# MultiAbRank: Benchmarking De Novo Antibody Design Under Multi-Objective Constraints

**Abdula Maher** [*1] **Anand Shanker** [*2] **Mikayla M. Young** **Sydney Eze** [3] **Adnan Ahmed** [4] **Kiran Nijjer** [5]

## Abstract

In recent years, diffusion and sequence-first models have been created to rapidly produce antibody candidates. Through physics-based scoring metrics that analyze features such as structural stability and interface geometry, methods for benchmarking and evaluating these candidates have also emerged. Proper evaluation of antibody candidates and computational design models requires examining binding, structural plausibility, biological realism, and novelty. However, de novo antibody evaluation methods rely on reconstruction-based criteria or structural confidence proxies, and do not satisfy these objectives, failing to properly address the underlying challenge in de novo antibody design that plagues *in vitro* and *in vivo* translation. We introduce a benchmark for de novo antibody design models with three complementary tasks, examine current models on supported tasks, and evaluate candidates produced from these three tasks based on a multi-objective scoring method. We also expose weaknesses in antibody candidate evaluation, offer a standardized framework for future computational design models.

## 1. Introduction

Computer-aided drug design has seen the development of large protein language models, diffusion-based generators, and structure-aware learning methods integrated into robust models that enable the generation of novel antibody sequences that satisfy biologically relevant intrinsic fitness criteria such as foldability, stability, and expression potential. These properties of sequences can be manipulated and evaluated without resource and time-extensive wet-lab discovery and validation. As a result, *in silico* antibody generation has shifted from sequence reproduction toward claims of de novo therapeutic design (Meng et al., 2024). Despite these advances, existing benchmarks evaluate narrow properties in isolation, rather than the multi-objective criteria that govern experimental viability and clinical translation (Liu et al., 2025).

In practice, affinity-related scores, structural stability measures, and novelty constraints can exert competing pressures on model output, such that improvement along one axis coincides with decline along another. Therefore, models that perform well under existing benchmarks fail when tasked with designing antibodies that are simultaneously novel, antigen-conditioned, specific, and biologically realistic in wet-lab environments. This implies that evaluating candidates using a single-metric paradigm is misaligned with the underlying physical and biological constraints that govern antibody behavior, thereby obscuring meaningful differences between model approaches (Chungyoun et al., 2024).

In our work, we evaluate de novo antigen-conditioned antibody designs under multi-objective constraints by aggregating complementary metrics into a comprehensive scoring protocol. We propose a benchmark comprising three complementary tasks: (i) Novel CDR-H3 generation, (ii) Framework region design, and (iii) Global antibody sequence and structure design. The novelty of our work lies in evaluating de novo antibody design across a rigorous set of biologically relevant metrics to bridge the gap between *in silico* antibody design and practical, real-world application.

## 2. Related Work

Current antibody generation models fall into two main categories: sequence-first and structure-first. Sequence-first models generate or predict antibody amino acid sequences, while structure-first models predict or optimize the geometry of the antibody-antigen complex, optimizing affinity, binding, and biological realism (Meng et al., 2024). AbBiBench,

[*]Equal contribution [1]Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada [2]Department of Computer Science, Purdue University, West Lafayette, Indiana, United States of America [3]Department of Neuroscience and Behavior, Barnard College, New York City, New York, United States of America [4]York University, Toronto, Ontario, Canada [5]Stanford University, Stanford, California, United States of America. Correspondence to: Abdula Maher <abdula.maher@mail.utoronto.ca>.

*Proceedings of the Workshop on Structured Data for Health at the $43^{rd}$ International Conference on Machine Learning*, Seoul, South Korea. Copyright 2026 by the author(s).

AbRank, and FLAb are benchmarks for targeting affinity maturation, pairwise binding ranking, and multi-property fitness landscapes, respectively, but do not assess the full scope of de novo antibody design (Zhao et al., 2025b) (Liu et al., 2025) (Uçar et al., 2024). Reconstruction-based metrics such as AAR give limited insight (Uçar et al., 2024) (Li et al., 2025) and assume known sequences are optimal, penalizing novelty (Kim et al., 2024). RMSD shares this limitation. Structural confidence proxies pLDDT and PAE indicate reliability but lack antigen context (Joubbi et al., 2025) (Varadi et al., 2022), while HADDOCK captures docking quality, but not specificity or developability (Kurkcuoglu et al., 2018). Novelty and biological realism can be assessed via OAS (Olsen et al., 2022) and the Therapeutic Antibody Profiler (Raybould et al., 2019).

