OpenFold3-preview2 Technical Report

The OpenFold3 Team

Contributing Institutions: Columbia University, Lawrence Livermore National Laboratory, Novo Nordisk, AWS, Seoul National University, Chan Zuckerberg Initiative, Absci, SandboxAQ, AMD, NVIDIA, University of Bristol, and OpenFold Consortium members.

Introduction

We describe OpenFold3-preview2 (OF3p2), the second prerelease version of the OpenFold3 system for general biomolecular structure prediction. OF3p2 substantially improves on its predecessor, OF3p, and is the most performant academic reproduction of AlphaFold3 (AF3) [1] to date. This second prerelease includes updated model weights, updated inference and training code, and all datasets needed for training and assessment, making OF3p2 the only functionally reproducible AF3 reproduction that is trainable from scratch and near AF3 parity. In this technical report, we provide new benchmarking results and a summary of improvements over OF3p. The goal of the OF3 project continues to be an open-source AF3 with performance parity across all molecular modalities.

Dataset Updates

Monomer Distillation Set

We generally followed the description provided by Google DeepMind for constructing the AF3 monomer self-distillation set. We used the clustered 2025_03 release of the Mgnify [2] database, selecting cluster representatives from clusters with three or more members. For approximately 5M sequences, we generated MSAs using the standard OF3 MSA pipeline which searches UniRef90, UniProt, and Mgnify sequence databases with JackHMMer [14] and an updated BFD database with HHblits [15]. For an additional set of 10M sequences, we generated sequences using the ColabFold MSA protocol with MMseqs2 [3]. To ensure high quality downstream structures we removed sequences with an N_{eff} <= 4.0, keeping 8M/10M sequences after this step. For all 13M sequences in this set, we predicted two structures with OpenFold [17], once using templates and once without (separate AF2 weights per structure, specifically “params_model_1_ptm” and “params_model_3_ptm”) and then selected the structure with the highest model confidence to include in the dataset.

RNA Distillation Set

Our procedure is based on the AF3 protocol for generating the RNA self-distillation set. Instead of RFAM v14.9 we used the current RFAM (v15.1) [5] and we filtered for cluster representative sequences from clusters containing three or more sequences. Structures were predicted using OF3p. After filtering for structures with an average PDE<2, approximately 125,000 predicted RNA structures remained.

Model Updates

Training

We adapted the method described in the AF3 SI for training OF3p2. We extended training relative to OF3p for a total of 155,000 training steps, with 131,500 steps for the initial stage, 8,000 steps for fine-tuning stage 1, and 15,500 steps for fine-tuning stage 2. Unlike AF3, we did not perform an additional fine-tuning stage 3 but instead trained the PAE confidence head from the start of initial training along with the rest of the model. As before, we terminated each training stage after an apparent maximum of the model selection metric was reached. We followed the hyperparameter configuration of the AF3 SI and trained OF3p2 on 256 NVIDIA H100 GPUs.

Bug fixes

We identified multiple model and data-related bugs in our previous version, OF3p. The template module provided an incorrect mask to the model by multiplying chain IDs instead of asserting identity. Our template pipeline was missing a filter that removes template structures with a high fraction of aligned unresolved residues. The plDDT computation in the confidence head incorrectly applied the 15Å inclusion radius, causing the set of evaluated atom pairs to vary with prediction quality. All these issues have been fixed in OF3p2.

Known errata

The version of the RNA distillation set used during training OF3p2 was generated using the rocBLASLt backend on MI300A APUs, which we retrospectively found to result in structures with decreased chemical validity. This can be corrected by using the rocBLAS backend (Nvidia GPUs are unaffected). The RNA self-distillation set we have made available uses corrected structures generated with rocBLAS.

Results

We evaluated performance on multiple types of biological molecules across several published benchmarks, and compared them against Protenix-v1 [6], Boltz-1 [8], Boltz-2 [10], Chai-1 [7], and AF3 [1], the last through an academic collaborator. We make all of our benchmarking code available here. We ran all OF3p2 assessments on MI300A APUs. The AMD implementation is based on a version of OF3p2 leveraging Triton kernels introduced in the MegaFold work, which we adapted for the OF3 inference stack [16]. For the other models, evaluations were performed on Nvidia-based GPUs.

