and -0.0085 ($[-0.0199, 0.0026]$) for the absolute-total criterion. Both Top-5 intervals also include zero. These results do not show a reliable ranking improvement from either single-atom correction.

C.5 A Simple Close-Contact Statistic

The tested geometry score counts cross-chain atom pairs with distances in $[3, 4)$ Å and ranks larger counts first. It obtains DockQ@1/5 of $0.1820/0.2655$, below the fixed confidence and weighted interface-energy selectors on the same pool. This result concerns the specified shell-count statistic and does not establish a general limitation of geometry-based assessment.

C.6 Coverage Loss from a Single-Cluster Gate

Complex poses are clustered at 5 Å. Sorted AF3 seed identifiers are assigned alternately to sets for support estimation and held-out evaluation. The gate selects the single cluster with greatest support among estimation seeds, breaking cluster ties using the best confidence score in the estimation set. It then ranks held-out candidates in the selected cluster using the same confidence score. The direct baseline ranks all held-out candidates by that score. Held-out poses participate in clustering and evaluation, but their seeds do not contribute to support estimation.

The gate retains 21.8% of held-out candidates on average. Table 10 reports whether a qualifying candidate remains available after filtering. The best DockQ in the selected cluster is on average 0.220 below the unrestricted held-pool oracle. For iPTM, the additional mean gap from the selected-cluster oracle to the retrieved Top-1/Top-5 quality is $0.047/0.028$. This separates loss of candidate coverage from ranking errors within the retained cluster.

For iPTM, the Top-1/5 differences are $-0.0710/-0.0935$. Their 95% confidence intervals are $[-0.0933, -0.0501]$ and $[-0.1173, -0.0706]$, respectively. PAE-based ranking gives the same direction of change. The result concerns this hard single-cluster rule; it does not test soft use of ensemble support or retention of multiple clusters.

Table 10: Candidate availability after the single-cluster gate. Eligible held pools contain at least one candidate meeting the threshold before filtering.

DockQ threshold	Eligible held pools	Retaining a candidate	Retention (%)
0.23	313	174	55.6
0.49	245	152	62.0
0.80	121	82	67.8

C.7 Exploratory Applicability Analyses

Additional analyses examine H3 length and the fractions of positively charged, polar, and aromatic residues as target-level attributes. Native contact count is used only as a reference-based diagnostic. Controlled regressions include target-level covariates and antigen-family block intervals where available. No composition interaction survives multiplicity correction. Predicted charge complementarity is available for only 68 targets, and its controlled family interval includes zero. These analyses do not support categorical claims about antibody subgroups.

The broader signal comparisons remain exploratory and are not globally corrected for feature selection. Their uncertainty estimates concern variation across targets, not model-training randomness. Both public pools evaluate the same biological targets, and all conclusions concern reference-structure quality rather than binding affinity or function.

D AQA-Score Implementation Details

This section supplements the architecture and training objectives in Section 4. Pretraining-data construction and provenance are described in Appendix A.1.

D.1 Pretraining Configuration

The single-chain encoder (Module A) uses 16 atom features, 27 residue features, and 16 radial basis functions. Its intra-chain atom graph uses a 4.5 Å neighborhood and three message-passing layers, followed by atom-to-residue mean pooling. Stage 1 trains interface-site prediction for five epochs with positive-site binary cross-entropy plus 0.25 times background binary cross-entropy. Training uses global batch size 128 and AdamW with learning rate 3×10^{-4} and weight decay 10^{-4}. The resulting development macro site AUPRC is 0.388.

The cross-chain compatibility model (Module B) uses two bidirectional cross-attention layers with four heads and a rank-32 pair head. For residue-contact supervision, positive pairs have Cα distance below 8 Å, negative pairs have distance at least 12 Å, and pairs in the intervening range are excluded. Stage 2 starts from the Stage 1 checkpoint and samples at most 512 positive residue pairs per interface, with four negatives per positive. Where available, half of the negatives are drawn from the hard-negative pool. Training runs for 10 epochs with learning rate 2×10^{-4} and site auxiliary weight 0.2. The epoch-8 checkpoint selected by development loss reaches pair AUPRC 0.447, compared with 0.351 for the product of the two site probabilities. This checkpoint initializes the frozen binding representations used in candidate assessment.

