ABGATE: LENGTH-DECOUPLED ANTIGEN-CONDITIONED ANTIBODY DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Antibodies are important therapeutics, but designing them for a chosen antigen from sequence alone remains challenging. The input specifies neither the binding site nor the geometry of the antibody–antigen interface. We propose AbGate, which conditions a pretrained antibody language model on the antigen sequence alone. A learned pooler compresses the antigens of different lengths into 32 tokens, and cross-attention supplies these tokens to every decoder layer without occupying the decoder's autoregressive input. Using approximately 1.1% of MAGE's trainable parameters, AbGate performs on par with MAGE, a fully finetuned generator trained on the same corpus, in mean RosettaFold3 ipTM and pLDDT on a shared target. Across 502 held-out targets AbGate generates antibodies with 95.6% filter retention and 100% novelty against the training pairs. In a matched evaluation across 496 targets, antibodies generated by AbGate perform comparably to experimentally validated binding antibodies on binding interface confidence. These results support compact antigen conditioning as an efficient approach to adapting antibody language models for candidate generation.

1 INTRODUCTION

Computational protein design has advanced quickly through generative models of protein sequences and structures (Watson et al., 2023; Ingraham et al., 2023; Dauparas et al., 2022). These advances have extended to antibody design, where methods use structural information to redesign antibody loops or generate binders against a chosen epitope (Luo et al., 2022; Swanson et al., 2025; Levine et al., 2026; Bennett et al., 2026; Kim & Baek, 2026). Sequence models offer a complementary approach that can learn from large antibody repertoires and from antigen–antibody sequence pairs without resolved complex structures (Olsen et al., 2022; Wasdin et al., 2025). By generating sequences separately from complex structures, they also allow structural evaluation to focus on selected candidates. We focus on this setting, in which the input is an antigen sequence and the output is a pair of variable antibody regions from the heavy and light chains (VH/VL).

Pretrained antibody language models provide a starting point for this task by learning sequence patterns from antibody repertoires (Nijkamp et al., 2023). To generate antibodies for a particular antigen, the model must also receive information about that antigen. MAGE (Wasdin et al., 2025), for example, supplies the antigen sequence as a prefix to a general protein language model and finetunes the full generator. Antigen lengths vary across targets, so this representation gives the decoder a different number of target tokens for each antigen. Longer prefixes occupy more of the decoder's context and increase its attention cost. This motivates a common conditioning interface that represents antigens of different lengths with a fixed number of tokens and adapts the model while keeping its pretrained Transformer blocks frozen.

We introduce AbGate, an antibody generation model that uses this compact interface to condition a pretrained decoder on an antigen sequence. A learned attention pooler compresses residue embeddings from a frozen ESM-C encoder (Candido et al., 2026) into 32 antigen tokens. The ProGen2-OAS decoder (Nijkamp et al., 2023) reads these tokens through cross-attention at every layer as it generates the paired antibody sequences. The number of conditioning tokens is therefore the same for every antigen, and these tokens occupy no positions in the decoder's autoregressive input.

With this compact conditioning interface, AbGate generates paired antibody sequences for 502 antigens held out by a cluster split at 30% sequence identity. Across these targets, 95.6% of generated

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.

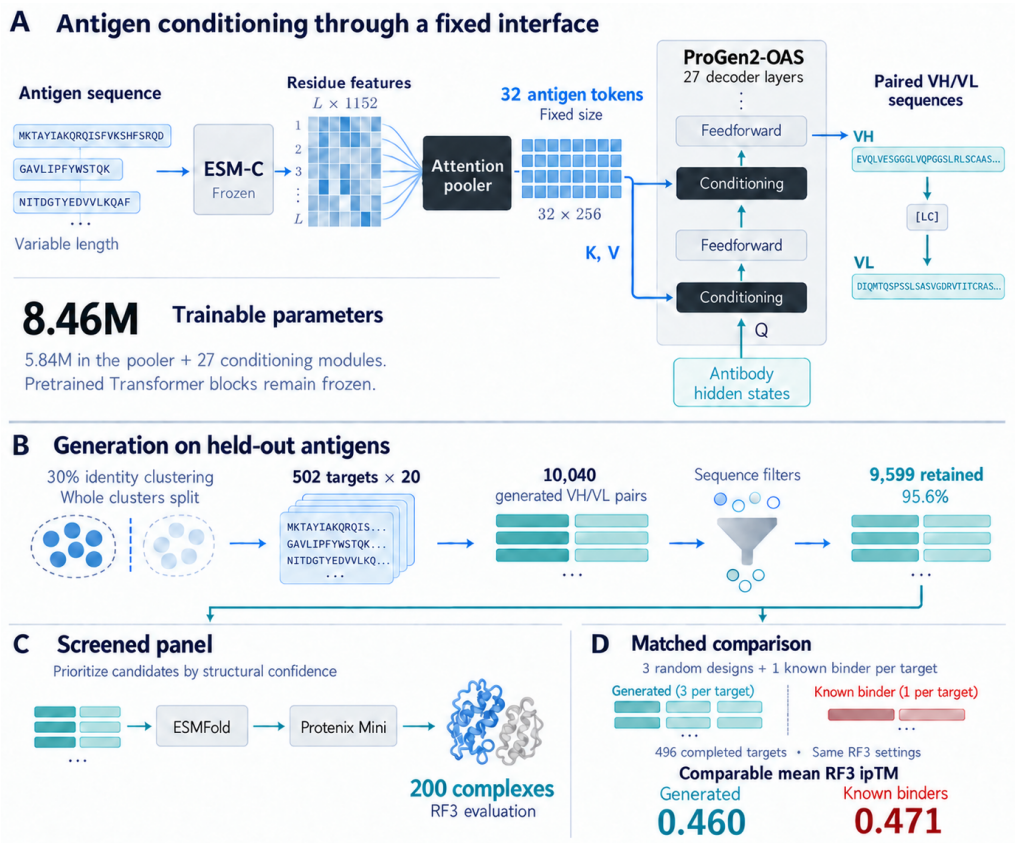


Figure 1: AbGate architecture and evaluation workflow. (A) A frozen ESM-C encoder and a learned attention pooler map each antigen sequence to 32 tokens of dimension 256. Conditioning modules read these tokens through cross-attention and update antibody hidden states before each decoder feedforward network. Other decoder operations are omitted from the schematic. (B) Sampling 20 pairs for each of 502 held-out targets yields 10,040 VH/VL pairs, of which 9,599 pass sequence filters. (C) ESMFold and Protenix Mini screening yields 200 complexes for RF3 evaluation. (D) For each target, a separate matched panel compares three designs randomly sampled from the filtered pool with one known binder under the same RF3 settings. Across 496 completed targets, generated antibodies and known binders achieve comparable mean ipTM (0.460 and 0.471, respectively). Sequence snippets and ribbon drawings are schematic.

candidates pass our antibody sequence filters. None of the candidates passing the numbering and length checks duplicates a training VH/VL pair. In a matched panel of 496 targets, generated antibodies also achieve mean RosettaFold3 (RF3) (Corley et al., 2025) ipTM comparable to known binders across antigen lengths. In a separate comparison on a shared SARS-CoV-2 receptor binding domain (RBD) target, mean ipTM is 0.434 for AbGate and 0.435 for the fully finetuned MAGE (Wasdin et al., 2025) model under common settings for structure prediction. AbGate achieves this confidence with approximately 1.1% of MAGE's trainable parameters (Section 3.3).

2 Method

To condition an antibody language model on antigens of different lengths, we separate antigen encoding from antibody generation. AbGate uses a learned pooler to summarize antigen residue features into 32 tokens, which cross-attention modules make available at every decoder layer. This gives the decoder the same number of conditioning tokens for every target without extending its autoregressive input as antigen length increases. The decoder generates the heavy chain, a `[LC]` separator token, then the light chain in sequence. Figure 1 shows how this architecture fits into the generation and evaluation workflow.

2.1 Antigen Encoding and Pooling

We first encode each antigen with ESM-C (`esmc_600m`) (Candido et al., 2026), obtaining a feature vector of dimension 1152 for each residue. These features are cached and reused during training and generation. A learned attention pooler then converts the residue features into $K = 32$ tokens of dimension $d = 256$. Following attention pooling with learned queries (Jaegle et al., 2021; Li et al., 2023), each output token is a weighted summary of the antigen residues.