Unless otherwise noted, we use the following inference procedure for all benchmarks: we subsample the input MSA to 1,024 sequences, provide structural templates as an input, run 10 trunk recycles, and run 5 MSA seeds each with 5 diffusion samples; this approach follows the inference procedure outlined in the AF3 SI. Additionally, as we observe sporadic failures during inference or downstream assessment in our own and other models, we report performances on a strict common subset of targets successfully across all models to enable fair comparisons.

Protein-small molecule complexes

We assessed OF3p2 on protein-small molecule complexes via the Runs N' Poses (RnP) benchmark [9] (Figures 1-2.). After inference, we evaluate predicted structures using OpenStructure following the RnP procedure. For Protenix-v1, we use the results reported in their github repo, and for all other models we obtain results from the RnP Zenodo page. For ranked evaluation, we select structures based on the average ipTM between a ligand and all chains in contact with it.

OF3p2 exhibits similar performance to Protenix v1.0 on most similarity bins (Figure 1) and continues to narrow the gap with AF3, which remains the best model measured by overall performance. We find that the ranked performance gap to AF3 is larger than the oracle performance gap, indicating that the underlying generator of OF3p2 is capable of producing high-quality samples, but can fail to select them in the confidence-based ranking step.

On a subset of Runs N' Poses adjusted to Boltz-2's training cutoff (Figure 2), we find OF3p2 to be closely matched to Boltz-2, despite Boltz2 having access to a larger corpus of training data. While oracle performance is overall similar between both models, we find that Boltz-2 retains a slight edge on higher-similarity bins on ranked performance, whereas OF3p2 seems ahead on the lower-similarity [0, 20] and [40, 50] SuCOS-pocket bins.

