AntiFold: Improved antibody structure design using inverse folding

Alissa M. Hummer[1]*, **Magnus Haraldson Høie**[2*], **Tobias H. Olsen**[1], **Morten Nielsen**[2], **Charlotte M. Deane**[1†]
[1] Department of Statistics, University of Oxford, United Kingdom
[2] Department of Health Technology, Section for Bioinformatics
Technical University of Denmark, Denmark

Abstract

The design and optimization of antibodies, important therapeutic agents, requires an intricate balance across multiple properties. A primary challenge in optimization is ensuring that introduced sequence mutations do not disrupt the antibody structure or its target binding mode. Protein inverse folding models, which predict diverse sequences that fold into the same structure, are promising for maintaining structural integrity during optimization. Here we present AntiFold, an inverse folding model developed for solved and predicted antibody structures, based on the ESM-IF1 model. AntiFold achieves large gains in performance versus existing inverse folding models on sequence recovery across all antibody complementarity determining regions (CDRs) and framework regions. AntiFold-generated sequences show high structural agreement between predicted and experimental structures. The tool efficiently samples hundreds of antibody structures per minute, providing a scalable solution for antibody design. AntiFold is freely available as a downloadable package at: `https://opig.stats.ox.ac.uk/data/downloads/AntiFold`.

1 Introduction

Antibodies are one of the largest classes of therapeutics, used to treat diseases ranging from cancers to viruses (Lu et al., 2020). Therapeutic antibody design is complex, requiring the optimization of numerous properties related to efficacy, manufacturability and safety (Rabia et al., 2018).

Computational, and in particular machine learning, tools demonstrate promise for accelerating multiple steps in the antibody development pipeline (Hummer et al., 2022). These approaches can be used to design antibodies by reducing liabilities such as immunogenicity and aggregation, or to rationally optimize for desirable properties such as binding affinity and developability (Marks et al., 2021; Prihoda et al., 2022; Tennenhouse et al., 2023; Makowski et al., 2022, 2023; Harvey et al., 2022). However, most current approaches only focus on one or a very small number of properties, and any changes to the antibody sequence may detrimentally impact other features.

A guiding consideration in antibody optimization is to select mutations which retain the structure, and thus biophysical characteristics such as stability and the target (antigen) binding mode. There is therefore a need for models that can suggest mutations which will be structurally tolerated at particular positions. Inverse folding models are trained to predict sequence given structure (Ingraham et al., 2019), and could therefore be used to generate novel sequences without altering the antibody structure.

There have been many advances in the development of inverse folding models for general

[*] AMH and MHH contributed equally to this work
[†] Corresponding author Charlotte M. Deane, deane@stats.ox.ac.uk

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

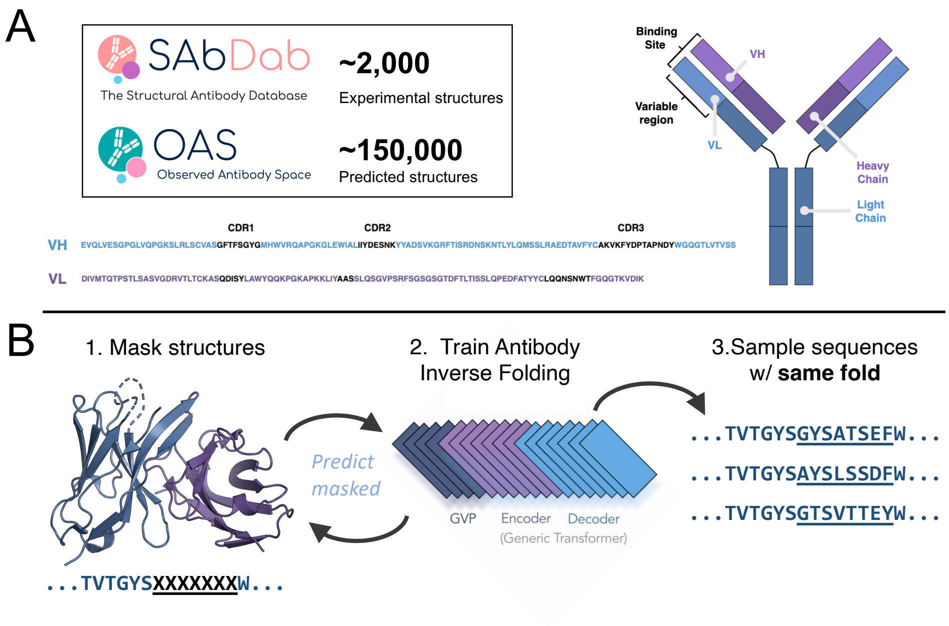
A
SAbDab
The Structural Antibody Database
~2,000
Experimental structures
OAS
Observed Antibody Space
~150,000
Predicted structures
Binding Site
Variable region
VH
VL
Heavy Chain
Light Chain
CDR1
CDR2
CDR3
VH EVQLVESGPGLVQPGKSLRLSCVASGFTFSGYGMHWVRQAPGKGLEWIALIIYDESNKYYADSVKGRFTISRDNSKNTLYLQMSSLRAEDTAVFYCAKVKFYDPTAPNDYWGQGTLVTVSS
VL DIVMTQTPSTLSASVGDRVTLTCKASQDISYLAWYQQKPGKAPKKLIYAASSLQSGVPSRFSGSGSGTDFTLTISSLQPEDFATYYCLQQNSNWTFGQGTKVDIK
B
1. Mask structures
2. Train Antibody Inverse Folding
3.Sample sequences w/ same fold
Predict masked
GVP
Encoder
Decoder
(Generic Transformer)
...TVTGYSGYSATSEFW...
...TVTGYSAYSLSSDFW...
...TVTGYSGTSVTTEYW...
...TVTGYSXXXXXXXW...