## 3. Methods

### 3.1. Baseline Model Methods and Workflows

Sequence-first models, IgLM and ProtT5-FT, are antibody-specialized protein language models used here for antigen-conditioned CDR-H3 and framework infilling. IgLM (decoder-only, masked infilling) and ProtT5-FT (encoder-decoder, OAS fine-tuned) serve as sequence-first baselines.

Structure-first models, RFDiffusion and BindCraft, instantiate diffusion- and AF2-Multimer–based pipelines for antigen-conditioned binder design; RFDiffusion generates CDR backbones or full VH/VL variable domains in complex with the antigen followed by ProteinMPNN inverse folding. BindCraft performs AF2-Multimer hallucination and gradient-based sequence optimization to obtain putative binders.

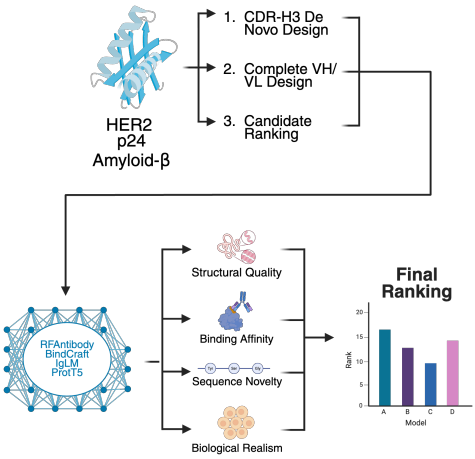


*Figure 1.* Overview of the benchmark pipeline.

### 3.2. Benchmarking Tasks

**Task 1: CDR-H3 De Novo Design:** Given a target antigen structure, a target epitope, and a fixed antibody framework, the model under test (MUT) designs a novel CDR-H3 loop. Task 1 evaluates whether the MUT can construct a functional binding interface. The MUT should generate 30 CDR-H3 sequences per antigen.

**Task 2: Framework Design:** Given a target antigen structure and a CDR-H3 sampled from Task 1, the MUT designs the VH/VL framework (excluding the provided CDR-H3). Task 2 assesses whether models can construct compatible frameworks that maintain the CDR-H3 context. Framework design is repeated for each of the CDR-H3 from Task 1.

**Task 3: Complete VH/VL De Novo Design:** Task 3 evaluates binding efficacy, folding, sequence realism, and biological realism. The MUT generates 30 complete VH/VL sequences per antigen.

## 4. Evaluation

We establish thresholds across four complementary objectives. Candidates failing any single threshold are disqualified and assigned a composite score of zero. Candidates clearing all thresholds receive a weighted composite score. Information on all scoring methods, threshold selection, weighting rationale, and weight-sensitivity ablation can be found in section A.

### 4.1. Scoring Metrics

**Novelty Score** In Task 1, novelty is measured on the CDR regions as a whole. For Task 2 and 3, novelty is calculated for the VH/VL sequences individually and averaged. Novelty is measured by sequence identity to the nearest OAS neighbor, using KA-Search (Olsen et al., 2023). While an antibody candidate can have good scores, if it already exists, the generation was unnecessary.

**Biological Realism Score** is computed as: $S_{\text{bio}} = \alpha \cdot S_{\text{TAP}} + (1-\alpha) \cdot S_{\text{imm}}$. $S_{\text{TAP}} = 1 - \frac{N_{\text{amber}} + 2 \cdot N_{\text{red}}}{10}$, with candidates scoring below 0.8 disqualified, and the immunogenicity score is $S_{\text{imm}} = 1 - \frac{f_{\text{strong}}}{0.1}$. For our scoring, $\alpha = 0.5$.

**Structural Score** is the mean of pLDDT and PAE subscores: $S_{\text{struct}} = \frac{1}{2}\left(S_{\text{pLDDT}} + S_{\text{PAE}}\right)$. $S_{\text{pLDDT}} = \frac{\text{pLDDT} - 75}{25}$ (candidates below pLDDT 80 are disqualified, yielding a range of $[0.2, 1.0]$) and $S_{\text{PAE}} = 1 - \frac{\text{PAE}}{10}$ (candidates above PAE 10 are disqualified, yielding a range of $[0, 1]$).

**Binding Score** was calculated using PRODIGY (Vangone and Bonvin, 2017) (Xue et al., 2016) on the top five HADDOCK poses for a given antigen-antibody pair, and applying the formula found in section A.4.