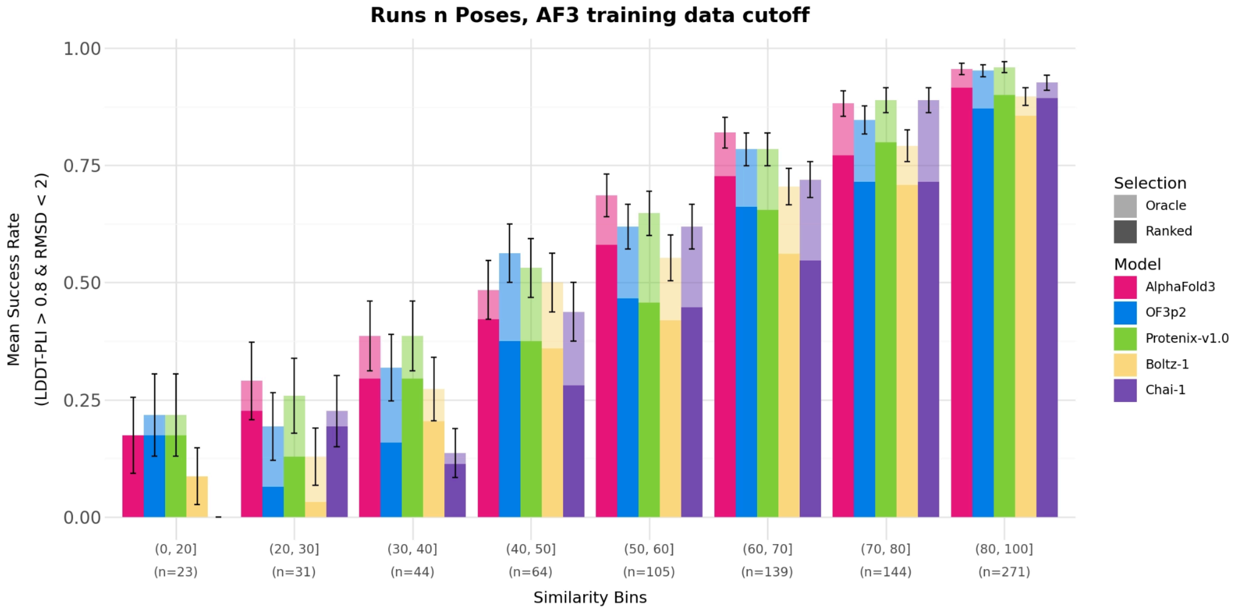

Figure 1. Performance of OF3p2 and other models on the Runs N' Poses set of protein-ligand complexes as a function of similarity (SuCOS-pocket) to training data, on structures held out with respect to the AF3 training date cutoff.

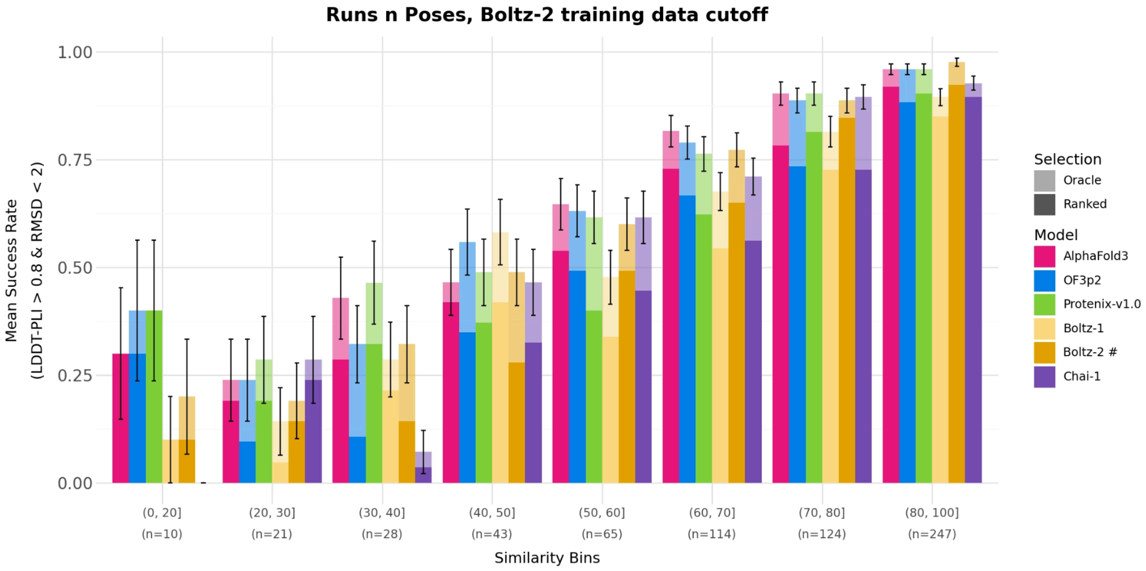

Figure 2. Performance of OF3p2 and other models on the Runs N' Poses set of protein-ligand complexes as a function of similarity (SuCOS-pocket) to training data. Here, we show results on a subset of RnP held out structures released after the Boltz-2 training cutoff which also do not change in their similarity scores between the two comparisons. "#" model used a newer training dataset than the standard AF3 cutoff date and is thus data-advantaged.

Biological polymers

We use the FoldBench benchmark [11] to evaluate performance for monomers (protein, DNA, RNA; Figure 3.) and multimers (protein-protein, protein-peptide, protein-DNA, protein-RNA; Figures 4.). We run OF3p2 using our standard inference procedure but use the sample ranking score to rank structures in order to remain faithful to the FoldBench protocol. We obtained results for AF3, Boltz-1, and Chai-1 from the FoldBench paper and for Protenix-v1 from the Protenix repository (downsampled from 20 to 5 MSA seeds). We ran Boltz-2 using its default inference procedure.

OF3p2 performs comparably to AF3 and Protenix-v1, slightly lagging behind on certain modalities (nucleic acid monomers, protein-protein interactions and protein-DNA interactions), while being on par or ahead on others (protein monomers, protein-RNA interactions), mostly within statistical variation.