D.2 Sparse Pairs and Candidate Features

The pair union in Equation 5 combines cross-chain residue pairs within 10 Å minimum heavy-atom distance, bidirectional Top-8 compatibility proposals, and bidirectional Top-8 AF3 contact-probability proposals. The union is deduplicated before applying a 4,096-pair cap. Truncation prioritizes formed contacts below 5 Å, followed by the remaining geometry, CP, and prior evidence. Prior- or CP-proposed pairs outside the formed-contact set are retained as unformed hypotheses. Reference contacts never propose pairs or alter the candidate graph.

Each pair combines pretrained and partner-conditioned residue states with a frozen compatibility logit, 19 distance/orientation features, five directional PAE/CP and availability features, and 62 candidate-context features. The context includes residue confidence, antibody-region identity, and

local Rosetta summaries. The candidate interface model performs two pair–residue interaction rounds at hidden width 128. Separate attention pooling summarizes formed contacts and unformed hypotheses before quality prediction. An empty pool contributes an all-zero pooled vector.

D.3 Quality Aggregation and Local Supervision

Let F_{ci}, I_{ci}, and L_{ci} denote the reference AbAg-DockQ components defined in Equation 8. The contact-recovery head predicts $\widehat{F}_{ci}$ through a sigmoid. Interface and antibody residue heads predict non-negative errors $\widehat{e}^{I}_{ci,r}$ and $\widehat{e}^{L}_{ci,r}$. For a nonempty candidate-defined residue set M^{t}_{ci}, the RMSD estimate is

$$\widehat{R}^{t}_{ci} = \left[\frac{1}{|M^{t}_{ci}|} \sum_{r \in M^{t}_{ci}} \left(\widehat{e}^{t}_{ci,r}\right)^2 \right]^{1/2}, \qquad t \in \{I, L\}. \tag{13}$$

The interface set M^{I}_{ci} contains residues participating in candidate cross-chain pairs within 10 Å, while M^{L}_{ci} contains all antibody residues. If the candidate interface set is empty, predicted interface RMSD is fixed to 20 Å. Reference residue correspondence determines valid local supervision without changing these inference masks.

With RMSDs measured in Å, the DockQ transforms give $\widehat{I}_{ci} = [1 + (\widehat{R}^{I}_{ci}/1.5)^2]^{-1}$ and $\widehat{L}_{ci} = [1 + (\widehat{R}^{L}_{ci}/8.5)^2]^{-1}$. Together with $\widehat{F}_{ci}$, these form the ranking score in Equation 6. Component regression for target c is

$$\begin{aligned} \mathcal{L}_{\mathrm{comp},c} = \frac{1}{N_c} \sum_{i=1}^{N_c} \Big[&(\widehat{F}_{ci} - F_{ci})^2 + (\widehat{I}_{ci} - I_{ci})^2 \\ &+ (\widehat{L}_{ci} - L_{ci})^2 \Big]. \end{aligned} \tag{14}$$

Valid local interface and antibody errors use SmoothL1 regression between predicted and reference errors in $\log(1 + e)$ space. Contact supervision is restricted to contacts formed by the candidate: a contact is positive if it also occurs in the reference structure and negative otherwise. Balanced binary cross-entropy assigns equal aggregate class weight when both classes are present. Missing native contacts affect the contact-recovery label but are not inserted into the forward graph.

D.4 Candidate-Assessment Optimization

The complete objective combines component regression, local supervision, and pool-wise selection:

$$\begin{aligned} \mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{comp}} + \lambda_{\mathrm{contact}} \mathcal{L}_{\mathrm{contact}} + \lambda_I \mathcal{L}_{I,\mathrm{local}} \\ + \lambda_L \mathcal{L}_{L,\mathrm{local}} + \lambda_{\mathrm{select}} \mathcal{L}_{\mathrm{select}}. \end{aligned} \tag{15}$$

Here, $\mathcal{L}_{\mathrm{select}}$ is the selection objective in Equation 7. Selection uses all 100 candidates of each training target, and losses are averaged equally across targets. We use

$$\begin{aligned} \lambda_{\mathrm{contact}} = \lambda_I = \lambda_L = 0.2, \qquad \tau = 0.1, \\ \lambda_{\mathrm{select}} \simeq 29.6387. \end{aligned} \tag{16}$$

The selection coefficient is twice the approximately 14.8193 scale obtained from a training-set gradient probe.

Stage 3 trains for 20 epochs using AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, and gradient clipping at 1.0. Each optimizer step uses four complete target pools across eight GPUs. Modules A and B, including the compatibility prior, remain frozen; the candidate interface model and quality heads are trainable. The instantiated model contains 1,582,556 parameters, of which 386,567 are trainable in Stage 3. At inference, candidates are scored independently before within-target sorting.