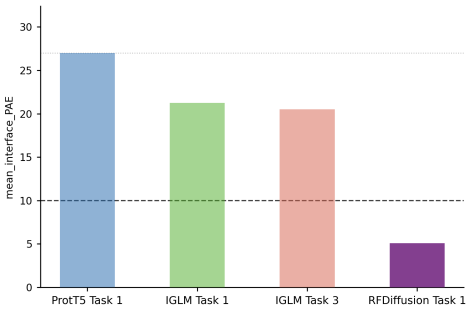


*Figure 2.* Mean PAE scores for the models tested. Models with a PAE score $> 10$ were disqualified.

### 4.2. Composite Score

For candidates clearing all thresholds $S_{\text{composite}} = \gamma \cdot S_{\text{nov}} + \delta \cdot S_{\text{bio}} + \epsilon \cdot S_{\text{bind}} + \zeta \cdot S_{\text{struct}}$ with default weights $(\gamma, \delta, \epsilon, \zeta) = (0.20, 0.40, 0.20, 0.20)$, which can be adjusted to stress-specific objectives.

### 4.3. Experimental Setup

We curated a panel of three high-value therapeutic targets spanning antigen categories across cancer (HER2), viruses (p24), and neurodegenerative diseases (amyloid-$\beta$) for the primary benchmark tasks. When applicable, all models were provided with identical antigen inputs, either as antigen sequences or the corresponding structure to the sequences.

### 4.4. Databases for Fine-Tuning

For sequence-first models, we used the Observed Antibody Space (OAS) and the Structural Antibody Database (SAbDab) for pretraining and fine-tuning. Both were created by the Oxford Protein Informatics Group. (Olsen et al., 2022) (Dunbar et al., 2014) (Schneider et al., 2022)

## 5. Results

To assess whether diffusion-generated antibodies exhibit both structural plausibility and predicted antigen binding, we evaluated AlphaFold structural confidence metrics alongside HADDOCK docking energies across the HER2 antibody design library. When scoring, if all antibodies from a particular model failed in one portion of the scoring (e.g., the structural score), the model was automatically disqualified.

As shown in Figure 2, IgLM and ProtT5 both failed Task 1, as all their generated antibodies had PAE scores exceeding 10. IgLM's average PAE score was 21.241, and ProtT5's average PAE score was 27.001. On Task 2, all candidates generated by ProtT5 failed to parse through TAP; ProtT5 failed on Task 2. For Task 3, IgLM-generated antibodies all had PAE scores greater than 10, again disqualifying them. BindCraft also failed Task 1 and 3 due to internal constraints,

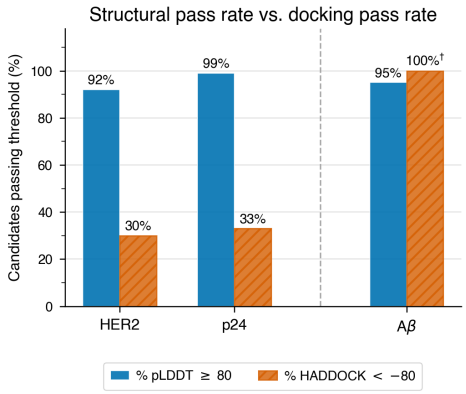


*Figure 3.* The A$\beta$ docking pass rate of 100% reflects the permissiveness of the disordered antigen topology rather than genuine high-affinity binding; the $-80$ threshold is non-discriminating for this antigen class.

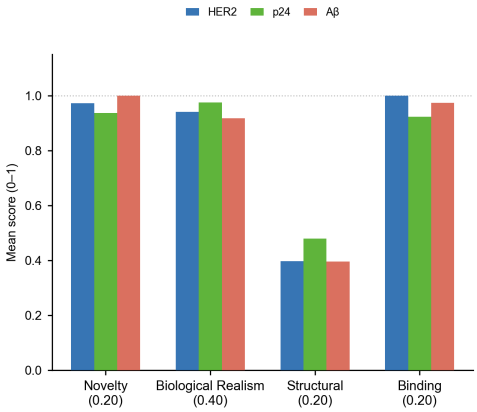


*Figure 4.* Evaluation pipeline scores for RFDiffusion-generated designs.

for which more information is provided in section A.6.

Most candidates displayed high structural confidence, with mean pLDDT values exceeding 80, but docking scores ranged from below -140 to above 0, indicating that structural plausibility alone does not guarantee strong binding. Candidates with lower interface PAE generally exhibited favorable docking energies, suggesting that accurate prediction of antibody–antigen orientation is a key determinant of binding quality.