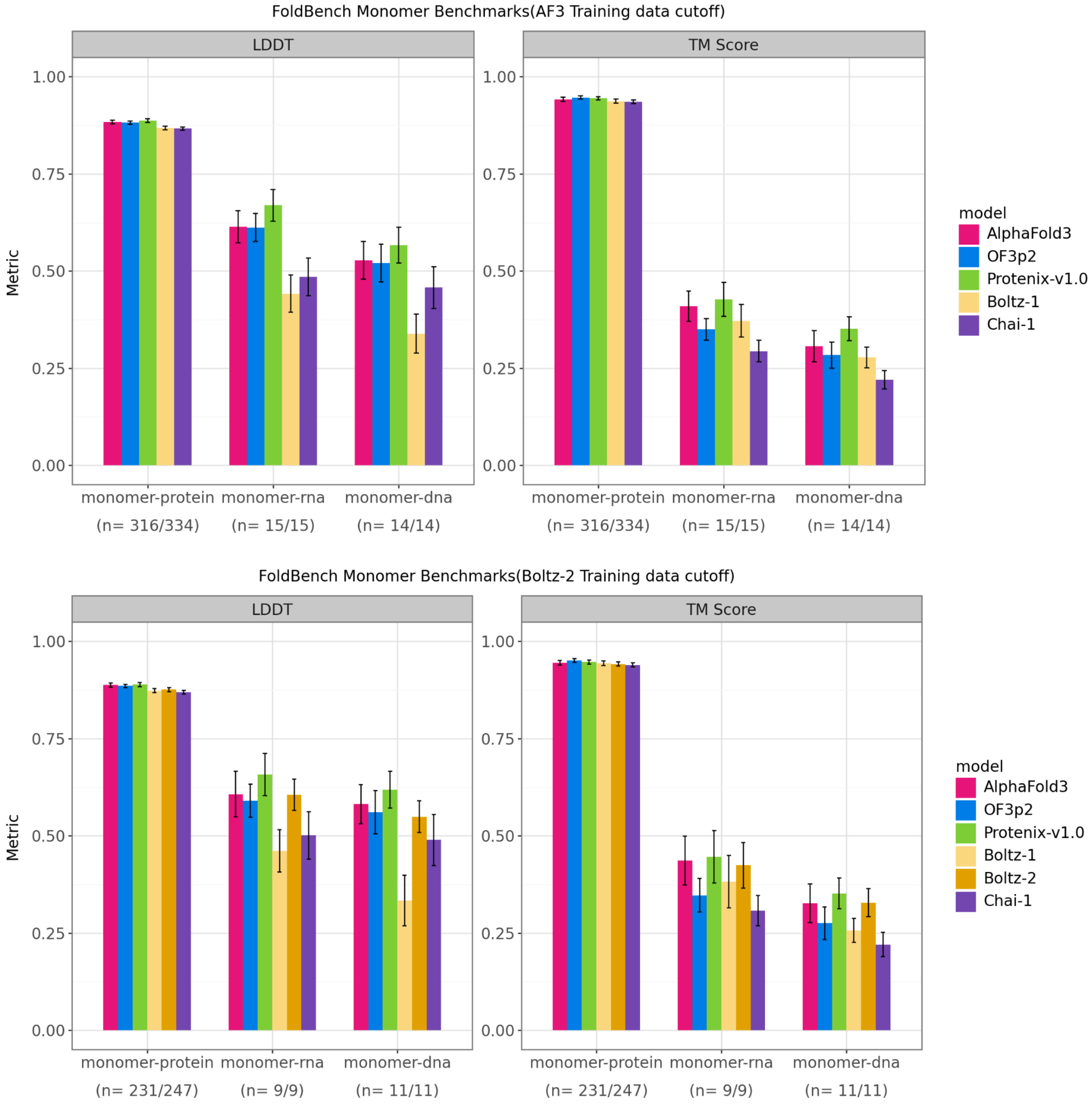

Figure 3. Model performances on monomeric modalities of the FoldBench benchmark on data held out with respect to the AF3 training cutoff (top) and the Boltz-2 training cutoff(bottom). Below each modality, we indicate the number of targets successfully predicted across all models out of all targets.