Table 11: Overall selection performance on AF3-AQA-100. DockQ@1 and DockQ@5 measure the AbAg-DockQ of the first-ranked candidate and the best candidate among the selected five, respectively. SR@.23/1 and SR@.23/5 are acceptable-quality success rates, reported as percentages over all 391 targets. Bold marks the largest displayed point estimate among practical selectors; Random and Oracle are reference rows. PAE-H is the unidirectional H1–H3-to-antigen mean; interface energy is the ref2015-weighted total.

Method	DockQ@1	DockQ@5	SR@.23/1	SR@.23/5
Random	0.245	0.348	35.3	48.1
iPTM	0.345	0.380	49.6	52.9
ipSAE (Dunbrack Jr, 2025)	0.330	0.385	48.6	52.2
PAE-H mean	0.342	0.382	49.6	52.4
CP-H3 max	0.337	0.383	48.1	51.9
Interface energy	0.328	0.385	46.3	51.7
DeepRank-Ab	0.333	**0.404**	48.8	**56.0**
ABAG-Rank	0.285	0.354	41.7	49.4
ARID-sf	0.227	0.312	32.7	42.7
AQA-Score	**0.363**	0.402	**51.9**	54.7
Oracle	0.502	0.502	70.1	70.1

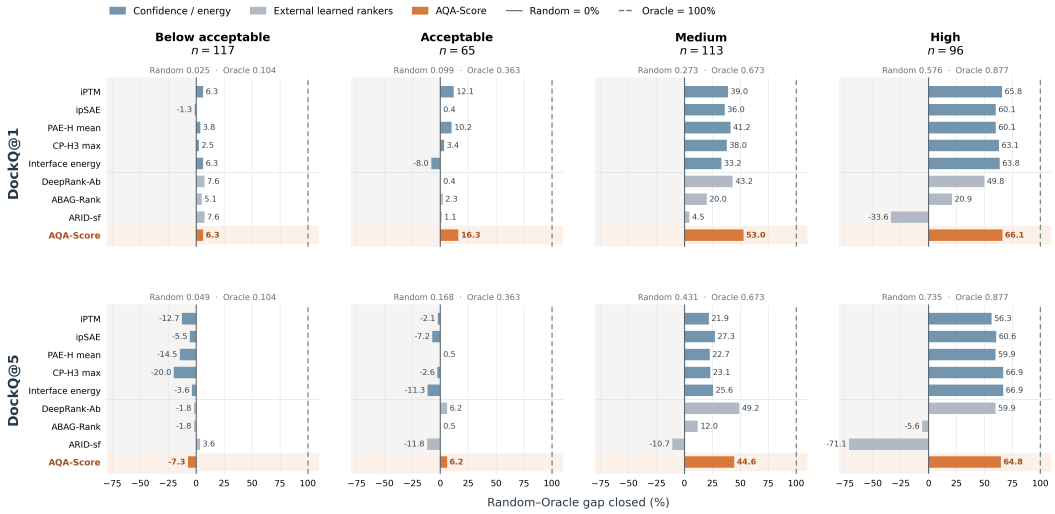


Figure 5: Normalized visualization of the stratified results in Table 12. Columns correspond to oracle-quality strata; rows show DockQ@1 and DockQ@5. Bars report the percentage of the Random–Oracle gap closed: $100(D_{\text{method}} - D_{\text{Random}})/(D_{\text{Oracle}} - D_{\text{Random}})$, computed separately for each stratum and metric. Random and Oracle correspond to 0 and 100, respectively. Negative values indicate performance below Random. Values are derived from the rounded table entries.

E Detailed Structure-Selection Results

Table 11 reports overall selection performance on AF3-AQA-100, complementing the stratified comparison in Table 4. All metrics give equal weight to the 391 test targets. DockQ@1/5 uses the AbAg-DockQ labels defined in Section 3.2.

Figure 6 shows that gains concentrate in Medium pools. These 113 targets constitute 28.9% of the benchmark and account for 88.6% of the summed Top-1 gain and 70.7% of the summed Top-5 gain over iPTM. AQA-Score reaches DockQ@1/5 of $0.485/0.539$ in this stratum, compared with $0.429/0.484$ for iPTM and $0.446/0.550$ for DeepRank-Ab. Thus, improved first-choice quality coexists with a remaining Top-5 advantage for DeepRank-Ab.