Let $\tilde{E} \in \mathbb{R}^{L\times 1155}$ contain the residue features augmented with three positional features, p, $\sin(\pi p)$, and $\cos(\pi p)$, where p is a normalized position index. With learned queries $Q \in \mathbb{R}^{K\times d}$, the pooler computes

$$T = \mathrm{LayerNorm}\left[\mathrm{softmax}\left(\frac{Q\phi_K(\tilde{E})^\top}{\sqrt{d}} + M\right)\phi_V(\tilde{E})\right] \in \mathbb{R}^{K\times d}, \tag{1}$$

where ϕ_K and ϕ_V are learned affine projections into d dimensions and M masks padded residues. Layer normalization is applied to each pooled token. Appendix A specifies the positional features and padding convention.

2.2 Conditioning the Antibody Decoder

Each decoder layer accesses this representation through a conditioning module inserted before its feedforward network. Within the module, antibody hidden states query the pooled antigen tokens through cross-attention, allowing the attention weights to depend on the sequence generated so far and on the current position. We add these modules to all 27 layers of ProGen2-OAS (Nijkamp et al., 2023), an antibody language model pretrained on the Observed Antibody Space (Olsen et al., 2022). This follows prior work on adapting pretrained networks through small trainable modules (Houlsby et al., 2019; Alayrac et al., 2022).

For hidden states $H \in \mathbb{R}^{n\times 1536}$, the module computes an antigen context A and a residual update:

$$\begin{aligned} A &= \mathrm{softmax}\left(\frac{(HW_Q)(TW'_K)^\top}{\sqrt{64}}\right)TW'_V \in \mathbb{R}^{n\times 16}, \\ H' &= H + \mathrm{GELU}([\,HW_{\mathrm{down}}\,;\,A\,])\,W_{\mathrm{up}} + b_{\mathrm{up}}. \end{aligned} \tag{2}$$

The query and key projections have dimension 64. Both the antigen context and the projected antibody state have dimension 16, and $[\,\cdot\,;\,\cdot\,]$ denotes concatenation. The output projection maps their combined representation back to the decoder width. We initialize W_{up} and b_{up} to zero, so each conditioning module initially leaves its input unchanged and learns a residual update during training.

The pooler and the 27 conditioning modules contain 5.84 million parameters, included in the implementation's total of 8.46 million parameters marked trainable. The pretrained Transformer blocks remain frozen, while the pooler, conditioning modules, and token parameters used for chain separation are adapted. Appendix A gives the parameter setting and implementation details.

2.3 Training and Candidate Evaluation

We train this conditioning interface on paired antigen and antibody sequences using teacher forcing. The objective is cross-entropy over the antibody sequence, including the chain separator and end token. Annotated CDR positions receive three times the weight of other sequence positions. Padding is excluded from the loss. After training, we compute the antigen representation once and reuse it to sample multiple paired antibody sequences. Section 3.1 gives the data split, sampling settings, and sequence filters. Appendix A.3 provides the training settings.

We evaluate candidates retained after sequence filtering in two settings (Figure 1C–D). The screened panel uses antibody fold confidence and a preliminary complex prediction to prioritize candidates for RF3 evaluation. The matched panel instead samples candidates at random and evaluates them alongside known binders against the same targets. Section 3 reports the construction and results of both panels.

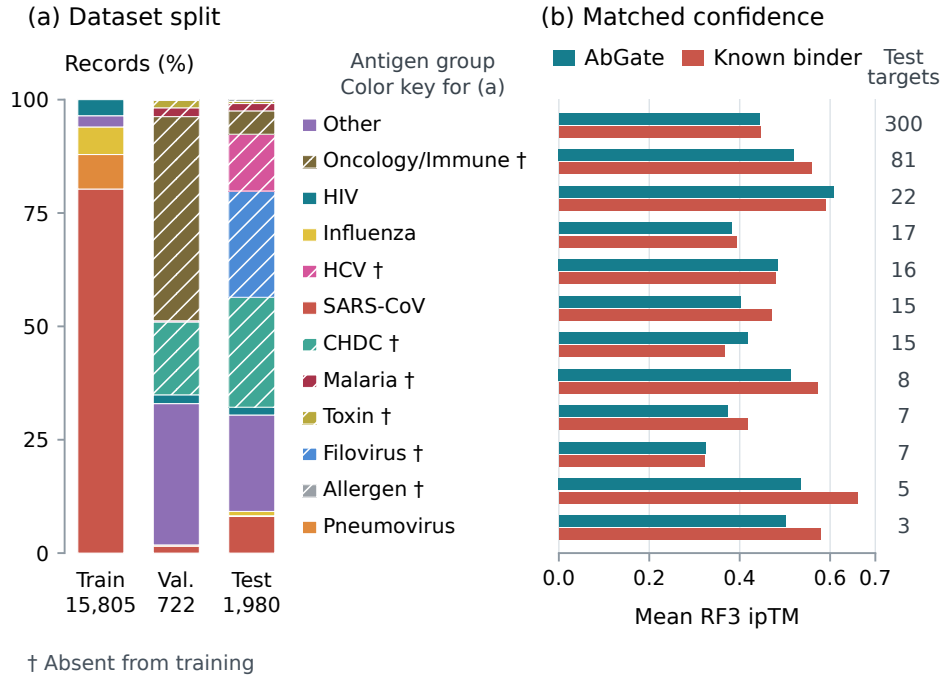


Figure 2: Antigen groups in the dataset and matched evaluation. (a) Group composition of antibody–antigen records after the 30% antigen-homology split. Counts below the bars give records in each split. (b) Mean RF3 ipTM for AbGate designs and known binders. Each target contributes the mean of three designs and the score of one known binder. Numbers on the right give target counts. Hatching and † mark groups absent from training. Other denotes antigens without a group annotation.

3 Experiments

3.1 Experimental setup

We use the paired antibody–antigen corpus released with MAGE (Wasdin et al., 2025), containing 18,507 records and 1,336 distinct antigen sequences. To evaluate generation beyond the training antigens, antigen sequences were grouped with MMseqs2 (Steinegger & Söding, 2017) at 30% sequence identity. Whole clusters were then assigned to training, validation, and test sets. The split contains 15,805 training, 722 validation, and 1,980 test pairs, with no overlap in clusters. Here, held-out antigens are those whose clusters are excluded from training. The test set spans twelve antigen groups, seven of which are absent from training (Figure 2a).

We select the checkpoint with the lowest validation loss and generate VH/VL pairs using nucleus sampling with $p = 0.9$ and temperature 1.0. Before structural evaluation, AbGate candidates pass four sequence filters: successful antibody numbering of both chains (F1), chain lengths within specified ranges (F2), absence of an exact VH/VL pair in the training set (F3), and identifiable CDR3 sequences in both chains (F4). Appendix A.3 gives the filter thresholds and additional training settings.

Within each target, RF3 predicts all antigen–antibody complexes using the same antigen sequence, antigen alignment, and prediction settings. We report ipTM and overall pLDDT (scaled to 0–1), together with minimum chain-pair predicted aligned error (PAE, in Å) for the antigen–antibody and VH–VL interfaces. Higher ipTM and pLDDT indicate greater structural confidence, while lower PAE indicates greater confidence in the relative placement of the chains. Appendix A.3 specifies the paired and unpaired analyses, including bootstrap intervals and rank tests.

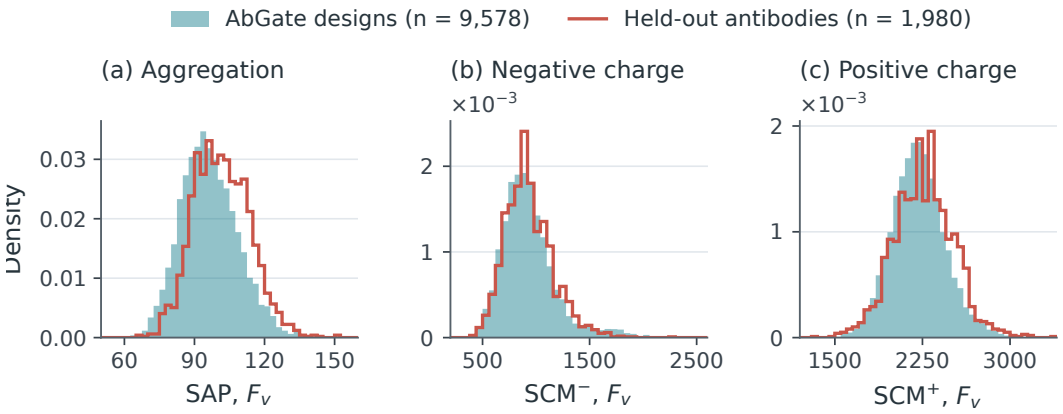


Figure 3: DeepSP descriptor distributions for the complete variable regions (F_v) of 9,578 generated and 1,980 held-out antibodies with valid DeepSP predictions. Panels show (a) spatial aggregation propensity (SAP), (b) negative and (c) positive spatial charge map (SCM) scores. Histograms use common bins within each panel and are normalized separately to unit area. Appendix B reports all seven descriptors.