Figure 1: (A) AntiFold is trained and evaluated on solved antibody structures from SAbDab (Dunbar et al., 2014; Schneider et al., 2021) and structures of antibody sequences from OAS (Kovaltsuk et al., 2018; Olsen et al., 2022) modeled with ABodyBuilder2 (Abanades et al., 2023). Antibodies consist of heavy (blue) and light (purple) chains. Target binding is primarily mediated by complementarity-determining regions (CDRs), in the variable region. Examples of heavy (VH) and light (VL) variable domain sequences are shown. (B) AntiFold is initialized with weights from ESM-IF1 (Hsu et al., 2022), then fine-tuned on antibody variable domain structures. AntiFold can generate diverse sequences maintaining the fold of the input structure. Figure adapted from (Olsen et al., 2022; Hsu et al., 2022). Structure and sequence from PDB 3W2D (Tian Xia and Guo, 2014).

proteins in recent years (Ingraham et al., 2019; Strokach et al., 2020; Anand et al., 2022; Jing et al., 2021; Hsu et al., 2022; Dauparas et al., 2022). Antibodies, however, have distinct structure and sequence properties (Stanfield and Wilson, 2014) (Figure 1A). The framework (FR) regions, which are mostly germline-encoded, are relatively conserved while the complementarity-determining region (CDR) loops are hypervariable and structurally less well-conserved (Figure 1A). The CDR loops are especially challenging for structure prediction and modeling tasks, but are of great interest as they form most of the antigen binding site. Training inverse folding models specifically on antibody structures should improve our understanding of the immunoglobulin fold sequence-structure relationship.

An antibody inverse folding model, AbMPNN (Dreyer et al., 2023), based on ProteinMPNN (Dauparas et al., 2022), has recently been released. It is, to the best of our knowledge, the only antibody fine-tuned inverse folding model. While this model demonstrated that performance gains can be realized from fine-tuning, the sequence recovery on the CDR loops was limited. Additionally, this architecture has several features, including the occasional reordering of antibody heavy and light chains, reversal of residues in the CDRH3 112 positions, and insertion of residues into gaps in IMGT-numbered antibodies, incompatible with antibody structures.

Here we present AntiFold, an antibody inverse folding model based ESM-IF1 (Hsu et al., 2022), an architecture which has successfully been applied to protein-protein interaction, small molecule binding site (Carbery et al., 2023) and B-cell epitope prediction (Høie et al., 2023). AntiFold is fine-tuned on both solved and predicted antibody structures, and achieves state-of-the-art performance on antibody sequence recovery across framework and CDR regions. Structural models of the predicted sequences show high structural similarity with the experimentally solved structures. The use of AntiFold in tandem with other property prediction tools, to guide mutations, could therefore improve the success rates of *in silico* antibody optimization.

2 Results

2.1 Data

We fine-tuned ESM-IF1 on solved and predicted antibody structures. To enable a direct comparison with AbMPNN, we trained, validated and tested our model on the same data: 2,074 solved complexes from the Structural Antibody Database (SAbDab) (Dunbar et al., 2014; Schneider et al., 2021), and 147,458 structures of sequences from the Observed Antibody Space (OAS) paired database (Kovaltsuk et al., 2018; Olsen et al., 2022) modeled with ABodyBuilder2 (Abanades et al., 2023). Each dataset was split using a 90% concatenated CDR sequence identity cutoff (80/10/10 train/validation/test) (Dreyer et al., 2023).

2.2 Fine-tuning strategy

Fine-tuning from a general protein inverse folding model enabled us to benefit from existing knowledge learned by ESM-IF1, which was trained on millions of structures. We explored the effect of multiple parameters on fine-tuning ESM-IF1 on antibody structures.

When fine-tuning on a new task or domain, there is a risk of "catastrophically forgetting" previously learned knowledge. We therefore applied a strategy of layer-wise learning rate decay, successfully used to fine-tune BERT models (Sun et al., 2019). We evaluated exponentially decaying the learning rate from the last to the first layer, preserving the weights of earlier parts of the model during training (see Figure 1B and Methods). Layer-wise learning rate decay did not further improve sequence recovery (Appendix Table A1-3), however we retained it for subsequent training to reduce the risk of overfitting and maintain generalization towards untested properties.

We also investigated different masking schemes in training. Shotgun masking hides the coordinates of randomly selected single positions, while span masking is applied to a consecutive stretch of positions. As FR and CDR regions in the antibody structure have different levels of variability, we tested biasing the selection of masked positions towards the more variable CDR residues (IMGT-weighted masking). In total, 15% of the backbone residues were masked during training (for more details on the masking parameters, see Methods). As previously reported (Hsu et al., 2022), we found stronger performance for shotgun than span masking on unmasked test structures. However, span masking improved CDR sequence recovery for test cases with masked CDR loops, a realistic design use case (Appendix Table A1-3). IMGT-weighted masking further improved performance on CDR loops, while only slightly reducing sequence recovery on FR regions (Appendix Table A1-2).

We included a large dataset of predicted structures in our fine-tuning strategy, in an aim to boost performance by training on more diverse antibodies. We tested the effects of adding Gaussian noise at a scale of 0.1 Å to the modeled protein backbone, previously found to improve performance (Hsu et al., 2022; Dauparas et al., 2022). We found no substantial effect, but have included it in our final model for robustness towards minor variations in input structures (Appendix Table A3).

Based on these results, we chose to train the final AntiFold model with IMGT-weighted shotgun and span masking, layer-wise learning rate decay and added Gaussian noise on predicted structures. We note that these augmentations, along with the use of the larger pre-trained ESM-IF1 architecture (142M parameters) instead of ProteinMPNN (1.7M parameters), comprise the main differences with AbMPNN. We split the training of AntiFold into two phases. First we fine-tuned ESM-IF1 on one pass of the training dataset of predicted structures from OAS. Next we fine-tuned the model on the solved training dataset, stopping training when there was no further improvement in validation loss for 10 epochs. This model, termed AntiFold, was used for all subsequent analysis.

2.3 Fine-tuning improves amino acid recovery on antibody sequences

AntiFold demonstrated a dramatic improvement in amino acid recovery (AAR) on the test set (solved structures) as compared to the original ESM-IF1 model (43 to 60% for CDRH3; Figure 2A, Appendix Table A1). AntiFold also outperformed AbMPNN across all CDR regions (Antifold 75-84%, AbMPNN 63-76% excluding CDRH3, Figure 2A) and most framework regions (Antifold 87-94%, AbMPNN 85-89%, Appendix Figure A1). Performance was lowest across all models for CDRH3 (AntiFold 60% AAR), corresponding with the challenge of predictive tasks for this loop.