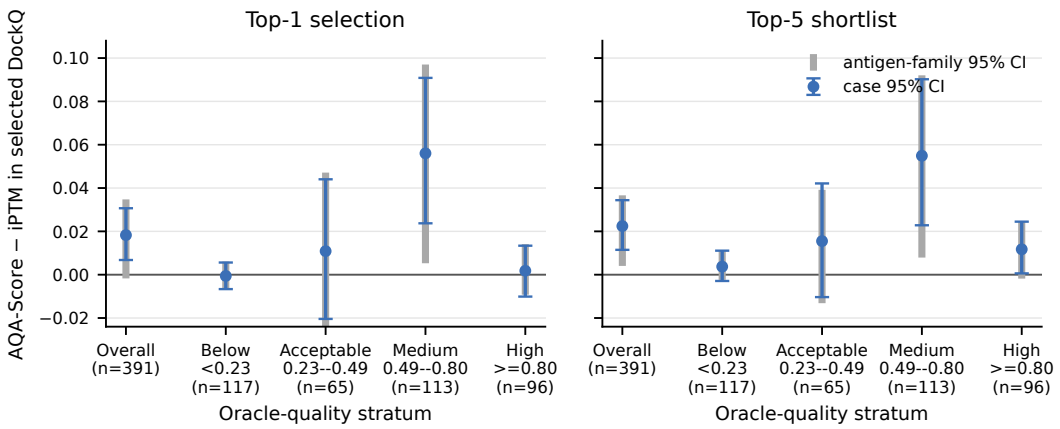


Figure 6: AQA-Score minus iPTM in selected DockQ across AF3-AQA-100 oracle-quality strata. Blue intervals resample targets; gray intervals resample antigen families. Overall estimates use all 391 targets, and stratum sizes are shown below their point estimates.

Changes in other strata are smaller or less certain. In Acceptable pools, mean selected DockQ increases, but acceptable-quality Top-5 success falls from 35.4% to 32.3%. High-pool Top-1 changes only slightly. The tables below report complete point estimates, paired intervals, and retrieval conditional on candidate availability.

E.1 Complete Results by Pool-Oracle Stratum

Table 12 gives the complete AF3-AQA-100 comparison within each pool-oracle stratum. The overall results are reported in Table 11. Strata are defined using reference quality only for analysis, and are not available to the selectors. Success rates in the table use all targets in the corresponding stratum as the denominator. Below-acceptable pools contain no candidate with DockQ ≥ 0.23, so their success rates are zero for every method.

Table 12: Complete AF3-AQA-100 results by pool-oracle stratum. Success rates are percentages within each stratum. Bold marks the largest displayed point estimate among practical selectors, retaining displayed ties; Random and Oracle are reference rows. Zero success rates are not highlighted.

Method	DockQ@1	DockQ@5	SR@.23/1	SR@.23/5
Below acceptable ($n = 117$)				
Random	0.025	0.049	0.0	0.0
iPTM	0.030	0.042	0.0	0.0
ipSAE	0.024	0.046	0.0	0.0
PAE-H mean	0.028	0.041	0.0	0.0
CP-H3 max	0.027	0.038	0.0	0.0
Interface energy	0.030	0.047	0.0	0.0
DeepRank-Ab	**0.031**	0.048	0.0	0.0
ABAG-Rank	0.029	0.048	0.0	0.0
ARID-sf	**0.031**	**0.051**	0.0	0.0
AQA-Score	0.030	0.045	0.0	0.0
Oracle	0.104	0.104	0.0	0.0
Acceptable ($n = 65$)				
Random	0.099	0.168	16.9	29.2
iPTM	0.131	0.164	**26.2**	**35.4**
ipSAE	0.100	0.154	21.5	30.8
PAE-H mean	0.126	0.169	23.1	33.8
CP-H3 max	0.108	0.163	18.5	30.8

Continued on next page

Table 13: Paired AQA-Score–iPTM differences on AF3-AQA-100. Intervals are 95% confidence intervals.