For Task 1, RFDiffusion-guided CDR design, 29 of 30 (96.7%) candidates in the p24 cohort qualified, yielding a mean composite score of 82.9%. Against amyloid-$\beta$, 28 of 30 candidates qualified (93.3%), yielding a cohort mean composite score of 78.4%. The HER2 outputs exhibited the greatest generative attrition, with only 23 of 30 candidates qualifying (76.7%); this elevated disqualification rate is directly reflected in the cohort mean score of 65.2%.

*Table 1.* Structural confidence, binding performance, and composite score components across antigens for RFDiffusion-generated designs.

| ANTIGEN | $n$ | AF2-MULTIMER | | PRODIGY, TOP-5 POSES | | | SCORE COMPONENTS | | |
|---|---|---|---|---|---|---|---|---|---|
| | | PLDDT | INT. PAE | MEAN±SD | RANGE | BINDING | NOVELTY | BIOSCORE | **COMPOSITE** |
| HER2 | 23 | $82.8 \pm 1.6$ | $5.18 \pm 0.51$ | $-17.4 \pm 0.9$ | $[-18.7, -15.0]$ | 1 | 0.972 | 0.941 | **65.2%** |
| p24 | 29 | $86.5 \pm 1.5$ | $4.99 \pm 0.54$ | $-14.9 \pm 1.2$ | $[-18.2, -13.0]$ | 0.913 | 0.937 | 0.975 | **82.9%** |
| A$\beta$ | 28 | $82.0 \pm 1.0$ | $4.85 \pm 0.43$ | $-19.3 \pm 2.7$ | $[-24.9, -10.9]$ | 0.975 | 1.000 | 0.917 | **78.5%** |

An analogous trend arose in p24 designs: only a fraction of structurally plausible antibodies achieved strongly favorable docking scores, indicating that structural correctness alone is insufficient for predicting functional binding interfaces. As seen in Figure 3, this disconnect is not antigen-specific, and plausibility alone does not ensure strong antigen binding.

Our observations demonstrate the need for multi-objective evaluation in de novo designs. While diffusion-based methods reliably generate structurally plausible candidates, filtering using interface geometry and binding proxies is needed to identify designs with strong predicted antigen affinity.

## 6. Discussion

RFDiffusion demonstrated antigen-dependent generative feasibility across the three target cohorts, with pass rates ranging from 76.7% to 96.7%. The elevated attrition observed in the HER2 cohort was driven primarily by structural failures: four candidates fell below the pLDDT threshold, and three failed sequence parsing. However, when composite scores are restricted to qualified candidates, HER2 (0.8501) approaches p24 (0.8580) and exceeds amyloid-$\beta$ (0.8408), indicating that the pipeline produces comparably high-quality designs across all antigens when generative trajectories succeed. The gap between HER2 and the other targets is therefore attributable almost entirely to disqualification, consistent with HER2's well-folded, conformationally constrained epitope, which presents a more challenging geometry for diffusion-guided CDR design than the capsid surface of p24 or the disordered topology of amyloid-$\beta$. The complete failure of BindCraft to produce any accepted designs across all three target antigens further underscores this point: AF2-Multimer hallucination-based pipelines face a substantially steeper feasibility barrier under multi-objective filtering than diffusion-guided CDR design, and their trajectory attrition is not recoverable solely through downstream sequence redesign.

Decomposition of score components among qualified candidates reveals further antigen-specific trends. The p24 cohort achieved the highest mean pLDDT score (86.5±1.5), reflecting the greater structural confidence AlphaFold assigns to capsid-targeting designs, while amyloid-$\beta$ and HER2 candidates were comparably lower (82.0±1.0 and 82.8±1.6, respectively). In contrast, interface PAE scores were similar across all three antigen classes, suggesting that once a candidate clears the qualification threshold, interface geometry is predicted with comparable reliability regardless of target. Notably, the amyloid-$\beta$ cohort achieved perfect novelty scores across all qualified candidates, whereas p24 and HER2 exhibited slight convergence toward known sequence motifs, possibly as a consequence of scaffold-imposed constraints on the CDR sequence space. Critically, high AlphaFold pLDDT scores did not reliably predict favorable docking energies for either HER2 or p24, confirming that structural confidence is decoupled from predicted binding and that pLDDT and PAE thresholds, while necessary, are insufficient criteria for de novo antibody evaluation. This decoupling was not antigen-specific and underscores the importance of the threshold-and-aggregate scoring protocol employed here. By disqualifying candidates that fail any individual threshold across novelty, biological realism, binding, and structure before computing composite scores, the protocol surfaces model-specific failure modes that would be obscured under single-metric assessment. Sensitivity analysis over weight configurations (Appendix A.8) also shows antigen-level rankings are invariant to weight perturbation, supporting the robustness of composite score.