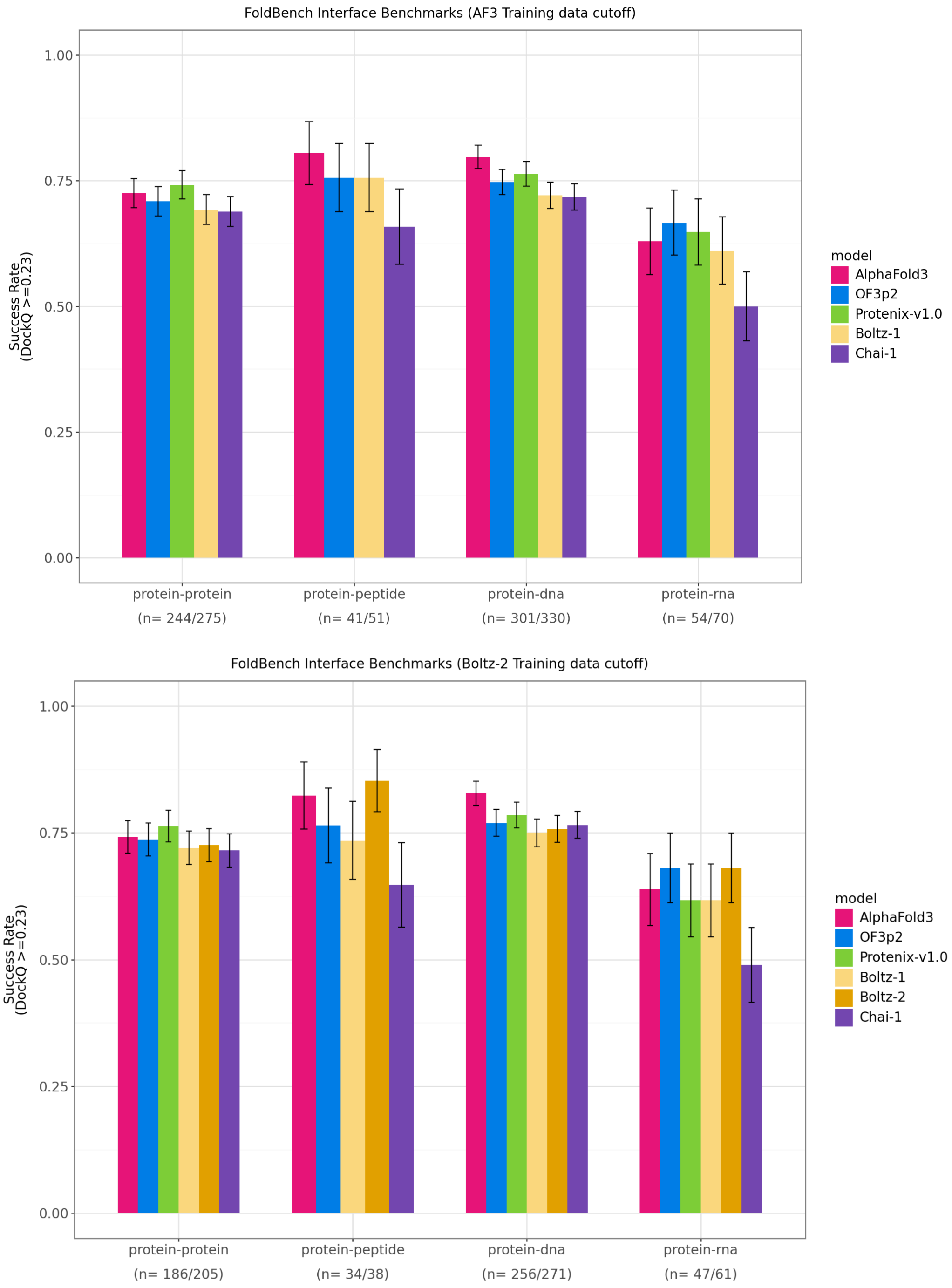

Figure 4. Model performances on multimeric modalities of the FoldBench benchmark on a held-out subset after the AF3 train cutoff (top) and the Boltz-2 training date cutoff (bottom). Below each modality, we indicate the number of targets successfully predicted across all models out of all targets. Protein-peptide performance is omitted for Protenix-v1, as it was not provided by the authors.