Stratum	Metric	Difference	Target CI	Family CI
Overall	Top-1	0.0183	$[0.0068, 0.0307]$	$[-0.0017, 0.0348]$
	Top-5	0.0224	$[0.0115, 0.0344]$	$[0.0041, 0.0366]$
Below acceptable	Top-1	-0.0006	$[-0.0067, 0.0056]$	$[-0.0062, 0.0060]$
	Top-5	0.0037	$[-0.0029, 0.0111]$	$[-0.0024, 0.0109]$
Acceptable	Top-1	0.0109	$[-0.0204, 0.0440]$	$[-0.0286, 0.0471]$
	Top-5	0.0155	$[-0.0103, 0.0421]$	$[-0.0132, 0.0392]$
Medium	Top-1	0.0560	$[0.0237, 0.0908]$	$[0.0053, 0.0970]$
	Top-5	0.0549	$[0.0228, 0.0902]$	$[0.0079, 0.0920]$
High	Top-1	0.0018	$[-0.0101, 0.0134]$	$[-0.0092, 0.0141]$
	Top-5	0.0117	$[0.0006, 0.0245]$	$[-0.0018, 0.0243]$

Table 12 continued

Method	DockQ@1	DockQ@5	SR@.23/1	SR@.23/5
Interface energy	0.078	0.146	10.8	29.2
DeepRank-Ab	0.100	**0.180**	16.9	**35.4**
ABAG-Rank	0.105	0.169	15.4	33.8
ARID-sf	0.102	0.145	15.4	24.6
AQA-Score	**0.142**	**0.180**	**26.2**	32.3
Oracle	0.363	0.363	100.0	100.0
Medium ($n = 113$)				
Random	0.273	0.431	45.1	70.8
iPTM	0.429	0.484	71.7	77.9
ipSAE	0.417	0.497	70.8	77.9
PAE-H mean	0.438	0.486	73.5	77.0
CP-H3 max	0.425	0.487	70.8	77.0
Interface energy	0.406	0.493	69.0	77.0
DeepRank-Ab	0.446	**0.550**	77.0	**88.5**
ABAG-Rank	0.353	0.460	62.8	74.3
ARID-sf	0.291	0.405	49.6	65.5
AQA-Score	**0.485**	0.539	**79.6**	85.8
Oracle	0.673	0.673	100.0	100.0
High ($n = 96$)				
Random	0.576	0.735	79.2	92.7
iPTM	0.774	0.815	**100.0**	**100.0**
ipSAE	0.757	0.821	**100.0**	**100.0**
PAE-H mean	0.757	0.820	**100.0**	**100.0**
CP-H3 max	0.766	**0.830**	**100.0**	**100.0**
Interface energy	0.768	**0.830**	**100.0**	**100.0**
DeepRank-Ab	0.726	0.820	96.9	**100.0**
ABAG-Rank	0.639	0.727	85.4	90.6
ARID-sf	0.475	0.634	64.6	80.2
AQA-Score	**0.775**	0.827	**100.0**	**100.0**
Oracle	0.877	0.877	100.0	100.0

E.2 Paired Differences and Conditional Retrieval

Table 13 reports AQA-Score minus iPTM with both target- and antigen-family-resampled intervals. The Medium stratum has positive Top-1 and Top-5 intervals under both schemes. The overall Top-1 family interval includes zero, while the overall Top-5 interval remains positive. Differences and intervals are computed from unrounded target-level results, so they need not equal differences between the displayed rounded means.

Table 14 gives AQA-Score retrieval counts among targets containing a qualifying candidate. On AF3-AQA-100, the medium-quality Top-1 count increases from 159 for iPTM to 170 for AQA-Score:

Table 14: AQA-Score retrieval conditional on candidate availability. Entries are successful targets divided by eligible targets; the denominator changes with pool and threshold.

Pool	DockQ threshold	Top-1	Top-5
AF3-AQA-100	0.49	170/209	179/209
	0.80	43/96	69/96
AF3-AQA-1000	0.23	224/325	243/325
	0.49	185/260	205/260
	0.80	57/137	77/137

14 iPTM failures are recovered and three successes are lost. The corresponding Top-5 counts are 169 and 179. At the high-quality threshold, Top-1 changes from 44 to 43, whereas Top-5 increases from 56 to 69.

On the complete AF3-AQA-1000 pool, AQA-Score obtains DockQ@1/5 of $0.4004/0.4518$, compared with an oracle of 0.6015. The complete-pool comparison includes cached fixed confidence and energy scores. External structure-based rankers are evaluated on AF3-AQA-100 and are not included in the full AF3-AQA-1000 comparison.