3.2 Generation for held-out antigens

Of the 504 test antigens, 502 have cached ESM-C embeddings. We generate 20 VH/VL pairs per antigen, yielding 10,040 candidates. Of these, 9,599 pass the sequence filters, a 95.6% pass rate, covering all 502 targets with a median of 20 candidates per target. None of the candidates passing the numbering and length checks exactly matches a training VH/VL pair. Appendix C reports counts at each evaluation stage.

The retained candidates have sequence properties similar to those of the held-out antibodies. Mean CDR3 lengths are 16.11 versus 16.19 residues for the heavy chain and 10.05 versus 9.63 for the light chain (Table 3). DeepSP (Kalejaye et al., 2024) predicts antibody aggregation propensity and surface charge profiles. These profiles resemble those of the held-out antibodies (Figure 3). Lower SAP indicates lower predicted aggregation propensity, while SCM describes negative and positive surface charge patches. Interquartile ranges overlap across all seven descriptors. Appendix B provides summaries of all seven descriptors and sequence similarity analyses.

We first assess the antibody structures with ESMFold (Lin et al., 2023), a model that predicts protein structures from sequence. Predictions were obtained for the VH and VL chains of all 9,599 candidates, with a mean pLDDT of 78.47 on a 0–100 scale. From the top 40% of the filtered pool by mean VH/VL pLDDT, we retain up to five candidates per antigen, yielding 2,478 candidates. Protenix Mini (Gong et al., 2025) then predicts their antigen–antibody complexes, completing 2,470 predictions. We select 200 complexes across 110 targets for RF3 evaluation using the Protenix ranking score, antigen group quotas, and a limit of three designs per target. This screened panel achieves a mean ipTM of 0.521, with 61 complexes scoring at least 0.6 and 37 scoring at least 0.7. The highest ipTM is 0.841 (Figure 4b). Appendix C gives the detailed screening settings.

To compare generation with known binders across test targets, we form a separate matched panel. For each target, we draw three designs at random from the sequence-filtered pool and compare their mean scores with those of one known binder. Across the 496 targets with complete RF3 predictions (1,984 complexes), AbGate achieves mean ipTM and overall pLDDT comparable to known binders of the same targets. Mean ipTM is 0.460 for AbGate and 0.471 for known binders, while mean overall pLDDT is 0.787 and 0.788, respectively (Figure 4c).

AbGate achieves comparable RF3 confidence in several antigen groups absent from training (Figure 2b). In Oncology/Immune, the largest such group with 81 targets, mean ipTM is 0.520 for AbGate and 0.559 for known binders. The means are nearly identical for HCV (0.485 versus 0.479) and Filovirus (0.324 versus 0.322), while AbGate scores higher for CHDC (0.416 versus 0.367). AbGate and known binders also show similar mean ipTM across all five antigen length groups (Figure 5). Scores decrease with antigen length in both groups.

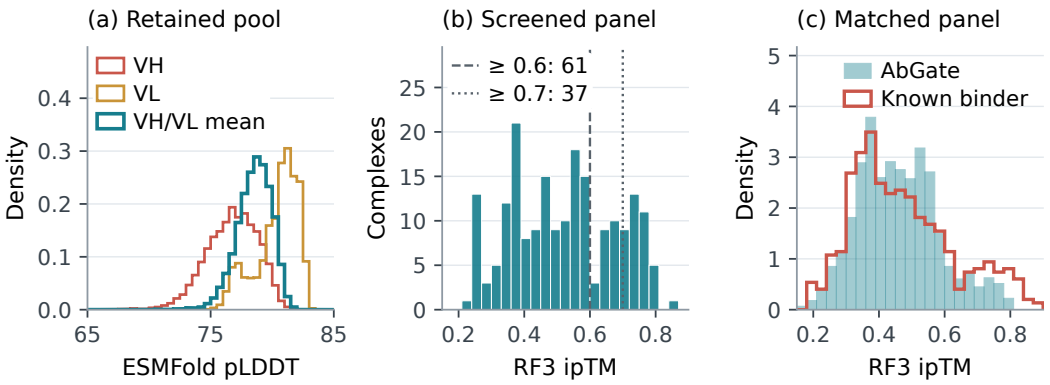
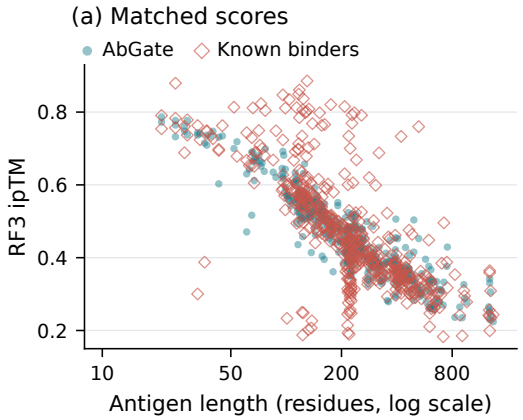


Figure 4: Antibody and complex structure confidence. (a) ESMFold pLDDT for VH, VL, and their mean across 9,599 retained candidates. (b) RF3 ipTM for 200 complexes selected by structural screening. (c) RF3 ipTM for three randomly sampled AbGate designs and one known binder per target across 496 targets with complete predictions. Histograms in (c) show individual complex scores.



(b) Mean ipTM by length

Length (residues)	Targets	AbGate	Known binders
≤100	50	0.682	0.701
101–200	155	0.527	0.557
201–400	192	0.416	0.415
401–800	84	0.340	0.340
>800	15	0.271	0.274

Figure 5: RF3 ipTM across antigen lengths in the matched panel. (a) Each antigen contributes the mean of three AbGate designs and the score of one known binder, giving 496 points per group. The horizontal axis shows antigen length on a logarithmic scale. (b) Mean ipTM within five antigen length groups, with target counts.

The interface PAE scores distinguish antibody chain assembly from antigen placement (Figure 6). AbGate has lower mean minimum VH–VL PAE than known binders, at 3.02 versus 3.85 Å, indicating confident relative placement of its antibody chains. Known binders have lower mean minimum antigen–antibody PAE, at 13.71 Å compared with 15.42 Å for AbGate. The generated antibodies therefore have comparable ipTM and overall pLDDT and confident VH–VL assembly, with greater uncertainty in their placement relative to the antigen. Appendix D reports the paired statistics and additional antigen length analyses.

3.3 Comparison with MAGE

We compare AbGate with MAGE on the SARS-CoV-2 RBD from MAGE's released generation run, a target represented in both training datasets. From MAGE's 991 released designs, we select the 180 with the highest ESMFold pLDDT. For AbGate, we generate 20 additional VH/VL pairs and evaluate all 20 without selection by structural confidence.

On this shared target, AbGate matches the fully finetuned MAGE model (Wasdin et al., 2025) in mean RF3 ipTM and overall pLDDT with approximately 1.1% of MAGE's trainable parameters (Table 1). Mean ipTM is 0.434 for AbGate and 0.435 for MAGE, while mean pLDDT is 0.787 and 0.789, respectively. Mean minimum antigen–antibody PAE is higher for AbGate, at 18.98 Å

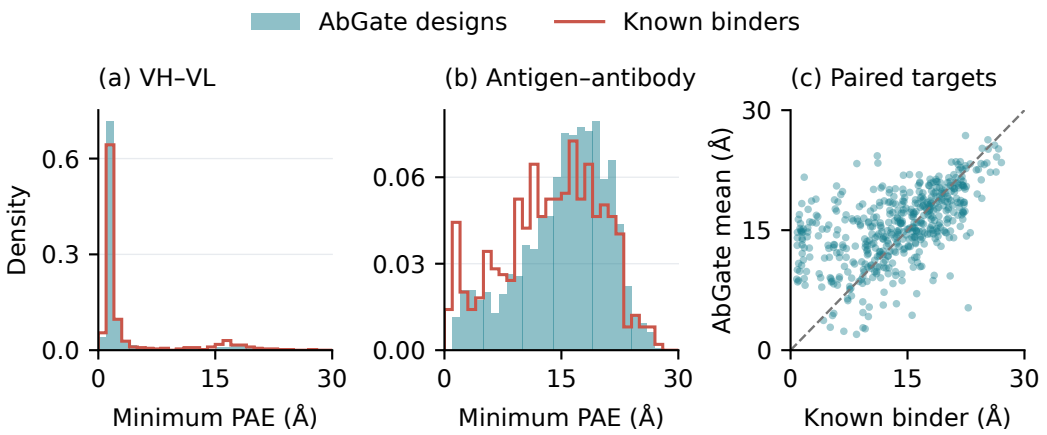


Figure 6: RF3 minimum chain-pair PAE in the matched panel. (a) VH–VL PAE and (b) antigen–antibody PAE for generated designs and known binders. Lower values indicate greater confidence. (c) Antigen–antibody PAE compared within each of the 496 targets. Each point compares the mean of three AbGate designs with one known binder. The dashed line denotes equal values.

compared with 17.47 Å for MAGE. The parameter comparison uses AbGate's full implementation tally of 8.46 million against MAGE's 764 million (Appendix A).