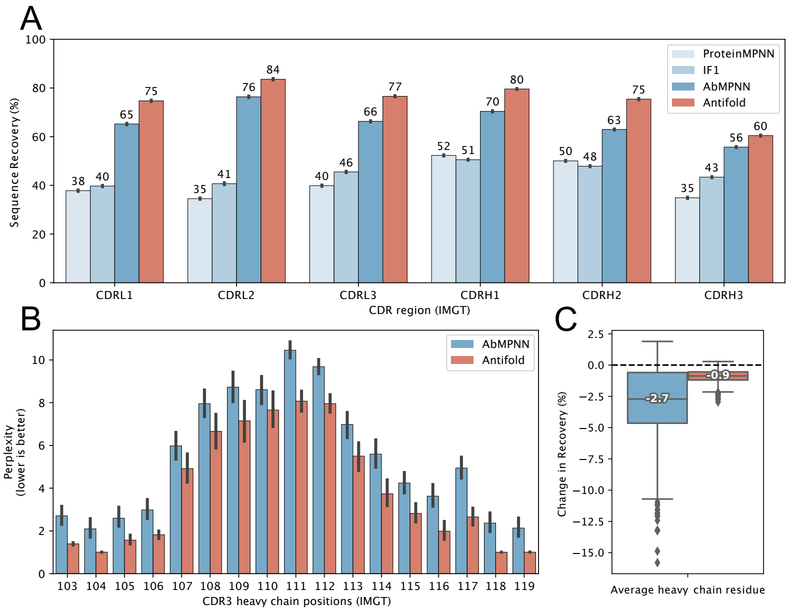

Figure 2: AntiFold performance on CDR design versus AbMPNN, ProteinMPNN and ESM-IF1. (A) Mean amino-acid recovery (AAR) across CDRs for antibody heavy and light chains. (B) Perplexity across the CDRH3 loop. Error bars indicate 95% confidence intervals after bootstrapping 1000 times with replacement. (C) Percent change in AAR when swapping predicted structures for solved structures (same sequences).

We note that the pre-trained, ESM-IF1 model slightly outperforms ProteinMPNN on CDR and FR sequence recovery. This aligns with previous findings on CATH 4.2 / 4.3 test sets (Dawson et al., 2016), showing similar ESM-IF1 sequence recovery versus ProteinMPNN (Gao et al., 2023).

We confirmed AntiFold can be accurately applied to modelled structures by testing on ABodyBuilder2 predictions of structures in the test set. AntiFold achieved similar AAR for solved and predicted structures, unlike AbMPNN which performed slightly worse on solved structures (Figure 2C, Appendix Figure A2).

We also calculated the perplexity, representing the average number of amino acid suggestions per position, across positions in the solved structures (see Methods). A random model (assigning equal probability to all 20 possible amino acids) would have a perplexity of 20, while an oracle model, assigning 100% confidence to a single amino acid, would have a perplexity of 1. AntiFold suggests on average ∼2-8 mutations in the CDRH3 which are likely to preserve the fold of the loop, versus ∼3-10 for AbMNN, reflecting AntiFold's improved accuracy (Figure 2B).

2.4 Predicted sequences have good structural agreement with experimental structures

To assess whether suggested mutations preserve the fold of the CDRs, we identified 56 high-quality antibody structures in the test set, solved using X-ray crystallography and with a resolution below 2.5 Å. Next, we sampled 20 sequences for each antibody using ESM-IF1, AbMPNN and AntiFold. We used a sampling temperature of 0.20, a hyperparameter which controls how often higher probability residues are selected versus less likely mutations, the same default as used by ProteinMPNN and AbMPNN (see Methods).

We modeled these sequences using ABodyBuilder2, aligned them with the framework backbone of their experimentally solved counterpart, then calculated root-mean-square deviation (RMSD) over the CDR residues (for more details, see Methods). As a baseline, we modeled the true sequences with ABodyBuilder2 (native, Appendix Figure A3). AntiFold generates sequences with high structural similarity to the original backbone, with a median CDR region RMSD of 0.67 (versus native RMSD 0.48, AbMPNN 0.74, ESM-IF1 0.75). The alignment does not take into account the modeling accuracy of side-chain atoms.

3 Conclusions

Therapeutic antibody optimization depends on the ability to identify mutations that retain the antibody fold and related structural properties. Fine-tuning a general protein inverse folding model, ESM-IF1, on antibody structures allowed us to take advantage of what the model learned from the millions of structures in its training dataset, while improving performance further on antibody structures.

The developed model, AntiFold, achieved state-of-the-art sequence recovery, outperforming general protein and fine-tuned antibody inverse folding models. These results highlight the value of fine-tuning for improving task-specific performance. The sequences sampled from AntiFold exhibit high structural similarity with experimental structures, suggesting this model could be used to guide antibody optimization and reduce development liabilities.

AntiFold samples $\sim$300 antibody structures per minute on a Nvidia GTX 1080 Ti GPU. The model is freely available at `https://opig.stats.ox.ac.uk/data/downloads/AntiFold`.

4 Methods

4.1 Data

4.1.1 Experimental antibody structures from SAbDab

The AbMPNN dataset contains 2,074 structures of antibodies in complex with a protein antigen, after filtering for redundancy and experimental resolution <5 Å (Dreyer et al., 2023). We obtained structures of the corresponding variable fragment (Fv) domains (Figure 1A), numbered with the IMGT antibody numbering scheme (Lefranc et al., 2003), from SAbDab (Dunbar et al., 2014; Schneider et al., 2021). We modeled structures of the validation and test set using ABodyBuilder2 (Abanades et al., 2023) to evaluate AntiFold performance on solved and predicted structures. One and three structures were removed from the validation and test datasets, respectively, as these could not be modeled with ABodyBuilder2 (Abanades et al., 2023).