## 7. Conclusion

There are limitations of this evaluation framework that warrant consideration. The scoring protocol's fixed thresholds and tool dependencies reflect current practice but introduce methodological biases. External validation remains constrained by a small anchor set insufficient to establish robust score-to-affinity correlations. The computationally intensive pipeline limits the number of candidates that can be evaluated per antigen and favors generative models that integrate naturally with prevailing structure prediction tools, raising the concern that observed performance differences may reflect resource or pipeline compatibility advantages over algorithmic merit. More broadly, the absence of a reproducible community benchmarking standard means that results of this kind cannot yet reliably distinguish model-specific progress from pipeline-dependent artifacts. Future work should address this gap, expand the antigen panel, and investigate closer integration of generation and evaluation.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning as applied to computational biology. The benchmark and scoring protocol introduced here are intended to improve the rigor of de novo antibody design evaluation, with potential downstream implications for therapeutic drug discovery. We note that no computationally generated antibody candidate should be considered clinically viable without experimental validation; this work is positioned as a step toward more principled model comparison, not a substitute for wet-lab confirmation. Open publication of benchmark tasks, scoring thresholds, and evaluation pipelines is intended to support reproducibility across resource-constrained research settings.

## A. Appendix

### A.1. Novelty Scoring Formula

Our novelty formulas found below were partially motivated by earlier papers and their use of sequence similarity in pruning the Observed Antibody Space. (Li et al., 2025) (Zhao et al., 2025a)

Sequences that are extremely similar ($\geq 90\%$ to those in the OAS were penalized slightly, though as long as they were not an exact match, they were accepted. Sequences that were dissimilar ($70 \leq s \leq 80\%$, $70 \leq s \leq 85\%$, respectively), were penalized to a larger degree as they could be biologically implausible, due to the large amount of sequences in the OAS. Sequences that were extremely dissimilar ($\leq 70\%$) were disqualified for the same reason. Due to the variability of the CDR regions, particularly CDR-H3, we didn't penalize generated CDR-H3 with moderate similarity. Finally, a similar non-penalty was applied for the frameworks, but due to lower variability in these regions, the lower bound was increased.

Due to compute limitations, all matching was done using OAS-aligned-small, found at the KA-Search Github. (Olsen et al., 2023)

$$S_{\mathrm{T1}}(s) = \begin{cases} 0 \text{ and disqualify}, & s < 0.70, s = 1 \\ \dfrac{s - 0.70}{0.10}, & 0.70 \leq s < 0.80, \\ 1, & 0.80 \leq s < 0.90, \\ 1 - (s - 0.9), & 0.90 \leq s < 1.00, \end{cases}$$

$$S_{\mathrm{T2/T3}}(s) = \begin{cases} 0 \text{ and disqualify}, & s < 0.70, s = 1 \\ \dfrac{s - 0.70}{0.15}, & 0.70 \leq s < 0.85, \\ 1, & 0.85 \leq s < 0.90, \\ 1 - (s - 0.9), & 0.90 \leq s < 1.00, \end{cases}$$

### A.2. Biological Realism Scoring Formula

#### A.2.1. TAP Score

As noted in (Raybould et al., 2019) and (Raybould et al., 2024), amber flags represent an extreme value for the category, while a red flag is previously unobserved. Our scoring method for TAP is a heuristic designed to tolerate mild but not extreme risk (e.g. 2 amber flags or 1 red flag).

#### A.2.2. Immunogenicity Score

NetMHCPan (Nilsson et al., 2023) reports strong and weak binders using a percentile-rank score. Following the conventions listed in (Reynisson et al., 2020), we've set the

percentile-rank cutoffs for strong and weak binders to $2\%$ and $10\%$, respectively. Note that this was done using the NetMHCIIpan 4.3 web server with the following list of alleles: DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0701, DRB1*0801, DRB1*1101, DRB1*1301, DRB1*1501. The alleles used were motivated by similar use in (Thrift et al., 2024).

Then, $f_{strong}$ is calculated as the number of peptide windows classified as strong binders divided by the total number of evaluated peptide windows. As $f_{strong}$ increases, the immunogenicity score decreases, and with $\geq 10\%$ of binders being strong binders, we reach an immunogenicity score of 0.