Antibody-antigen complexes

We used the test set described in the original AF3 publication [1] for assessing our performance on antibody-antigen complexes (Figures 5-6.). For each model evaluated (OF3p2, AF3, Protenix-v1, Boltz-1, and Chai-1), we generated inputs corresponding to the full bioassembly for each structure in the dataset using the corresponding model's default inference protocol. We ran 1,000 MSA seeds for OF3p2 and Protenix-v1, 600 seeds for Boltz-1, and 300 seeds for Chai-1. We obtained 1,000 MSA seed samples for AF3 from a collaborator. We used fewer than 1,000 seeds for models that appeared unlikely to close the gap at 1,000 seeds. To generate scaling curves, we follow the subsampling and bootstrapping procedure described in the AF3 paper.

OF3p2 exhibits notable inference time scaling behavior, albeit at a lower rate than AF3. We observe that OF3p2 shows stronger oracle scaling than Protenix-v1, whereas Protenix-v1 displays better ranked performance. Our results also indicate that while both OF3p2 and Protenix-v1 narrow the gap significantly between AF3 and its reproductions, they still fail to match exactly its ability to scale with higher MSA seed count.

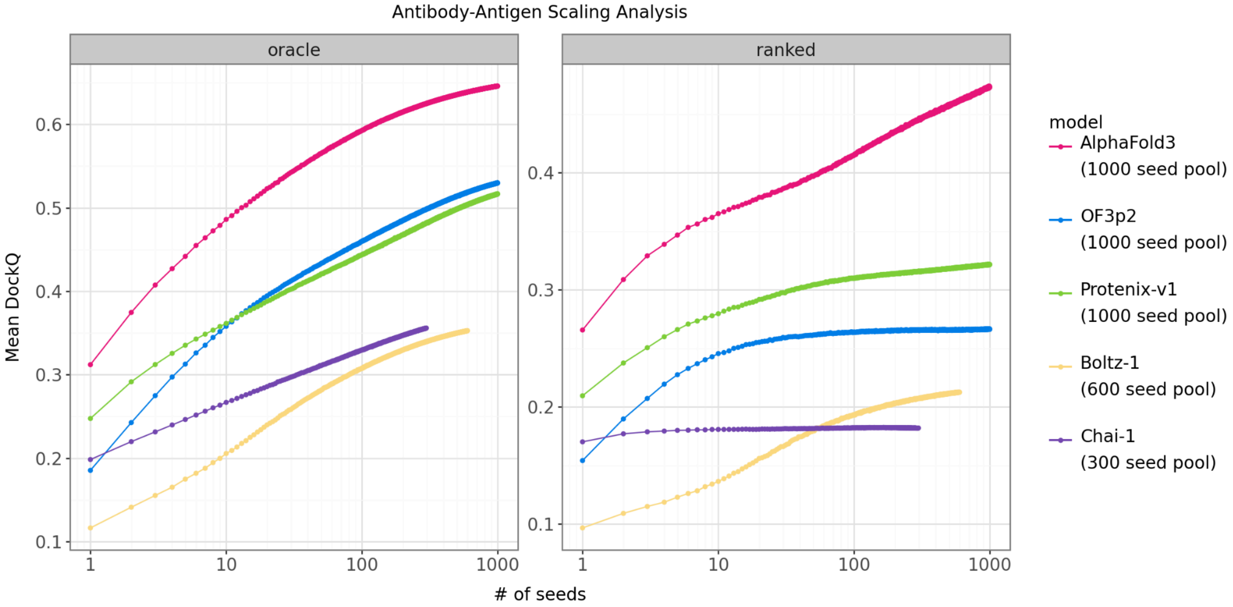

Figure 5. Mean DockQ as a function of MSA seed count on the Google DeepMind antibody-antigen test set.