Table 1: Trainable parameters and mean RF3 scores on the shared RBD target. MAGE includes the 180 designs with the highest ESMFold pLDDT among its 991 released designs. AbGate includes all 20 additional samples without selection by structural confidence. PAE is in Å. Parameter accounting is detailed in Appendix A.

	AbGate ($n = 20$)	MAGE ($n = 180$)
Parameters marked trainable (M)	8.46	764
ipTM	0.434	0.435
Overall pLDDT	0.787	0.789
Minimum antigen–antibody PAE	18.98	17.47

We also compare the generated chains with the training antibodies for this RBD. Mean VH sequence similarity to the nearest chain among the 3,000 training pairs is 0.919 for AbGate and 0.935 for MAGE (Figure 7a). Exact light-chain matches occur in 1 of 20 AbGate designs (5%) and 51 of 180 MAGE designs (28%). AbGate thus generates antibody chains that are more distinct from these training sequences while maintaining similar mean ipTM and pLDDT.

In the larger sequence pools, the generators have similar mean OASis percentiles (Prihoda et al., 2022), reported on a 0–1 scale where higher values indicate greater humanness. The mean is 0.881 for AbGate's 9,599 filtered designs across 502 test targets and 0.878 for MAGE's 991 released RBD designs (Figure 7b). Appendix E reports the complete structural comparison and sequence analysis methods.

Together, these results show that AbGate generates antibody candidates across held-out antigens with sequence properties similar to those of known antibodies. It achieves mean RF3 ipTM and pLDDT comparable to known binders on the matched test panel and to MAGE on the shared RBD target. The MAGE comparison shows that this level of structural confidence can be reached with approximately 1.1% of MAGE's trainable parameters.

4 Discussion

AbGate provides a compact way to add antigen conditioning while keeping the pretrained decoder blocks frozen. Pooling keeps the number of conditioning tokens fixed across antigen lengths, and

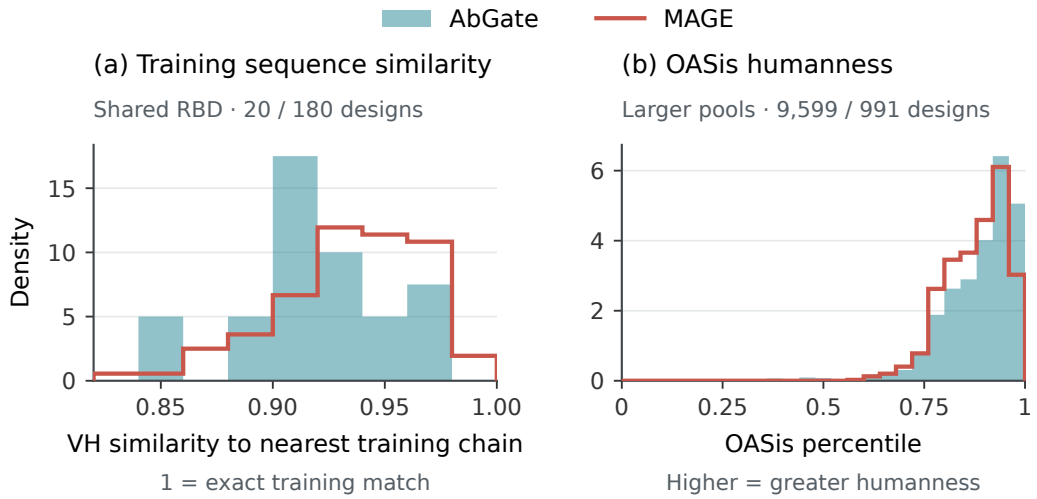


Figure 7: Sequence similarity and humanness distributions. (a) VH similarity to the nearest chain among the 3,000 training pairs for the shared RBD, for 20 AbGate and 180 MAGE designs. A similarity of 1 denotes an exact match to a training VH. (b) OASis percentiles for 9,599 AbGate designs across 502 test antigens and 991 MAGE RBD designs. Histograms use common bins within each panel and are normalized separately to unit area.

cross-attention supplies target information throughout generation. On the shared RBD target, AbGate is on par with MAGE in mean ipTM and pLDDT while using far fewer trainable parameters (Table 1).

The held-out evaluations support using this interface for candidate generation across diverse targets. Generated antibodies resemble known antibodies in sequence properties and achieve similar mean ipTM across antigen lengths and in several groups absent from training (Figures 2b and 5).

PAE shows modest differences between AbGate and known binders in the matched panel (Figure 6). AbGate has lower mean minimum VH–VL PAE but higher antigen–antibody PAE, indicating greater confidence in antibody chain assembly and greater uncertainty in placement relative to the antigen.

Binding affinity and specificity require experimental validation. The MAGE comparison covers one shared RBD target. A comparison across held-out antigens requires a common antigen cluster split and generation protocol. Our evaluation covers paired VH/VL antibodies, and the sequence split does not establish generalization to new interface geometries. RF3 also exceeded GPU memory for six long targets, leaving their structural evaluation incomplete (Appendix D).

During development, we also trained mean pooling and learned token pooling variants without decoder cross-attention. We compared their validation losses and responses to altered antigen inputs. These diagnostics did not isolate how pooling and cross-attention affect the quality of generated antigen–antibody complexes. In our related work, preliminary experiments with different token counts provided no clear basis for selecting K under a limited training budget. We did not complete a controlled comparison through AbGate's full generation and evaluation workflow because each variant required repeating the costly structural screening stages.

5 Conclusion

AbGate connects antigen sequences to a pretrained antibody decoder through 32 pooled tokens and cross-attention at every layer. It generates paired antibody candidates across 502 held-out antigens with a 95.6% sequence filter pass rate. Mean RF3 ipTM and pLDDT are comparable to known binders in the matched test panel and to MAGE on the shared RBD target. In the MAGE comparison, AbGate achieves these scores with approximately 1.1% of MAGE's trainable parameters. These results support compact antigen conditioning as an efficient approach to adapting antibody language models for candidate generation before structural evaluation and experimental testing.

ETHICS STATEMENT

This study is entirely computational and uses antibody–antigen sequence records from previously published resources. It involves no prospective collection of human-participant data, no interaction or intervention with human participants, and no animal experiments. The generated antibody sequences and predicted complexes are research hypotheses, not experimentally validated binders or therapeutics. They should not be used for clinical or biological deployment without independent experimental validation, biosafety assessment, and appropriate institutional and dual-use review.

REPRODUCIBILITY STATEMENT

Sections 2 and 3 describe the model, data split, candidate generation, and evaluation. Appendices A–G provide implementation and statistical settings, complete sequence and structural results, and details of the comparison methods. Source datasets, pretrained models, and prediction tools are identified by citation.

AI USE STATEMENT

In accordance with the conference policy on the use of AI, we disclose the following. Large language model assistants (Anthropic Claude, OpenAI ChatGPT) were used in this work to implement and debug the evaluation and analysis codes, and to edit and polish writing of the manuscript. OpenAI ChatGPT also assisted with the preparation of Figure 1. No synthetic data was generated by the assistants, and no reported result originates from them: every number in this paper comes from the authors' own training, generation, and folding runs. The authors reviewed all AI-assisted content, verified each quantitative claim against the underlying run artifacts, and take full responsibility for the contents of this paper.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. doi: 10.52202/068431-1723.

Nathaniel R. Bennett, Joseph L. Watson, Robert J. Ragotte, et al. Atomically accurate de novo design of antibodies with RFdiffusion. *Nature*, 649(8095):183–193, 2026. doi: 10.1038/s41586-025-09721-5.

Salvatore Candido, Thomas Hayes, Alexander Derry, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, June 2026. doi: 10.64898/2026.06.03.729735.

Nathaniel Corley, Simon Mathis, Rohith Krishna, et al. Accelerating biomolecular modeling with AtomWorks and RF3. *bioRxiv*, pp. 2025.08.14.670328, August 2025. doi: 10.1101/2025.08.14.670328.

Daniel Cutting, Frédéric A. Dreyer, David Errington, et al. De novo antibody design with SE(3) diffusion. *Journal of Computational Biology*, 32(4):351–361, 2025. doi: 10.1089/cmb.2024.0768.