### A.3. Structural Score

For the structural score, we wanted to balance between pLDDT and PAE, representing local structure and positioning. As such, we take the average of the two scores.

Our pLDDT score was based on Alphafold documentation (Google DeepMind and EMBL-EBI, 2026), where it is noted that pLDDT scores above 70 are expected to be modeled well, and above 90 are expected to be modeled to high accuracy. For PAE, below 5 angstroms indicates excellent confidence, and between 5 and 15 indicates moderate confidence locally, and lower confidence in long-range and inter domain positioning. As such, we made the cutoffs for pLDDT and PAE to be 80 and 10, respectively. By doing so, we filtered for antibody candidates with strong local structure and near excellent confidence.

One note here is that even if the cutoff for PAE had been 15 instead of 10, the same models would have failed Task 1. For Task 3, IGLM wouldn't have completely failed, but only 11 of 90 generated antibodies would have been fully scored. Even assuming a perfect score on everywhere else, IGLM would've received a composite score below 12% for Task 3.

### A.4. Binding Score

$S_{bind} = \frac{(\Delta G - (-8))}{(-15)-(-8)}$

$\Delta G$ is calculated as the mean PRODIGY-predicted binding free energy (kcal/mol) from the top five HADDOCK poses.

A binding energy higher than -8 indicates weak predicted binding, while a binding score lower than -15 indicates very strong predicted binding. As such, our formula was made so binding energies of -8 or greater are scored as 0, binding energies of -15 or less are scored as 1, and energies from -8 to -15 are increasing in score.

### A.5. Antigens

The antigens used for this paper were P24, HER2, and Amyloid-Beta.

### A.6. BindCraft Results

BindCraft optimises binder backbone geometry via AlphaFold2-Multimer backpropagation across four sequential stages, after which successful trajectories undergo ProteinMPNN sequence redesign and AF2 monomer complex reprediction against a predefined filter set. Of eight trajectories targeting the HIV-1 p24 capsid protein (PDB: 1E6J), four completed all optimisation stages; of twenty-four trajectories targeting the amyloid-$\beta$ peptide (PDB: 6CO3), two passed the trajectory quality filter and entered ProteinMPNN sequence redesign; of fifty-seven trajectories targeting the HER2 receptor ectodomain (PDB: 1N8Z), trimmed to a 75-residue epitope-proximal crop, thirty-three completed all optimisation stages and were scored, while fourteen and ten were classified as clashing and low-confidence, respectively. None of the resulting sequences passed downstream AF2 validation filters, yielding zero accepted designs for all three targets.

### A.7. Weighting Rationale

Default weights were chosen under the reasonable assumption that Biological Realism is by far the most important criteria. Ultimately, no matter how good predicted structural, binding, and novelty scores are, if the antibody isn't realistic, there was no point in scoring it. This is particularly important when it comes to TAP usage, which actually informs us when antibody candidates are extremely far from the norm or previously unseen.

### A.8. Weighting Ablation

As IgLM, BindCraft, and ProtT5 failed upstream scoring thresholds and did not receive composite scores, sensitivity analysis is only performed on fully scored RFDiffusion candidates.

**Default weights:** $S_{\text{composite}} = 0.20S_{\text{nov}} + 0.40S_{\text{bio}} + 0.20S_{\text{bind}} + 0.20S_{\text{struct}}$.

**Equal weights:** $S_{\text{composite}} = 0.25S_{\text{nov}} + 0.25S_{\text{bio}} + 0.25S_{\text{bind}} + 0.25S_{\text{struct}}$.

**Binding-heavy:** $S_{\text{composite}} = 0.15S_{\text{nov}} + 0.25S_{\text{bio}} + 0.40S_{\text{bind}} + 0.20S_{\text{struct}}$.

**Structure-heavy:** $S_{\text{composite}} = 0.15S_{\text{nov}} + 0.25S_{\text{bio}} + 0.20S_{\text{bind}} + 0.40S_{\text{struct}}$.

**Novelty-heavy:** $S_{\text{composite}} = 0.40S_{\text{nov}} + 0.25S_{\text{bio}} + 0.15S_{\text{bind}} + 0.20S_{\text{struct}}$.