E.3 Scope of the Available Ablations

Table 5 reports two input-channel and two loss ablations on AF3-AQA-100. The input ablations remove AF3 confidence or local Rosetta energy features. The loss ablations replace separate component regression with regression on the mean predicted score, or remove the pool-wise selection loss $\mathcal{L}_{\text{select}}$.

All variants are evaluated on the same candidates and reference quality labels for the same 391 targets. Each result is from a single training run, so these comparisons do not quantify training-seed variation. They also do not separately isolate the contributions of binding-related pretraining, unformed hypotheses, or local auxiliary supervision.

F Sampling-Budget Protocol and Results

The budget analysis uses only AF3-AQA-1000. A continuous PCG64 random-number stream with seed 20260925 generates one permutation of the 1,000 candidates for each target and repeat. Prefixes of the same permutation define budgets $N \in \{20, 100, 250, 500, 1000\}$, with 100 repeats per target. Permutation operates on individual candidates, rather than grouping the five samples from an AF3 seed. All methods receive identical prefixes within a target and repeat, and score definitions and checkpoints remain fixed across budgets.

For each budget, repeated measurements are averaged within a target before computing target-level means and bootstrap intervals. The repeats are not treated as additional biological targets. The $N = 100$ prefix is a subset of AF3-AQA-1000 and must not be identified with the independently generated public AF3-AQA-100 pool.

The full 1,000-candidate prefix reproduces the complete-pool AQA-Score result, and a replay of the cached iPTM protocol matches to tolerance 10^{-12}. The subpool oracle is non-decreasing along every nested permutation. Selected quality need not be monotonic, because a newly sampled candidate can receive a higher score while having lower reference quality. This distinction underlies the remaining gap between additional sampling coverage and the quality ultimately retrieved.

Figure 7 reports selection performance across budgets. AQA-Score improves from DockQ@1/5 of $0.353/0.401$ at $N = 20$ to $0.399/0.449$ at $N = 250$. Performance changes little thereafter, reaching $0.401/0.452$ at $N = 500$ and $0.4004/0.4518$ at $N = 1000$. Over the same range from 20 to 1,000 candidates, the mean oracle increases from 0.442 to 0.6015.

On the complete 1,000-candidate pool, iPTM obtains $0.388/0.431$, PAE-H mean obtains $0.396/0.437$, CP-H3 maximum obtains $0.394/0.435$, and weighted interface energy obtains $0.374/0.440$. External learned rankers are not included in this comparison. The remaining oracle gap shows that larger

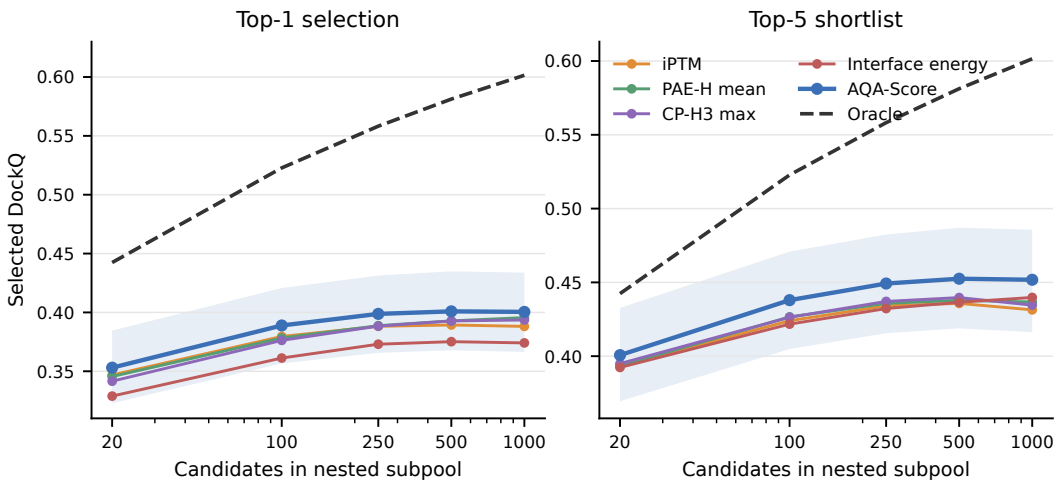


Figure 7: Structure selection under increasing sampling budgets. Methods receive identical nested subsets of AF3-AQA-1000. The blue band gives target-bootstrap uncertainty for AQA-Score. Selected quality begins to plateau while the best available candidate continues to improve.

candidate pools provide additional attainable quality that is only partly recovered by the evaluated scoring rules.