4.1.2 Predicted antibody structures from ABodyBuilder2

The structures of 148,832 paired antibody sequences from OAS (Kovaltsuk et al., 2018; Olsen et al., 2022) modeled using ABodyBuilder2 were released as part of ImmuneBuilder (Abanades et al., 2023). Filtering out structures with identical concatenated CDRs, as in AbMPNN (Dreyer et al., 2023), resulted in a dataset of 147,458 structures.

4.2 Fine-tuning strategy

We trained AntiFold by fine-tuning the ESM-IF1 inverse folding architecture (Hsu et al., 2022) (Figure 1B) on antibody structures. The inverse folding problem can be formalized as learning the conditional probability distribution, $p(Y|X)$, of the protein sequence, Y, consisting of amino acids $(y_1, \ldots, y_i, \ldots, y_n)$, given the structure, X, with spatial coordinates of the backbone atoms (N, C_α and C) $(x_1, \ldots, x_i, \ldots, x_{3n})$ (Equation (1)) (Hsu et al., 2022).

$$p(Y|X) = \prod_{i=1}^{n} p(y_i|y_{i-1}, \ldots, y_1; X) \tag{1}$$

The ESM-IF1 architecture consists of 4 GVP-GNN (Graph Neural Network Geometric Vector Perceptron) layers (Jing et al., 2021), 8 generic Transformer (Vaswani et al., 2017) encoder layers and 8 decoder layers (Hsu et al., 2022). The architecture is invariant to rotation and translation of the input coordinates.

The ESM-IF1 model is trained only on single chain structures. In order to represent complexes of antibody heavy and light chains, we concatenated the backbone coordinates with a 10 position padding of “gap” tokens, represented as missing coordinates in the input structure.

More details about the training approach, the fine-tuning parameters we evaluated (layer-wise learning rate decay, masking scheme and Gaussian noise added to modeled structure coordinates), and model evaluation are included in the Appendix (Supplementary Methods).

4.3 Model availability

AntiFold is available at `https://opig.stats.ox.ac.uk/data/downloads/AntiFold/`.

References

Abanades, B., Wong, W. K., Boyles, F., Georges, G., Bujotzek, A., and Deane, C. M. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Communications Biology*, 6(1):1–8, 2023. ISSN 23993642. doi: 10.1038/s42003-023-04927-7.

Anand, N., Eguchi, R., Mathews, I. I., Perez, C. P., Derry, A., Altman, R. B., and Huang, P. S. Protein sequence design with a learned potential. *Nature Communications*, 13(1):1–11, 2022. ISSN 20411723. doi: 10.1038/s41467-022-28313-9.

Carbery, A., Buttenschoen, M., Skyner, R., von Delft, F., and Deane, C. M. Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures. *bioRxiv*, 2023. doi: 10.1101/2023.09.07.556685.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615): 49–56, 2022. doi: 10.1126/science.add2187.

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., and Sillitoe, I. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, 45(D1):D289–D295, 11 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw1098.

Dreyer, F. A., Cutting, D., Schneider, C., Kenlay, H., and 1, C. M. D. Inverse folding for antibody sequence design using deep learning. The 2023 ICML Workshop on Computational Biology, 2023.

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., and Deane, C. M. SAbDab: The structural antibody database. *Nucleic Acids Research*, 42(D1):1140–1146, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1043.

Gao, Z., Tan, C., Chacón, P., and Li, S. Z. Pifold: Toward effective and efficient protein inverse folding. 2023.

Harvey, E. P., Shin, J.-E., Skiba, M. A., Nemeth, G. R., Hurley, J. D., Wellner, A., Shaw, A. Y., Miranda, V. G., Min, J. K., Liu, C. C., Marks, D. S., and Kruse, A. C. An in silico method to assess antibody fragment polyreactivity. *Nature Communications*, 13:7554, 2022. doi: 10.1038/s41467-022-35276-4.

Høie, M. H., Gade, F. S., Johansen, J. M., Würtzen, C., Winther, O., Nielsen, M., and Marcatili, P. Discotope-3.0 - improved b-cell epitope prediction using alphafold2 modeling and inverse folding latent representations. *bioRxiv*, 2023. doi: 10.1101/2023.02.05.527174.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022. doi: 10.1101/2022.04.10.487779.

Hummer, A. M., Abanades, B., and Deane, C. M. Advances in computational structure-based antibody design. *Current Opinion in Structural Biology*, 74:102379, 2022. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2022.102379.

Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., and Dror, R. Learning from protein structure with geometric vector perceptrons, 2021.

Kovaltsuk, A., Leem, J., Kelm, S., Snowden, J., Deane, C. M., and Krawczyk, K. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *The Journal of Immunology*, 201(8):2502–2509, 10 2018. ISSN 0022-1767. doi: 10.4049/jimmunol.1800708.

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., and Lefranc, G. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Developmental Comparative Immunology*, 27(1):55–77, 2003. ISSN 0145-305X. doi: https://doi.org/10.1016/S0145-305X(02)00039-3.

Lu, R. M., Hwang, Y. C., Liu, I. J., Lee, C. C., Tsai, H. Z., Li, H. J., and Wu, H. C. Development of therapeutic antibodies for the treatment of diseases. *Journal of Biomedical Science*, 27(1):1–30, 2020. ISSN 14230127. doi: 10.1186/s12929-019-0592-z.

Makowski, E. K., Kinnunen, P. C., Huang, J., Wu, L., Smith, M. D., Wang, T., Desai, A. A., Streu, C. N., Zhang, Y., Zupancic, J. M., Schardt, J. S., Linderman, J. J., and Tessier, P. M. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nature Communications*, 13(1), 2022. ISSN 20411723. doi: 10.1038/s41467-022-31457-3.

Makowski, E. K., Wang, T., Zupancic, J. M., Huang, J., Wu, L., Schardt, J. S., De Groot, A. S., Elkins, S. L., Martin, W. D., and Tessier, P. M. Optimization of therapeutic antibodies for reduced self-association and non-specific binding via interpretable machine learning. *Nature Biomedical Engineering*, 2023. doi: 10.1038/s41551-023-01074-6.