J. Dauparas, I. Anishchenko, N. Bennett, et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

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

Chengyue Gong, Xinshi Chen, Yuxuan Zhang, et al. Protenix-Mini: Efficient structure predictor via compact architecture, few-step diffusion and switchable pLM. *arXiv preprint arXiv:2507.11839*, July 2025. doi: 10.48550/arXiv.2507.11839.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, et al. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019.

John B. Ingraham, Max Baranov, Zak Costello, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023. doi: 10.1038/s41586-023-06728-8.

Andrew Jaegle, Felix Gimeno, Andrew Brock, et al. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021. URL https://proceedings.mlr.press/v139/jaegle21a/jaegle21a.pdf.

Lateefat Kalejaye, I-En Wu, Taylor Terry, et al. DeepSP: Deep learning-based spatial properties to predict monoclonal antibody stability. *Computational and Structural Biotechnology Journal*, 23:2220–2229, 2024. doi: 10.1016/j.csbj.2024.05.029.

Yunji Kim and Minkyung Baek. De novo epitope-specific antibody design via time-dependent guidance. *Bioinformatics*, 42(Supplement_2):btag455, 2026. doi: 10.1093/bioinformatics/btag455.

Clayton W. Kosonocky, Alex M. Abel, Aaron L. Feller, et al. Validation and analysis of 12,000 AI-driven CAR-T designs in the Bits to Binders competition. *bioRxiv*, pp. 2026.03.03.709355, March 2026. doi: 10.64898/2026.03.03.709355.

Simon Levine, Jonathan Edward King, Jacob Stern, et al. Origin-1: a generative AI platform for de novo antibody design against novel epitopes. *bioRxiv*, 2026. doi: 10.64898/2026.01.14.699389.

Junnan Li, Dongxu Li, Silvio Savarese, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Zeming Lin, Halil Akin, Roshan Rao, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Shitong Luo, Yufeng Su, Xingang Peng, et al. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9754–9767, 2022. doi: 10.52202/068431-0709.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, et al. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, 2023. doi: 10.1016/j.cels.2023.10.002.

Tobias H. Olsen, Fergus Boyles, and Charlotte 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.

David Prihoda, Jad Maamary, Andrew Waight, et al. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs*, 14(1):2020203, 2022. doi: 10.1080/19420862.2021.2020203.

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

Erik Swanson, Michael Nichols, Supriya Ravichandran, et al. mBER: Controllable de novo antibody design with million-scale experimental screening. *bioRxiv*, September 2025. doi: 10.1101/2025.09.26.678877.

Perry T. Wasdin, Nicole V. Johnson, Alexis K. Janke, et al. Generation of antigen-specific paired-chain antibodies using large language models. *Cell*, 188(25):7206–7221.e16, December 2025. doi: 10.1016/j.cell.2025.10.006.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Jason Yang, Aadyot Bhatnagar, Jeffrey A. Ruffolo, et al. Function-guided conditional generation using protein language models with adapters. *arXiv preprint arXiv:2410.03634*, June 2025. doi: 10.48550/arXiv.2410.03634. Version 2.

A Implementation Details and Parameter Accounting

A.1 Antigen embeddings and sequence training

The training loader pads antigen residue embeddings to the longest antigen in each minibatch, with padded length $L_{\max}$. For position i, the pooler uses $p_i = (i-1)/(L_{\max}-1)$, $\sin(\pi p_i)$, and $\cos(\pi p_i)$, with $p_1 = 0$ when $L_{\max} = 1$. All antigens in the minibatch share this positional grid, and padded residues are masked before the attention softmax. During generation for a single antigen, $L_{\max}$ is its length. Thus training normalizes position by the padded minibatch length, rather than by each antigen's length.

Training sequences concatenate a start token, the heavy chain, `[LC]`, the light chain, and an end token. The loss is a weighted mean of next-token cross-entropies. Annotated CDR residues have weight 3. Other residues, the separator, and the end token have weight 1, while start and padding positions have weight 0. Sequences without a matched CDR annotation use weight 1 for all residues. Generation samples the heavy chain followed by the light chain, inserting `[LC]` at the transition. Maximum generated lengths are 150 residues for the heavy chain and 140 for the light chain.

A.2 Trainable parameters

AbGate has 8,462,336 parameters marked trainable in the implementation, approximately 1.1% of MAGE's 764 million trainable parameters (Table 2). The pooler and 27 conditioning modules account for 5,839,872 of this total. The full count also includes two unused modules for alternative conditioning routes and both complete token matrices. Frozen decoder parameters and the frozen 600M ESM-C encoder are excluded.

Table 2: AbGate parameter counts. The pooler and 27 conditioning modules form the active conditioning pathway. The implementation trainable tally is used for the comparison with MAGE.

Component	Parameters
Attention pooler	600,576
27 conditioning modules	5,239,296
Active conditioning total	**5,839,872**
Unused conditioning MLP	360,960
Unused pooled-token bridge	2,163,200
Full input and output token matrices	98,304
Implementation trainable tally	**8,462,336**

The pretrained Transformer blocks remain frozen. Both token matrices retain trainable status, with gradient hooks masking rows 0–29 and permitting gradients for rows 30–31, including the `[LC]` separator. These rows contain 6,144 entries across the input embedding and output matrices. Together with the active conditioning pathway, they account for 5,846,016 gradient-eligible parameters. AdamW applies weight decay of 0.01 to the full token matrices, including rows with masked gradients. The adapter checkpoint stores the conditioning modules and token rows 30–31, which are loaded into a fresh pretrained decoder for generation. It does not store the earlier token rows.

A.3 Experimental settings

Training uses AdamW with learning rate 3×10^{-5}, weight decay 0.01, batch size 8, and a maximum antibody token length of 512. Early stopping has patience 10, and the checkpoint is selected by

overall validation loss. Losses for the Oncology, CHDC, Other, and Remainder groups are also recorded for monitoring.

Antigen clustering uses MMseqs2 with minimum sequence identity 0.30, coverage 0.8, coverage mode 2, and cluster mode 0. Training, validation, and test sets contain 5, 140, and 159 clusters, respectively, with no cluster shared between sets. Antigen groups are corpus annotations, not clustering units. The Other category contains antigens without an assigned group.

Sequence filter F1 requires AbNumber (Dunbar & Deane, 2016) numbering of both chains under the IMGT scheme. F2 accepts VH lengths of 80–160 residues and VL lengths of 75–135 residues. F3 excludes exact VH/VL pairs in the training split. F4 requires a nonempty CDR3 annotation for each chain.

In matched comparisons, each target contributes the mean score of three generated designs and the score of its known binder. We calculate percentile bootstrap intervals from 20,000 resamples of target-level differences and apply the Wilcoxon signed-rank test. The MAGE comparison contains different designs against one antigen, so we resample designs independently within each set and apply the two-sided Mann–Whitney test. The reported decision rule requires an interval excluding zero and a rank-test p-value below 0.05.

B Sequence Quality and Predicted Developability

B.1 Sequence retention and CDR3 novelty

Of the 10,040 generated candidates, 404 fail antibody numbering (F1) and another 37 fail the chain length limits (F2), leaving 9,599 antibody pairs. The training overlap filter (F3) and CDR3 annotation filter (F4) remove no additional candidates. Table 3 compares the retained sequences with held-out and training antibodies.

Table 3: Sequence properties of generated, held-out, and training antibodies. CDR3 distances are mean normalized edit distances to the nearest neighbor among 15,805 training antibodies. The final three rows report fractions above the training set's 95th percentile for each DeepSP (Kalejaye et al., 2024) descriptor. The training fraction is 5% by construction.

	our designs 9,599	real, held out 1,980	real, train 15,805
CDR3-H length	16.11	16.19	16.23
CDR3-L length	10.05	9.63	9.73
CDR3-H to nearest training	0.407	0.470	n/a
CDR3-H identical to one	0.01%	0.05%	n/a
CDR3-L to nearest training	0.161	0.249	n/a
CDR3-L identical to one	3.95%	1.97%	n/a
above p95, SAP Fv	2.85%	5.25%	5%
above p95, SCM^+ Fv	4.24%	6.06%	5%
above p95, SCM^- Fv	4.85%	2.47%	5%

Generated heavy-chain CDR3s have a mean nearest-training distance of 0.407, compared with 0.470 for held-out antibodies. Exact matches occur in 0.01% and 0.05%, respectively (Table 4). Against the held-out reference, generated heavy-chain CDR3s have a mean distance of 0.425 with no exact matches. Light-chain CDR3s are closer to both reference sets. Their mean nearest-training distance is 0.161, compared with 0.249 for held-out antibodies, and their exact-match rate is 3.95%, compared with 1.97%. Against held-out light-chain CDR3s, generated sequences have a mean distance of 0.155 and an exact-match rate of 8.27%.