*Table 2.* Weight-sensitivity ablation for RFDiffusion-generated candidates.

| Weight setting | HER2 | p24 | A$\beta$ | Ranking |
|---|---|---|---|---|
| Default | 65.2% | 82.9% | 78.5% | p24 $>$ A$\beta$ $>$ HER2 |
| Equal weights | 63.4% | 80.1% | 76.7% | p24 $>$ A$\beta$ $>$ HER2 |
| Binding-heavy | 66% | 82.1% | 79.1% | p24 $>$ A$\beta$ $>$ HER2 |
| Structure-heavy | 56.7% | 73.5% | 68.3% | p24 $>$ A$\beta$ $>$ HER2 |
| Novelty-heavy | 65.4% | 82.4% | 79.8% | p24 $>$ A$\beta$ $>$ HER2 |

## A.9. Model Usage

*Table 3.* Overview of baseline models used in the benchmark.

| Model | Architecture | Tasks | Workflow |
|---|---|---|---|
| IgLM | Decoder-only protein language model | Task 1, Task 3 | **Unconditional generation / infilling.** IgLM uses antibody-specific sequence grammar to generate antibody candidates. For Task 1, its masking and infilling capability is used to redesign the CDR-H3 loop within a fixed framework. |
| ProtT5-FT | Encoder–decoder protein language model | Task 1, Task 2 | **Conditional infilling.** ProtT5 is fine-tuned on an OAS subset for sequence-to-sequence translation. CDR-H3 infilling in Task 1 and framework infilling in Task 2 are performed by masking the target region and conditioning the encoder on the antigen sequence. |
| RFDiffusion | Diffusion model | Task 1 | **Antigen-conditioned generation.** The RFantibody pipeline is used to generate CDR backbones and antibody poses relative to the fixed antigen structure. ProteinMPNN inverse folding is then applied to convert generated backbones into amino acid sequences. |
| BindCraft | AF2-Multimer optimization | Task 1, Task 3 | **Hallucination pipeline.** BindCraft performs gradient-based optimization over binder sequence and structure to maximize AF2-Multimer confidence metrics such as ipTM and ipLDDT. ProteinMPNN inverse folding is used after backbone optimization. |

*Table 4.* Sampling hyperparameters for sequence-first language models.

| Parameter | IgLM | ProtT5-FT | Rationale |
|---|---|---|---|
| Temperature | 1.0 | 1.0 | A neutral value balancing novelty and coherence. |
| Top-$p$ | 1.0 | 0.85 | For IgLm, allows greater diversity in sequences. For ProtT5, restricts sampling to highly probable residues for sequence realism. |
| Number of samples | 30 | 30 | Required number of candidates per antigen and task. |
| Beam size | N/A | 1 | Beam search is used in ProtT5's decoder for controlled, high-quality conditional generation. |

*Table 5.* Sampling and optimization hyperparameters for structure-first models.

| Parameter | RFDiffusion | BindCraft | Rationale |
|---|---|---|---|
| Denoising steps | 150 | N/A | Fixed number of steps for high-quality, reproducible backbone generation. |
| Guidance scale | 1.0–2.0 | N/A | Controls adherence to antigen and epitope constraints. |
| Optimization goal | N/A | Maximize ipTM and ipLDDT | Drives the sequence and structure toward a highly confident binding interface. |
| Inverse folding | ProteinMPNN required | ProteinMPNN required | All structure-first designs use ProteinMPNN to translate final backbone coordinates into amino acid sequences. |

## References


M. Chungyoun, J. Ruffolo, and J. Gray. Flab: Benchmarking deep learning methods for antibody fitness prediction. bioRxiv preprint, 2024.

J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, and C. M. Deane. Sabdab: The structural antibody database. *Nucleic Acids Research*, 42 (D1):D1140–D1146, 2014. doi: 10.1093/nar/gkt1043.

Google DeepMind and EMBL-EBI. AlphaFold Protein Structure Database. https://alphafold.ebi.ac.uk/, 2026. Accessed: 2026-04-27.

S. Joubbi, A. Micheli, P. Milazzo, G. Ciano, S. M. Gagné, P. Liò, D. Medini, and G. Maccari. Enhancing antibody-antigen interaction prediction with atomic flexibility. *PLOS Computational Biology*, 21(10):e1013576, 2025. doi: 10.1371/journal.pcbi.1013576.

N. Kim, M. Kim, S. Ahn, and J. Park. Decoupled sequence and structure generation for realistic antibody design, 2024.

Z. Kurkcuoglu, P. I. Koukos, N. Citro, M. E. Trellet, J. P. G. L. M. Rodrigues, I. S. Moreira, J. Roel-Touris, A. S. J. Melquiond, C. Geng, J. Schaarschmidt, L. C. Xue, A. Vangone, and A. M. J. J. Bonvin. Performance of HADDOCK and a simple contact-based protein-ligand binding affinity predictor in the D3R grand challenge 2. *Journal of Computer-Aided Molecular Design*, 32(1), 2018. doi: 10.1007/s10822-017-0049-y.