Marks, C., Hummer, A. M., Chin, M., and Deane, C. M. Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics*, 37(22):4041–4047, 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab434.

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

Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., and Bitton, D. A. 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.

Rabia, L. A., Desai, A. A., Jhajj, H. S., and Tessier, P. M. Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. *Biochemical Engineering Journal*, 137:365–374, 2018. ISSN 1369-703X. doi: https://doi.org/10.1016/j.bej.2018.06.003.

Schneider, C., Raybould, M. I. J., and Deane, C. M. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Research*, 50(D1):D1368–D1372, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1050.

Schrödinger, LLC. The PyMOL molecular graphics system, version. November 2015.

Stanfield, R. L. and Wilson, I. A. Antibody structure. *Microbiology Spectrum*, 2(2): 10.1128/microbiolspec.aid–0012–2013, 2014. doi: 10.1128/microbiolspec.aid-0012-2013.

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., and Kim, P. M. Fast and flexible protein design using deep graph neural networks. *Cell Systems*, 11(4):402–411.e4, 2020. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2020.08.016.

Sun, C., Qiu, X., Xu, Y., and Huang, X. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583, 2019.

Tennenhouse, A., Khmelnitsky, L., Khalaila, R., Yeshaya, N., Noronha, A., Lindzen, M., Makowski, E. K., Zaretsky, I., Sirkis, Y. F., Galon-Wolfenson, Y., Tessier, P. M., Abramson, J., Yarden, Y., Fass, D., and Fleishman, S. J. Computational optimization of antibody humanness and stability by systematic energy-based ranking. *Nature Biomedical Engineering*, 2023. ISSN 2157846X. doi: 10.1038/s41551-023-01079-1.

Tian Xia, H. W. S. H. Y. S. X. Y. J. H. J. L. S. G. J. D. Z. L., Shuaiyi Liang and Guo, Y. Structural basis for the neutralization and specificity of staphylococcal enterotoxin b against its mhc class ii binding site. *mAbs*, 6(1):119–129, 2014. doi: 10.4161/mabs.27106.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

5 Appendix

5.1 Supplementary Methods

5.1.1 Fine-tuning parameter evaluation

We evaluated the effect of the parameters described below on model performance, as applied to the validation dataset.

Layer-wise learning rate decay
We decayed the learning rate for each previous layer in the ESM-IF1 architecture by an alpha factor:

$$LR_i = LR \times \alpha^i \tag{2}$$

where i ranges from zero to the number of layers in the model (20), and alpha is set to 0.85.

Masking
We masked portions of the input antibody structure for model training and calculated loss over model predictions for the masked positions. The coordinates of masked positions were hidden for input to the model.

We evaluated three different masking schemes:

- Shotgun masking: randomly selected individual positions for masking
- Span masking: masked spans (consecutive stretches of positions) by randomly selecting starting positions and sampling the span length from a geometric distribution where p = 0.05, with a maximum span length of 30 positions
- Shotgun plus span masking: 7.5% of the structure was first masked using span masking and a further 7.5% was subsequently masked using the shotgun approach

Antibody sequence/structure can be separated into FR and CDR regions (Figure 1A), with the former being more conserved and typically easier to predict. As our model loss is calculated over masked positions, we explored whether performance could be improved by biasing the selection of masked positions towards CDR residues. There are more than 2.5 times as many FR as CDR positions in the sequence. For shotgun masking, we implemented a 3:1 weighting for the selection of CDR vs FR positions. For span masking, we biased selection to be low (weight = 1) for most FR positions, high (weight = 3) for most CDR positions, and medium (weight = 2) for FR positions immediately preceding CDRs as well as CDR positions immediately preceding FRs.

Gaussian noise
In the case of predicted structures, we add noise to the backbone (N, C_α and C) 3D-coordinates, sampled from a Gaussian distribution with a scale of 0.1 Å, following the approach taken in ESM-IF1 (Hsu et al., 2022).

5.1.2 Early stopping

Training of models was stopped when validation loss did not decrease after 10 epochs. The model with the lowest validation loss was carried forward.

5.1.3 Model performance evaluation

Amino acid recovery (AAR) is calculated as the percent of positions for which the top predicted amino acid is identical to the true amino acid.

Model output probabilities are given by:

$$logits = raw\ model\ outputs \tag{3}$$

$$probabilities(i) = \frac{e^{\text{logits}(i)}}{\sum_{j=1}^{20} e^{\text{logits}(j)}} \tag{4}$$

Perplexity for each position is given by:

$$perplexities = 2^{-\sum_{i=1}^{20} probabilities(i) \times log_2(probabilities(i))} \tag{5}$$

During sequence sampling, we sampled residues for each position in the CDRs proportional to their probability, using a temperature of 0.20. We used the same method as ProteinMPNN (Dauparas et al., 2022) of applying temperature directly to the logits before converting to probabilities:

$$scaled\, logits = \frac{logits}{t} \tag{6}$$

ProteinMPNN (Dauparas et al., 2022) and AbMPNN (Dreyer et al., 2023) were run with default settings and the flags –conditional_probs_only, –sampling_temp 0.20, –num_seq_per_target 20 and –seed 37. Sampled sequences were then predicted with ABodyBuilder2 (Abanades et al., 2023) at default settings. We corrected for ProteinMPNN reordered chains, reversal of insertions in IMGT positions 112 and invalid gaps.

We calculated RMSD using Pymol's rms_cur method (Schrödinger, LLC, 2015) between the solved and predicted backbone (N, C$_\alpha$, and C atoms) for each region, after aligning on the framework.

5.1.4 Bootstrapping

For bootstrapping, we resampled with replacement 1000 times, with the bootstrapped values used to calculate means and confidence intervals.

5.2 Appendix Tables and Figures

Table A1: Fine-tuning parameter evaluation, applied to validation dataset (experimental, “Exp”, structures). The training (layer-wise learning rate decay, train masking) and testing (test masking) parameters are indicated. The values in the right side of the table represent amino acid recovery for a particular IMGT-region (FR: framework, CDR: complementarity-determining region). The highest value is shown in bold, the second-highest in italics.