B.2 Predicted aggregation and surface charge

DeepSP (Kalejaye et al., 2024) predicts broadly similar aggregation and surface charge distributions for generated and held-out antibodies (Table 5). Table 6 provides the training reference. Figure 8 compares the fractions above each descriptor's training-set 95th percentile. For whole-F_v SAP,

Table 4: CDR3 nearest-neighbor distances, weighted by family size. Distance is the mean normalized edit distance from each query CDR3 to its nearest neighbor in the reference pool.

query	reference	chain	distance	identical
our designs	training	H	0.407	0.01%
real, held out	training	H	0.470	0.05%
our designs	held out	H	0.425	0.00%
our designs	training	L	0.161	3.95%
real, held out	training	L	0.249	1.97%
our designs	held out	L	0.155	8.27%

2.85% of generated and 5.25% of held-out antibodies exceed this threshold. For whole-F_v negative SCM, the fractions are 4.85% and 2.47%.

DeepSP returns valid scores for 9,578 of the 9,599 generated candidates and all 1,980 held-out and 15,805 training antibodies. Distribution summaries use valid scores. Threshold percentages in Figure 8 and Table 3 use the full sequence sets as denominators. The 21 generated candidates without scores contribute to the denominator but not to the count above the threshold.

Table 5: DeepSP predictions for generated and held-out antibodies. Values are medians with the 25th and 75th percentiles in brackets. Sample counts include only sequences with valid scores. SAP denotes spatial aggregation propensity and SCM denotes the spatial charge map.

Descriptor	Generated ($n = 9{,}578$)	Held out ($n = 1{,}980$)
SAP, CDR	31.3 [24.8, 39.3]	34.7 [27.4, 41.7]
SAP, CDR3-H	15.0 [9.7, 21.6]	16.0 [10.1, 22.4]
SAP, F_v	95.5 [87.8, 104.2]	101.0 [92.8, 110.0]
SCM$^-$, CDR	354.9 [266.3, 466.2]	343.4 [253.5, 447.3]
SCM$^-$, F_v	896.6 [765.3, 1048.9]	898.1 [778.1, 1059.9]
SCM$^+$, CDR	275.4 [210.4, 357.1]	326.0 [246.6, 417.1]
SCM$^+$, F_v	2207.6 [2062.1, 2355.5]	2257.3 [2089.2, 2427.4]

Table 6: DeepSP descriptors for the 15,805 training antibodies. Values are medians with the 25th and 75th percentiles in brackets. Table 5 reports the corresponding generated and held-out distributions.

Descriptor	Training
SAP, CDR	33.6 [26.9, 41.0]
SAP, CDR3-H	14.6 [9.8, 21.3]
SAP, F_v	99.9 [91.9, 109.0]
SCM$^-$, CDR	384.3 [286.3, 508.5]
SCM$^-$, F_v	964.2 [815.3, 1136.0]
SCM$^+$, CDR	287.0 [215.1, 370.5]
SCM$^+$, F_v	2206.0 [2043.5, 2364.5]

B.3 Sequence composition and CDR3 cysteine content

We examined sequence composition features motivated by the Bits to Binders study (Kosonocky et al., 2026). For VH/VL pairs, we measured cysteine occurrence and the maximum combined fraction of lysine and glutamate within a 15-residue window. This K/E measure uses sequence windows, while the original study used structural information to identify K/E enrichment in alpha helices.

The K/E window flags 37.9% of generated antibodies, 27.4% of training antibodies, 24.7% of held-out antibodies, and 53.9% of validation antibodies (Table 7). This variation across reference splits limits its use as a general antibody quality filter. Whole-domain cysteine occurrence is also nearly universal because antibody frameworks contain conserved cysteines.

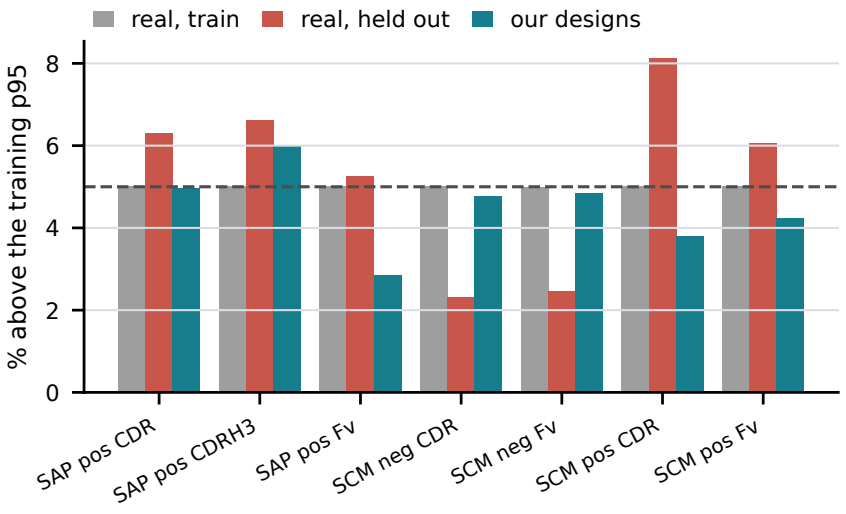


Figure 8: Fractions of generated, held-out, and training antibodies above each DeepSP descriptor's training-set 95th percentile. The same threshold is used for all sets, placing the training reference at approximately 5%. Percentages use the full sequence sets as denominators. Tables 5 and 6 report distribution medians and quartiles.

CDR3 cysteine content reveals a more specific difference. A cysteine occurs in at least one CDR3 loop in 27.8% of generated antibodies and 11.5% of held-out antibodies. Heavy-chain rates are similar at 8.8% and 8.6%, while light-chain rates are 20.8% and 3.0%. This light-chain enrichment, together with the greater CDR3-L overlap with reference sequences in Table 4, identifies light-chain generation as an area for improvement. CDR3 cysteines should be reviewed when selecting candidates for synthesis.

Table 7: Sequence composition annotations in generated antibodies and reference splits. The K/E flag indicates that the combined fraction of lysine and glutamate exceeds 30% in at least one 15-residue window in either chain. Cysteine rows report occurrence anywhere in VH/VL or within the specified CDR3 loops.

annotation	ours	real, train	real, test	real, val
Any cysteine in VH/VL	100%	100%	99.9%	100%
K/E window $>$ 30%	37.9%	27.4%	24.7%	53.9%
CDR3 cysteine	27.8%	14.5%	11.5%	4.7%
heavy chain CDR3	8.8%	10.2%	8.6%	3.3%
light chain CDR3	20.8%	4.6%	3.0%	1.4%

C Structural Screening and Predictor Comparisons

C.1 Candidate selection and RF3 results

The screened panel contains 200 complexes selected from the generated pool. Table 8 gives the candidate count at each stage, and Table 9 reports all RF3 metrics for this panel.

We screened candidates with Protenix Mini after ranking them by the mean ESMFold pLDDT of their heavy and light chains. We retained up to five candidates per target above the global 60th percentile, a pLDDT of 78.9. The procedure required at least two candidates per target, but every target met this minimum above the threshold. Of the 2,478 selected candidates, 2,470 completed Protenix Mini screening.

We then selected 200 complexes by Protenix ranking score, with antigen group quotas and at most three designs per target. The quotas were 10 each for Allergen, Malaria, and Toxin, 14 for HCV, 16 for Oncology/Immune, and 20 each for the seven remaining groups. Unfilled slots were reassigned across groups by ranking score while retaining the three-design limit. Pneumovirus filled 9 of its 20 slots, and the remaining 11 went to Other. The final panel covered 110 targets, and all 200 RF3 evaluations completed.

Table 8: Candidate counts through sequence filtering and structural screening.

Stage	Count
Generated VH/VL pairs	10,040
Passed sequence filters	9,599
Completed ESMFold predictions	9,599
Entered Protenix Mini screening	2,478
Completed Protenix Mini screening	2,470
Selected for and completed RF3 evaluation	200

Table 9: RF3 results for the screened panel of 200 complexes across 110 targets, using antigen-only alignments. pLDDT is on a 0–100 scale. PAE is in Å and lower values indicate greater confidence.

	mean	median	min	max
interface pTM (ipTM)	0.5211	0.5262	0.2175	0.8408
pTM	0.6556	0.6699	0.3280	0.8908
overall pLDDT	78.77	79.05	72.94	82.88
ranking score	0.5480	0.5527	0.2555	0.8508
target–antibody PAE	15.27	15.13	2.00	26.70
VH–VL PAE	1.52	1.23	0.85	10.89
counts over the 200				
ipTM ≥ 0.6 / ≥ 0.7 / ≥ 0.8				61 / 37 / 3
target–antibody PAE ≤ 15 Å				97
both ipTM ≥ 0.6 and PAE ≤ 15 Å				57
complexes with a steric clash				0

Mean minimum PAE is 1.52 Å for VH–VL and 15.27 Å for antigen–antibody placement. RF3 is therefore more confident in the relative placement of the antibody chains than in their placement against the antigen. Of the 61 complexes with ipTM at least 0.6, 57 also have minimum antigen–antibody PAE at most 15 Å. Figure 10 shows the association between ipTM and total complex length in the screened and matched panels.