Y. Li, Y. Lang, C. Xu, Y. Zhou, Z. Pang, and P. J. Greisen. Benchmarking inverse folding models for antibody CDR sequence design. *PLoS One*, 20(6):e0324566, 2025. doi: 10.1371/journal.pone.0324566.

C. Liu, A. Pelissier, Y. Shao, L. Denzler, A. Martin, B. Paige, and M. Martínez. Abrank: A benchmarking dataset and metric-learning framework for antibody–antigen affinity ranking, 2025.

F. Meng, N. Zhou, G. Hu, R. Liu, Y. Zhang, M. Jing, and Q. Hou. A comprehensive overview of recent advances in generative models for antibodies. *Computational and Structural Biotechnology Journal*, 23, 2024. doi: 10.1016/j.csbj.2024.06.016.

J. B. Nilsson, S. Kaabinejadian, H. Yari, M. G. D. Kester, P. van Balen, W. H. Hildebrand, and M. Nielsen. Accurate prediction of hla class ii antigen presentation across all loci using tailored data acquisition and refined machine learning. *Science Advances*, 9(47):eadj6367, 2023. doi: 10.1126/sciadv.adj6367.

T. H. Olsen, F. Boyles, and C. M. Deane. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022. doi: 10.1002/pro.4205.

T. H. Olsen, B. Abanades, I. H. Moal, and C. M. Deane. Ka-search, a method for rapid and exhaustive sequence identity search of known antibodies. *Scientific Reports*, 13(11612), 2023. doi: 10.1038/s41598-023-38108-7.

M. I. J. Raybould, C. Marks, K. Krawczyk, B. Taddese, J. Nowak, A. P. Lewis, A. Bujotzek, J. Shi, and C. M. Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, 116(10):4025–4030, 2019. doi: 10.1073/pnas.1810576116.

M. I. J. Raybould, O. M. Turnbull, A. Suter, B. Guloglu, and C. M. Deane. Contextualising the developability risk of antibodies with lambda light chains using enhanced therapeutic antibody profiling. *Communications Biology*, 7(62), 2024. doi: 10.1038/s42003-023-05744-8.

B. Reynisson, B. Alvarez, S. Paul, B. Peters, and M. Nielsen. Netmhcpan-4.1 and netmhciipan-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic Acids Research*, 48(W1):W449–W454, 2020. doi: 10.1093/nar/gkaa379.

C. Schneider, M. I. J. Raybould, and C. M. Deane. Sabdab in the age of biotherapeutics: Updates including sabdab-nano, the nanobody structure tracker. *Nucleic Acids Research*, 50(D1):D1368–D1372, 2022. doi: 10.1093/nar/gkab1050.

W. J. Thrift, J. Perera, S. Cohen, N. W. Lounsbury, H. R. Gurung, C. M. Rose, J. Chen, S. Jhunjhunwala, and K. Liu. Graph-pmhc: graph neural network approach to mhc class ii peptide presentation and antibody immunogenicity. *Briefings in Bioinformatics*, 25(3):bbae123, 2024. doi: 10.1093/bib/bbae123.

T. Uçar, C. Malherbe, and F. Gonzalez. Benchmarking generative models for antibody design exploring log-likelihood for sequence ranking. bioRxiv preprint, 2024.

A. Vangone and A. M. J. J. Bonvin. Prodigy: A contact-based predictor of binding affinity in protein-protein complexes. *Bio-protocol*, 7(3), 2017. doi: 10.21769/BioProtoc.2124.

M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi,

M. Figurnov, and S. Velankar. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1), 2022. doi: 10.1093/nar/gkab1061.

L. C. Xue, J. P. G. L. M. Rodrigues, P. L. Kastritis, A. M. J. J. Bonvin, and A. Vangone. Prodigy: A web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics*, 32(23):3676–3678, 2016. doi: 10.1093/bioinformatics/btw514.

S. Zhao, J. Moller, P. Quintero-Cadena, and L. van Niekerk. Guided generation for developable antibodies. In *Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025a.

X. Zhao, Y. Tang, A. Singh, V. Cantu, K. An, J. Lee, A. Stogsdill, I. Hamdi, A. Ramesh, Z. An, X. Jiang, and Y. Kim. Abbibench: A benchmark for antibody binding affinity maturation and design, 2025b.