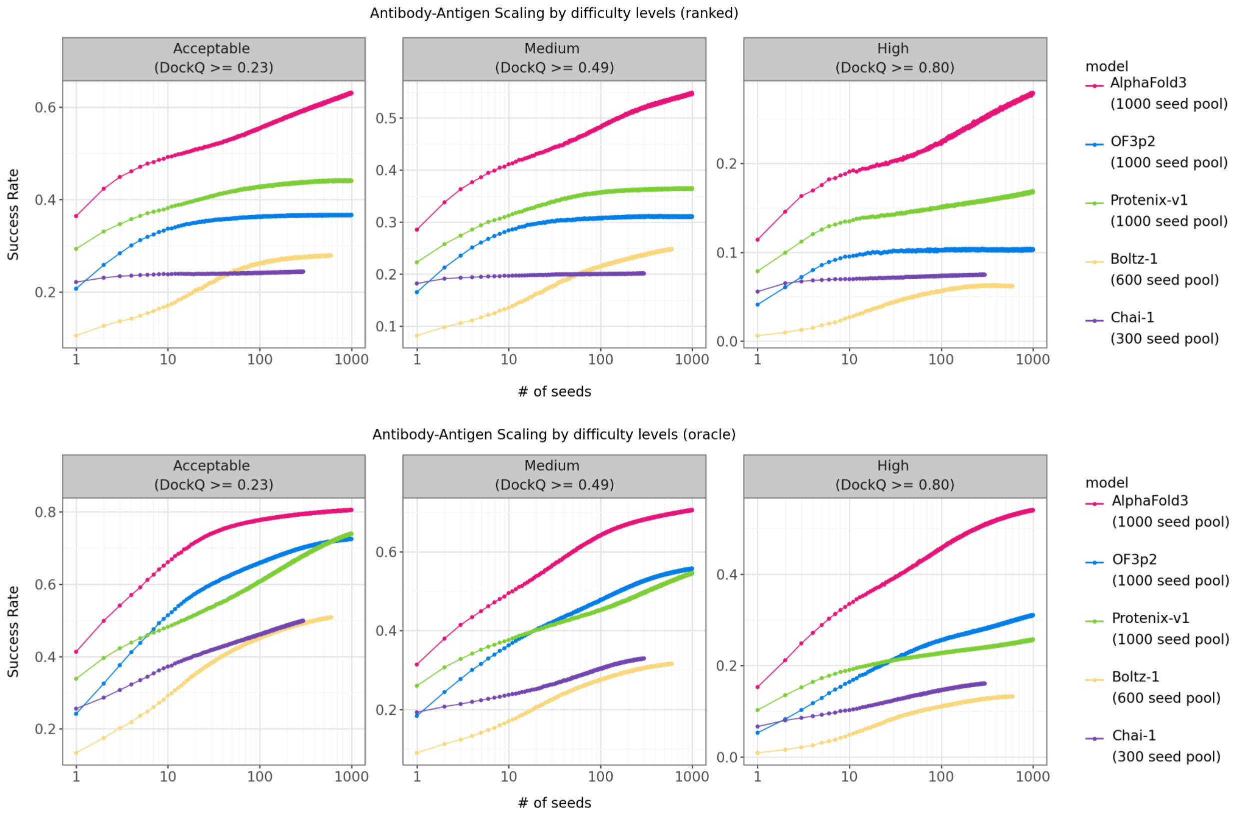

Figure 6. Success rate as a function of MSA seed count on the Google DeepMind antibody-antigen test set by DockQ category. We use the AF3 definition for success rate, the fraction of interfaces with DockQ higher than or equal to the threshold for the corresponding category.