C.2 Screening and target composition

In the screened panel, 30.5% of complexes have ipTM of at least 0.6. Reweighting its target-length composition to that of the candidate pool gives 10.7%. Within the screened panel, the yields for the five antigen-length bands defined in Figure 5b are 91%, 5%, 0%, 0%, and 0%, and all qualifying complexes have antigens of at most 129 residues. These values describe different target compositions, so the screened-panel yield is reported separately from the full candidate pool.

Protenix Mini and RF3 select the same top candidate for 26 of 64 targets, with a median within-target rank correlation of 0.00.

C.3 Comparison of RF3 and AlphaFold3

We evaluated the same 200 screened complexes with AlphaFold3 (Abramson et al., 2024), using the antigen-only alignments from the RF3 evaluation. Both predictors completed all complexes. AlphaFold3 and RF3 give similar mean ipTM scores, 0.5224 and 0.5211, with a rank correlation of 0.927 (Figure 9a). AlphaFold3 places 68 complexes above ipTM 0.6 and RF3 places 61 above this threshold, with 56 shared between them (Figure 9b). They also share 43 of their top 50 complexes by ipTM.

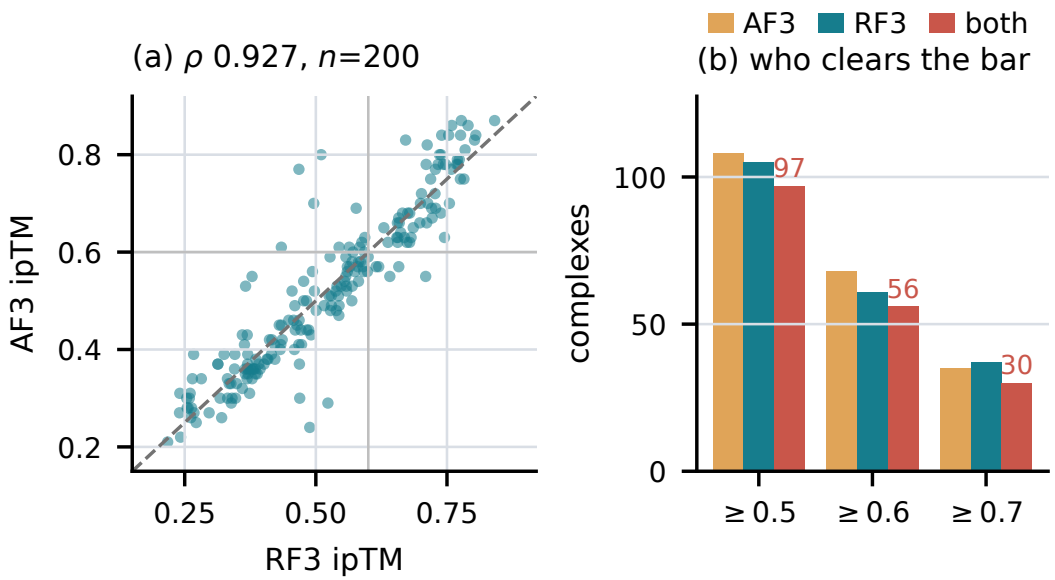


Figure 9: AlphaFold3 and RF3 results for the same 200 complexes with identical antigen alignments. (a) ipTM scores, with the dashed line marking $y = x$. (b) Counts above each ipTM threshold for each predictor and for both predictors.

Much of this agreement follows the association with complex length (Appendix D.1). The length correlations are -0.947 for AlphaFold3 and -0.965 for RF3. After controlling for length, their partial rank correlation falls from 0.927 to 0.150. Within targets, the predictors select the same top candidate for 30 of 64 targets with multiple candidates and agree on 60 of 105 pairwise orderings. Their antigen group means correlate at 0.986, largely reflecting differences in target length. Thus the predictors agree on the overall panel scores more strongly than on candidate rankings within individual targets.

C.4 Alignment loading and score sensitivity

Before RF3 inference, we load each input through the predictor's alignment resolution path and require every target chain to have an alignment attached. Checking only that alignment files exist is insufficient because RF3 resolves them using a hash of the query sequence.

An earlier run used antigen names for alignment filenames, so the target alignments did not load. Correcting the filenames increased mean pLDDT by 0.0292 on the 0–1 scale and mean ipTM by 0.0080. The correlation between antigen-group rankings before and after correction was 0.17 for pLDDT and 0.979 for ipTM. The final analyses use the corrected alignment loading procedure.

D Matched Evaluation Across Held-Out Antigens

Across 496 targets, AbGate and known binders have comparable mean RF3 ipTM and overall pLDDT. For each target, we compare the mean score of three generated designs with the score of one known binder. The generated mean ipTM is higher for 231 of the 496 targets. Table 10 gives paired differences, using the bootstrap and rank tests in Appendix A.3.

Minimum antigen–antibody PAE is the smaller of the minimum PAE values for the antigen–VH and antigen–VL chain pairs.

Figure 2b shows mean ipTM by antigen group, and Figure 6 compares chain-pair PAE.

Table 10: Paired RF3 score differences on the matched panel. Differences are generated minus known binder, averaged across 496 targets. pLDDT is on a 0–1 scale and PAE is in Å.

Metric	Mean difference	95% bootstrap interval	Wilcoxon p
ipTM	-0.0109	$[-0.0200, -0.0018]$	0.131
Overall pLDDT	-0.00074	$[-0.00165, +0.00017]$	0.439
Minimum antigen–antibody PAE	$+1.71$	$[+1.24, +2.19]$	2.47×10^{-8}

D.1 Antigen length and score variation

RF3 ipTM decreases with antigen length in both groups. For antigens of at most 100 residues, mean ipTM is 0.682 for designs and 0.701 for known binders. For antigens longer than 800 residues, the means are 0.271 and 0.274 (Figure 5b). Spearman correlations with total complex length are -0.851 and -0.758, respectively (Table 11, Figure 10). Total complex length includes the antigen and both antibody chains, and these correlations use all completed complexes in each group.

The PAE scores distinguish antibody chain assembly from antigen placement. Mean minimum VH–VL PAE is lower for AbGate than for known binders, at 3.02 versus 3.85 Å. Minimum antigen–antibody PAE is higher, at 15.42 versus 13.71 Å. In both groups, antigen–antibody PAE has a weaker association with complex length than ipTM.

Table 11: RF3 scores and correlations with total complex length in the matched panel. Means and Spearman correlations are calculated separately for 1,488 generated designs and 496 known binders. Complex length includes the antigen, VH, and VL. pLDDT is on a 0–1 scale and PAE is in Å.

	Mean score		Spearman ρ	
Metric	Designs	Known binders	Designs	Known binders
ipTM	0.460	0.471	-0.851	-0.758
Minimum antigen–antibody PAE	15.42	13.71	$+0.542$	$+0.546$
Minimum VH–VL PAE	3.02	3.85	$+0.508$	$+0.537$
Overall pLDDT	0.787	0.788	$+0.092$	-0.040

We summarize score variation by dividing the sample variance of the 496 target means by the mean sample variance of the three designs within each target. All sample variances use the $n - 1$ denominator. The ratio is 3.84 for ipTM and 1.17 for minimum antigen–antibody PAE (Table 12). Thus, variation between target means is larger relative to variation within targets for ipTM.

Table 12: Variation across and within targets for generated designs. The ratio divides the variance of target means by the mean variance within targets.

Metric	Variance ratio
ipTM	3.84
Minimum antigen–antibody PAE	1.17
Minimum VH–VL PAE	1.29
Overall pLDDT	6.85

D.2 Structural evaluation coverage

RF3 predictions are missing for 24 complexes from six antigens longer than 800 residues. The longest length band therefore contains 15 targets rather than 21. Memory demand depends on complex length, alignment depth, and memory retained between predictions. On a 40 GiB GPU, complexes of 1,570–1,574 residues completed at a 39.0 GiB peak, while complexes of 1,522–1,528 residues with deeper alignments failed. Retrying the latter on a 48 GiB GPU succeeded at 40.3–40.7 GiB. Complexes of at least 1,811 residues still failed, with peak memory reaching 47.4 GiB.