Exp/Pred	**Layer Decay**	**Train Masking**	**Test Masking**	**FR Avg.**	**CDR1H**	**CDR2H**	**CDR3H**	**CDR1L**	**CDR2L**	**CDR3L**
Exp	–	Shotgun	None	**0.845**	0.695	0.606	0.532	0.597	0.584	*0.609*
Exp	–	Span	None	0.814	0.635	0.506	0.364	0.521	0.516	0.505
Exp	–	Shotgun + Span	None	*0.842*	0.675	0.601	0.525	0.570	0.559	0.582
Exp	–	Shotgun – IMGT-Weighted	None	0.835	**0.708**	**0.640**	*0.543*	**0.613**	**0.628**	**0.626**
Exp	–	Span – IMGT-Weighted	None	0.807	0.636	0.511	0.365	0.535	0.521	0.516
Exp	–	Shotgun + Span – IMGT-Weighted	None	0.837	0.688	0.631	0.533	0.591	0.611	0.601
Exp	✓	Shotgun	None	*0.842*	**0.708**	0.620	*0.543*	0.601	0.567	*0.609*
Exp	✓	Span	None	0.803	0.621	0.500	0.364	0.513	0.492	0.487
Exp	✓	Shotgun + Span	None	0.838	0.684	0.609	0.538	0.587	0.577	0.596
Exp	✓	Shotgun – IMGT-Weighted	None	0.832	**0.708**	*0.636*	0.541	*0.611*	*0.614*	**0.626**
Exp	✓	Span – IMGT-Weighted	None	0.798	0.614	0.498	0.354	0.507	0.502	0.494
Exp	✓	Shotgun + Span – IMGT-Weighted	None	0.833	*0.699*	0.629	**0.544**	0.600	0.598	0.606
Exp	–	Shotgun	CDRs	**0.832**	0.520	0.388	0.310	0.439	0.438	0.437
Exp	–	Span	CDRs	0.811	*0.622*	0.507	0.348	0.521	*0.521*	0.485
Exp	–	Shotgun + Span	CDRs	**0.832**	0.587	0.477	*0.368*	0.506	0.484	0.485
Exp	–	Shotgun – IMGT-Weighted	CDRs	0.827	0.608	0.496	0.343	0.520	**0.545**	*0.499*
Exp	–	Span – IMGT-Weighted	CDRs	0.807	**0.623**	*0.512*	0.354	*0.532*	0.511	**0.509**
Exp	–	Shotgun + Span – IMGT-Weighted	CDRs	*0.828*	0.604	**0.532**	**0.380**	**0.541**	0.511	0.493
Exp	✓	Shotgun	CDRs	*0.828*	0.524	0.386	0.307	0.428	0.446	0.434
Exp	✓	Span	CDRs	0.800	0.599	0.483	0.330	0.494	0.467	0.470
Exp	✓	Shotgun + Span	CDRs	0.825	0.582	0.483	0.348	0.476	0.466	0.465
Exp	✓	Shotgun – IMGT-Weighted	CDRs	0.824	0.580	0.476	0.350	0.478	0.498	0.466
Exp	✓	Span – IMGT-Weighted	CDRs	0.795	0.606	0.508	0.343	0.490	0.479	0.485
Exp	✓	Shotgun + Span – IMGT-Weighted	CDRs	0.822	0.609	0.497	*0.368*	0.509	0.485	0.498

Table A2: Fine-tuning parameter evaluation, applied to validation dataset (predicted, "Pred", structures). The training (layer decay, train masking) and testing (test masking) parameters are indicated. The values in the right side of the table represent amino acid recovery for a particular IMGT-region (FR: framework, CDR: complementarity-determining region). The highest value is shown in bold, the second-highest in italics.

Exp/Pred	Layer Decay	Train Masking	Test Masking	FR Avg.	CDR1H	CDR2H	CDR3H	CDR1L	CDR2L	CDR3L
Pred	–	Shotgun	None	**0.856**	0.703	0.617	0.519	0.600	0.611	0.604
Pred	–	Span	None	0.816	0.639	0.505	0.373	0.531	0.506	0.499
Pred	–	Shotgun + Span	None	0.851	0.697	0.602	0.510	0.580	0.563	0.596
Pred	–	Shotgun – IMGT-Weighted	None	0.850	*0.708*	*0.640*	*0.520*	**0.636**	*0.625*	**0.635**
Pred	–	Span – IMGT-Weighted	None	0.810	0.643	0.506	0.377	0.545	0.519	0.516
Pred	–	Shotgun + Span – IMGT-Weighted	None	0.844	0.701	0.628	0.513	0.589	0.604	0.602
Pred	✓	Shotgun	None	*0.853*	**0.710**	0.626	*0.520*	0.585	0.597	0.603
Pred	✓	Span	None	0.808	0.618	0.487	0.361	0.503	0.464	0.481
Pred	✓	Shotgun + Span	None	0.848	0.693	0.615	0.507	0.587	0.585	0.593
Pred	✓	Shotgun – IMGT-Weighted	None	0.847	0.704	**0.645**	**0.526**	*0.620*	**0.632**	*0.624*
Pred	✓	Span – IMGT-Weighted	None	0.803	0.615	0.509	0.359	0.512	0.499	0.493
Pred	✓	Shotgun + Span – IMGT-Weighted	None	0.842	0.706	0.634	0.518	0.596	0.612	0.605
Pred	–	Shotgun	CDRs	**0.844**	0.535	0.395	0.327	0.438	0.444	0.444
Pred	–	Span	CDRs	0.814	0.618	0.501	0.351	0.517	0.508	0.486
Pred	–	Shotgun + Span	CDRs	0.840	0.603	0.481	0.374	0.517	0.473	0.478
Pred	–	Shotgun – IMGT-Weighted	CDRs	*0.841*	0.622	0.504	0.356	0.522	**0.534**	0.492
Pred	–	Span – IMGT-Weighted	CDRs	0.810	**0.630**	*0.512*	0.356	*0.536*	*0.529*	*0.499*
Pred	–	Shotgun + Span – IMGT-Weighted	CDRs	0.836	*0.627*	**0.536**	**0.394**	**0.537**	0.509	**0.502**
Pred	✓	Shotgun	CDRs	0.840	0.540	0.388	0.319	0.435	0.445	0.426
Pred	✓	Span	CDRs	0.805	0.600	0.488	0.341	0.498	0.481	0.464
Pred	✓	Shotgun + Span	CDRs	0.836	0.600	0.464	0.351	0.494	0.468	0.463
Pred	✓	Shotgun – IMGT-Weighted	CDRs	0.837	0.586	0.476	0.361	0.482	0.498	0.475
Pred	✓	Span – IMGT-Weighted	CDRs	0.800	0.610	0.503	0.344	0.493	0.496	0.487
Pred	✓	Shotgun + Span – IMGT-Weighted	CDRs	0.835	0.612	0.487	*0.377*	0.523	0.471	0.494