Discussion

OF3p2 presents significant improvements over OF3p and generally outperforms other benchmarked AF3 reproductions except the recently released Protenix-v1. Throughout our benchmarks, we observed that OF3p and Protenix-v1 match or outperform Boltz-2 despite its later training cutoff date. We believe that OF3p2 may benefit even further from inclusion of more training data, and reserve this as a direction for future versions. Despite recent advances across the field, we observe that no reproduction has fully matched AF3's performance on all modalities. We view OF3p2 as a meaningful step toward this goal. By publicly releasing our model weights, training and inference code, and complete distillation datasets, we aim to accelerate both our own continued development and broader community efforts toward full parity with AF3 and beyond.

Acknowledgement

We would like to acknowledge the AlphaFold3 team for helpful discussions and feedback.

References

[1] Abramson, Josh et al. “Accurate structure prediction of biomolecular interactions with AlphaFold 3.” Nature vol. 630,8016 (2024): 493-500. doi:10.1038/s41586-024-07487-w

[2] Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M. L., Burdett, T., ... & Finn, R. D. (2023). MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic acids research*, *51*(D1), D753-D759.

[3] Mirdita M., Schütze K., Moriwaki Y., Heo L., Ovchinnikov S., Steinegger M. ColabFold: Making protein folding accessible to all Nature Methods, doi: 10.1038/s41592-022-01488-1.

[4] Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

[5] Griffiths-Jones, Sam et al. “Rfam: an RNA family database.” Nucleic acids research vol. 31,1 (2003): 439-41. doi:10.1093/nar/gkg006

[6] Chen, Xinshi, et al. “Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction.” bioRxiv, 8 Jan. 2025, doi:10.1101/2025.01.08.631967.

[7] Boitreaud, J., et al. . “Chai-1: Decoding the Molecular Interactions of Life.” bioRxiv, preprint v2, 15 Oct. 2024, doi:10.1101/2024.10.10.615955.

[8] Wohlwend, Jeremy et al. “Boltz-1 Democratizing Biomolecular Interaction Modeling.” bioRxiv : the preprint server for biology 2024.11.19.624167. 6 May. 2025, doi:10.1101/2024.11.19.624167. Preprint.

[9] Škrinjar, Peter, et al. “Have Protein-Ligand Co-Folding Methods Moved Beyond Memorisation?” bioRxiv, 7 Feb. 2025, doi:10.1101/2025.02.03.636309.

[10] Passaro, Saro, et al. “Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction.” bioRxiv, preprint, 18 June 2025, doi:10.1101/2025.06.14.659707.

[11] Xu, Sheng, et al. “FoldBench: An All-Atom Benchmark for Biomolecular Structure Prediction.” bioRxiv, version 1, 22 May 2025, doi:10.1101/2025.05.22.655600.

[13] Ludaic, Marko, and Arne Elofsson. “Limits of Deep-Learning-Based RNA Prediction Methods.” bioRxiv, version 1, 5 May 2025, doi:10.1101/2025.04.30.651414.

[14] Johnson, L. Steven, Sean R. Eddy, and Elon Portugaly. "Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure." BMC Bioinformatics, vol. 11, no. 1, 2010, pp. 1–8, doi:10.1186/1471-2105-11-431.

[15] Remmert, Michael, Andreas Biegert, Andreas Hauser, and Johannes Söding. "HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment." Nature Methods, vol. 9, no. 2, 2012, pp. 173–175, doi:10.1038/nmeth.1818.

[16] La, Hoa, Ahan Gupta, Alex Morehead, Jianlin Cheng, and Minjia Zhang. "MegaFold: System-Level Optimizations for Accelerating Protein Structure Prediction Models." arXiv, 24 Jun. 2025, doi:10.48550/arXiv.2506.20686.

[17] Ahdritz, Gustaf, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, et al. "OpenFold: Retraining AlphaFold2 Yields New Insights into Its Learning Mechanisms and Capacity for Generalization." Nature Methods, vol. 21, 2024, pp. 1514–1524, doi:10.1038/s41592-024-02272-z.