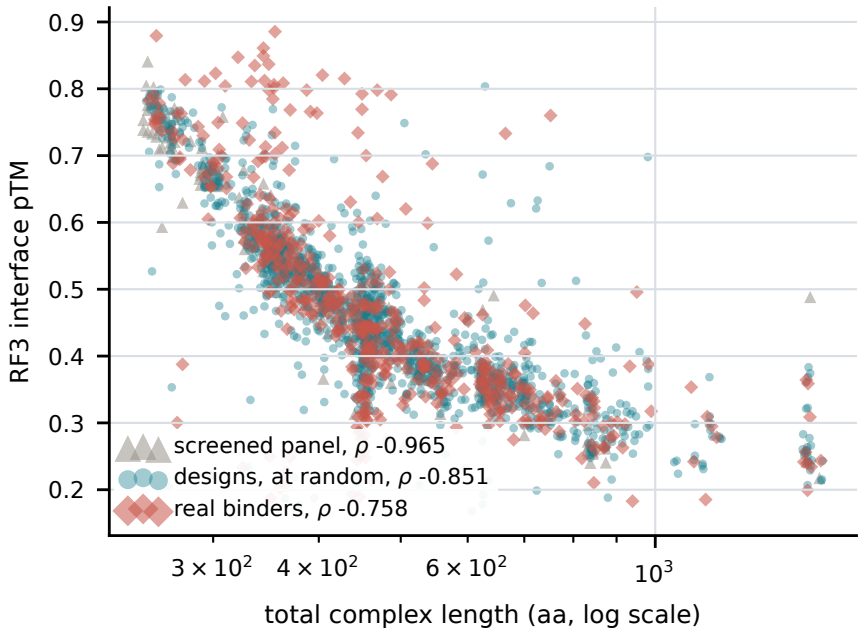


Figure 10: RF3 ipTM against total complex length in the screened and matched panels. Total length includes the antigen, VH, and VL. Spearman correlations are -0.965 for the screened panel, -0.851 for randomly sampled designs, and -0.758 for known binders in the matched panel.

E Additional MAGE Comparisons

E.1 Shared target and evaluation settings

We use the 223-residue SARS-CoV-2 RBD sequence from MAGE's released generation run. This sequence appears in 3,000 records in AbGate's training split and is absent from its validation and test splits. All 20 AbGate samples pass the sequence filters in Section 3.1. We compare them with the 180 designs with the highest ESMFold pLDDT among MAGE's 991 released designs. All 200 complexes complete RF3 prediction using the same antigen alignment and prediction settings.

The two groups contain different designs against one target. We use an unpaired bootstrap and a two-sided Mann–Whitney test, as described in Appendix A.3.

E.2 Structural scores and score distributions

AbGate and MAGE achieve comparable mean ipTM and overall pLDDT, while MAGE has lower mean antigen–antibody PAE (Table 13). For each complex, minimum antigen–antibody PAE is the smaller of the antigen–VH and antigen–VL chain-pair minima. The chain-averaged minimum PAE is the average of these two minima.

The median PAE values are closer than the means. Median minimum antigen–antibody PAE is 19.11 Å for AbGate and 18.93 Å for MAGE. For chain-averaged minimum PAE, the medians are 19.46 and 19.03 Å, respectively.

Fifteen of the 20 AbGate designs rank in the top 100 of the 200-complex panel by ipTM. The best AbGate design ranks 27th, and the median rank is 70.5. MAGE includes higher maximum scores: 0.853 compared with 0.463 for AbGate. In 20,000 subsamples of 20 designs drawn without replacement from the evaluated MAGE pool, the median maximum is 0.605. AbGate's maximum lies at the 3.7th percentile of this distribution.

Table 13: Structural scores for 20 AbGate and 180 MAGE designs on the shared RBD target. Values are means across designs. Differences are AbGate minus MAGE. Intervals are 95% bootstrap intervals for the mean difference. pLDDT is on a 0–1 scale. PAE is in Å, with lower values indicating greater confidence.

Metric	AbGate	MAGE	Mean diff.	95% interval	Rank p
ipTM	0.4343	0.4353	-0.0010	$[-0.0140, +0.0115]$	0.059
pTM	0.5605	0.5677	-0.0072	$[-0.0163, +0.0014]$	0.346
RF3 ranking score	0.4595	0.4618	-0.0023	$[-0.0144, +0.0094]$	0.080
Ag–Ab minimum PAE	18.98	17.47	$+1.51$	$[+0.56, +2.48]$	0.445
Chain-averaged min. PAE	19.33	17.88	$+1.46$	$[+0.62, +2.33]$	0.396
Overall pLDDT	0.7870	0.7890	-0.0020	$[-0.0044, +0.0004]$	0.112

E.3 Humanness and similarity to training sequences

The two sequence pools have similar mean OASis humanness percentiles: 0.881 for AbGate's 9,599 filtered designs across 502 test targets and 0.878 for MAGE's 991 released RBD designs. For AbGate, we calculate the fraction of 9-mers found in at least 10% of human OAS subjects and convert it to an OASis percentile using BioPhi (Prihoda et al., 2022). We use the released OASis percentiles for MAGE and report both on a 0–1 scale.

For the shared RBD target, AbGate designs are less similar to the training antibody chains than MAGE designs. Exact light-chain matches occur in 1 of 20 AbGate designs (5%) and 51 of 180 MAGE designs (28%). Mean VH similarity to the nearest training chain is 0.919 for AbGate and 0.935 for MAGE (Mann–Whitney $p = 0.026$). Neither group contains an exact copy of a training VH/VL pair. All six designs with ipTM at least 0.6 are from MAGE, and four have an exact light-chain match in the training reference set.

The reference set contains 3,000 training pairs with the same RBD sequence. For each generated chain, we take the highest Python `SequenceMatcher` ratio among reference chains whose lengths differ by at most 12 residues, with `autojunk=False`. The ratio is $2M/(L_q + L_r)$, where M is the number of residues in matching blocks and L_q and L_r are the two chain lengths.

E.4 Overlap with experimentally tested designs

Eight of MAGE's 20 experimentally tested designs occur in the 180-design structural panel. Their mean ipTM is 0.4202, compared with 0.4360 for the other 172 designs (Mann–Whitney $p = 0.97$). None of the six designs with ipTM at least 0.6 belongs to the tested subset. Appendix F reports the binding-label analysis for all 20 tested designs.

F Interface Confidence and Experimental Binding Labels

We compare RF3 confidence with experimental binding labels for MAGE's 20 tested RBD designs: nine ELISA binders and eleven nonbinders (Wasdin et al., 2025). All designs are predicted against the same 223-residue RBD using the antigen alignment and prediction settings of the MAGE comparison. Eight predictions come from that comparison panel, and the other twelve use the same protocol.

We report the area under the receiver operating characteristic curve (AUC). For this calculation, we orient the scores so that higher ipTM and pLDDT and lower minimum antigen–antibody PAE favor the binder class. The 95% intervals use 20,000 bootstrap resamples, drawn independently within the nine binders and eleven nonbinders, with seed 20260831.

ipTM has an AUC of 0.576 (Table 14). Selecting the top nine designs by ipTM retrieves five binders, compared with 4.05 expected under random selection. The wide intervals reflect the small size of this panel. Its ipTM scores range from 0.350 to 0.509.

Table 14: RF3 confidence against MAGE's 20 ELISA-labeled RBD designs. The panel contains 9 binders and 11 nonbinders. Intervals are stratified bootstrap intervals for AUC.

Metric	AUC	95% interval
ipTM	0.576	[0.293, 0.838]
Minimum antigen–antibody PAE	0.566	[0.283, 0.838]
Overall pLDDT	0.566	[0.303, 0.818]

G Task Settings of Related Methods

MAGE shares our task of generating paired antibody sequences from an antigen sequence. We therefore compare the two generators under common settings for complex prediction (Section 3.3). Table 15 summarizes the inputs and outputs of related methods. These methods cover different tasks, including sequence generation, loop design, and backbone generation.

Table 15: Inputs and outputs of related generative methods.

Method	Input	Output
AbGate	Antigen sequence	Paired VH/VL sequences
MAGE (Wasdin et al., 2025)	Antigen sequence as a prompt	Paired VH/VL sequences
ProCALM (Yang et al., 2025)	Function labels, taxonomy, or text	Protein sequence
DiffAb (Luo et al., 2022)	Antibody framework and target structure	CDR loops in a given complex
IgDiff (Cutting et al., 2025)	No input or an existing antibody	Antibody backbone
RFdiffusion (Watson et al., 2023)	Target structure	Binder backbone