Table A3: Final model parameter evaluation, applied to validation dataset (experimental, “Exp”, and predicted, “Pred”, structures). Each model was trained with IMGT-weighted shotgun plus span masking for 1 epoch on the large predicted OAS structure dataset, followed by training on the experimental SAbDab dataset. The other training parameters (layer-wise learning rate decay and application of Gaussian noise to the predicted OAS structures) are indicated. The values in the right side of the table represent amino acid recovery for a particular IMGT-region (FR: framework, CDR: complementarity-determining region). The highest value is shown in bold, the second-highest in italics.

Exp/Pred	**Layer Decay**	**OAS Gaussian Noise**	**Test Masking**	**FR Avg.**	**CDR1H**	**CDR2H**	**CDR3H**	**CDR1L**	**CDR2L**	**CDR3L**
Exp	-	-	None	**0.898**	0.731	**0.712**	0.569	**0.723**	*0.736*	0.718
Exp	-	✓	None	**0.898**	*0.735*	0.698	0.566	0.716	0.702	0.713
Exp	✓	-	None	*0.895*	**0.741**	0.700	**0.584**	0.716	**0.741**	*0.725*
Exp	✓	✓	None	0.894	0.727	*0.702*	*0.573*	*0.720*	0.728	**0.727**
Exp	-	-	CDRs	**0.894**	0.680	0.637	*0.432*	*0.677*	*0.689*	**0.661**
Exp	-	✓	CDRs	**0.894**	**0.696**	0.651	**0.434**	**0.692**	0.680	*0.659*
Exp	✓	-	CDRs	0.890	0.675	**0.657**	0.431	0.666	*0.689*	0.658
Exp	✓	✓	CDRs	*0.891*	*0.681*	*0.653*	0.430	0.666	**0.698**	0.655
Pred	-	-	None	**0.909**	**0.753**	*0.716*	*0.561*	0.738	0.731	*0.722*
Pred	-	✓	None	0.905	0.749	0.704	0.558	0.729	0.725	*0.722*
Pred	✓	-	None	*0.907*	*0.750*	**0.730**	**0.572**	**0.746**	**0.737**	**0.730**
Pred	✓	✓	None	0.903	0.744	0.713	0.554	*0.744*	*0.733*	0.718
Pred	-	-	CDRs	**0.904**	*0.706*	0.650	**0.445**	*0.691*	*0.687*	**0.665**
Pred	-	✓	CDRs	0.901	**0.709**	**0.657**	*0.435*	**0.701**	**0.690**	*0.658*
Pred	✓	-	CDRs	*0.903*	0.695	*0.654*	*0.435*	0.675	0.675	0.654
Pred	✓	✓	CDRs	0.898	0.699	0.647	0.433	0.682	0.682	*0.658*

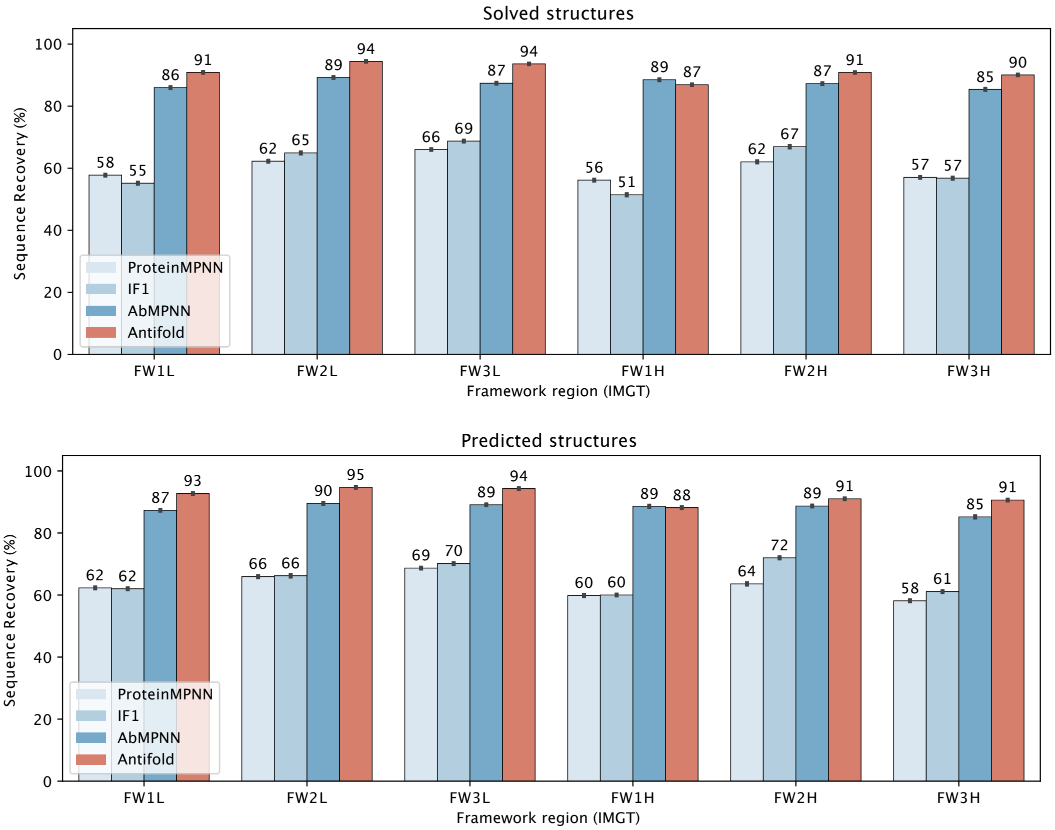

Figure A1: Framework (FW) amino acid sequence recovery, for solved (top) and predicted (bottom) structures in the test set.

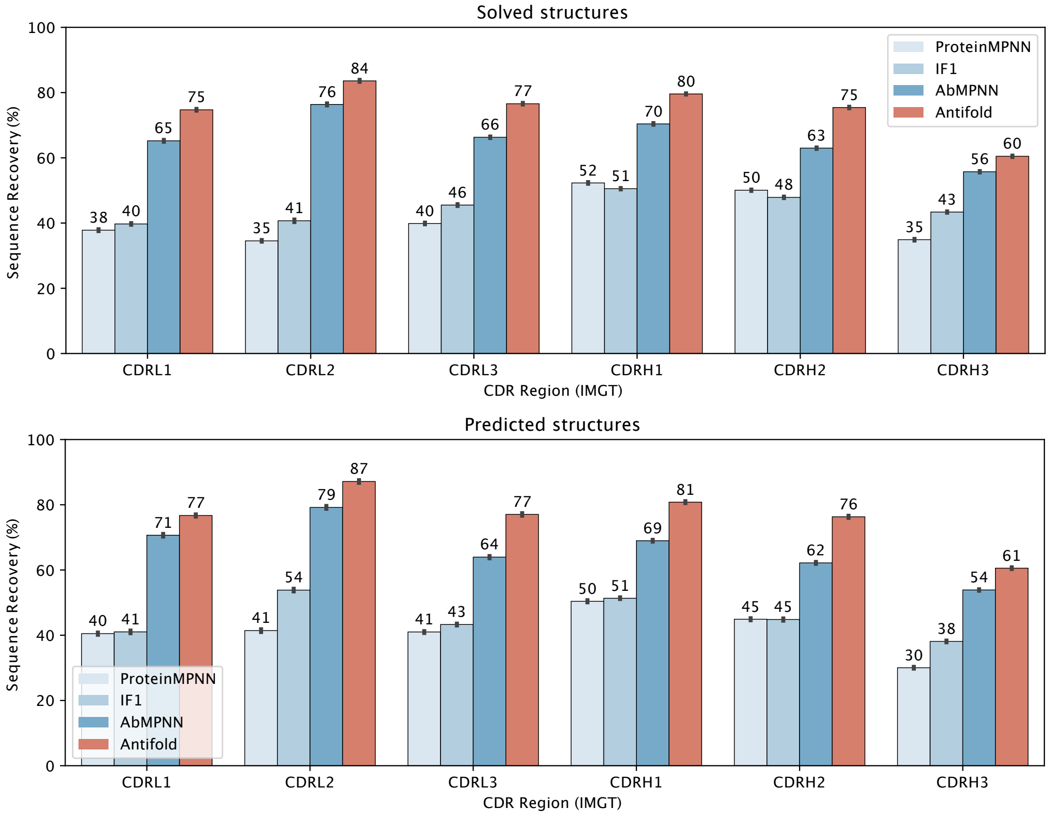

Figure A2: Complementarity determining region (CDR) amino acid sequence recovery for solved (top) and predicted (bottom) structures in the test set.

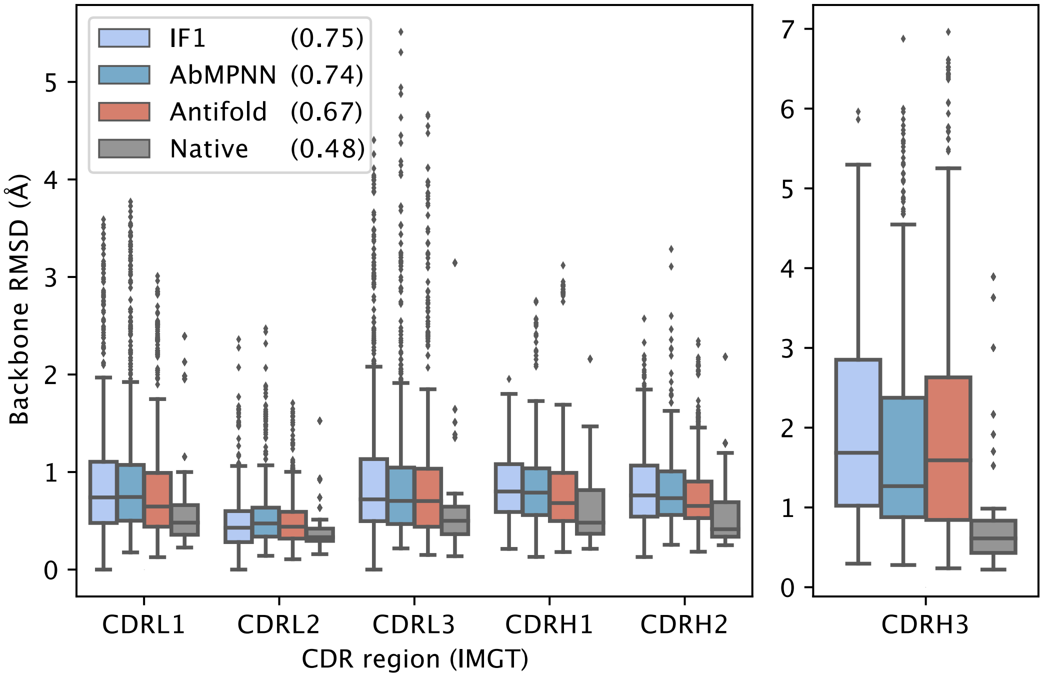

Figure A3: RMSD between ABodyBuilder2 predicted and experimental structure backbones, for sequences sampled with AntiFold and AbMPNN with a temperature of 0.20, and native sequences across IMGT regions (see Methods). Median CDR region RMSD values are shown in parentheses in the legend.