OpenFold3-preview technical white paper

The OpenFold3 Team

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

Introduction

We describe OpenFold3-preview (OF3p), a prerelease version of the OF3 system for general biomolecular structure prediction. OF3 aims to be an open-source, bitwise reproduction of AlphaFold3 (AF3) with performance parity across all molecular modalities. This white paper provides an overview of OF3p, including preliminary benchmarking and methodological differences from the AF3 Supplementary Information (SI). Our OF3p prerelease includes model weights, inference code, and (unsupported) training code.

Methods

OF3p was developed by following, to the best of our understanding, all details of the AF3 paper and SI [1], including model architecture, training datasets, and training regimens. We deviated in a few instances, as described below, to correct for unambiguous bugs or to use newer and more up-to-date databases.

Model differences. As noted by other reproductions, stable training of AF3-like models requires changes to the algorithms outlined in the SI [2-4]. We added architectural modifications matching what has been previously reported in Chen et al. [2; Table 2], with the exception of the binning described in Algorithm 31, Line 3 of the AF3 SI, which we change to 38 bins of equal width ranging from 3.25Å to 50.75Å, with an extra bin for larger distances. We add `LayerNorm` functions around inputs of the final linear projection of each confidence head, and remove `LayerNorm` biases throughout the diffusion module. Furthermore, we skip the MSA stack in the last MSA module block, as it does not influence the returned pair representation.

Training differences. We followed the multi-stage training procedure described in the AF3 SI. The precise number of training steps for each training stage were not disclosed. However, we followed the criteria described for model selection, terminating after an apparent maximum of the model selection metric was reached during each training stage. This resulted in 74,250 steps for initial training, 1,750 steps for fine-tuning 1, 250 steps for fine-tuning 2, and 1,750 steps for fine-tuning 3.

As we did not have a pretrained OF3 model, self-distillation sets for nucleic acids were not used. All other datasets were constructed as described in the AF3 SI, except for using newer protein sequence databases for MSA construction. Date cutoffs for structural training sets were the same as AF3.

Known errata. The template module in OF3p contains a bug when constructing the inter-chain mask. In Algorithm 16, Line 5 of the AF3 SI, our implementation provides an incorrect mask to the model by multiplying chain IDs instead of asserting identity.

Results

We assessed OF3p on the following molecular modalities: protein monomers and multimers (including antibody-antigen complexes), RNA monomers, and protein-ligand complexes. To predict structures for evaluation, we run five different (MSA) seeds and generate five (diffusion) samples per seed. For all molecular modalities except antibodies, we subsample the input MSA to 1,024 sequences and provide structural templates as an input; this approach follows the procedure outlined in the AF3 SI. For antibodies, we keep the paired MSA and only subsample the unpaired MSA.

We report both oracle (best predicted structure as assessed with respect to the known experimental structure) and ranking-based performance. For protein-ligand and protein-protein interfaces, we rank structures based on chain pair ipTM, an interface variant of the predicted TM (pTM) score that is computed exclusively using the inter-chain components of the two interacting chains. For antibody-antigen complexes, we use the bespoke ipTM, which averages the chain pair ipTM of each of the two chains constituting an interface with all other chains, as defined in the AF3 SI. For protein and RNA monomers, we use the sample ranking metric defined in the AF3 SI. For baselines, we consider AF3 [1], Protenix [2], Chai [4], and Boltz-1 [3]. We did not include Boltz-2 [5] as it uses a newer date cutoff for its training set than the other methods, making a direct comparison incommensurate. For all comparisons to AF3, we ran AF3 without templates unless stated otherwise.

We found OF3p to be close to the best current AF3 reproduction for any given modality, and at parity with AF3 on RNA. However, AF3 remains substantially more performant than any reproduction on multiple modalities, including antibody-antigen and protein-ligand complexes, particularly ones with highly novel pockets.

Proteins. For monomer assessment, we evaluated OF3p on one protein set, derived from all publicly released monomeric protein structures in the CASP16 [6] experiment**.** CASP16 occurred after the AF3 (and therefore OF3p) training set date cutoff and thus serves as an appropriate test set. OF3p achieves slightly worse performance than AF3 but is comparable with other reproductions (Fig. 1A).

For multimeric assessment, we considered both general protein-protein complexes and antibody-antigen complexes. For the former, we used a CASP16-based set and one based on FoldBench [7]. For the latter, we used the antibody-antigen test set from the AF3 paper. On general protein-protein complexes, AF3 is slightly better than other reproductions while OF3p is comparable (Fig. 1B). On antibody-antigen complexes, AF3 maintains a considerable lead (Fig. 1C).

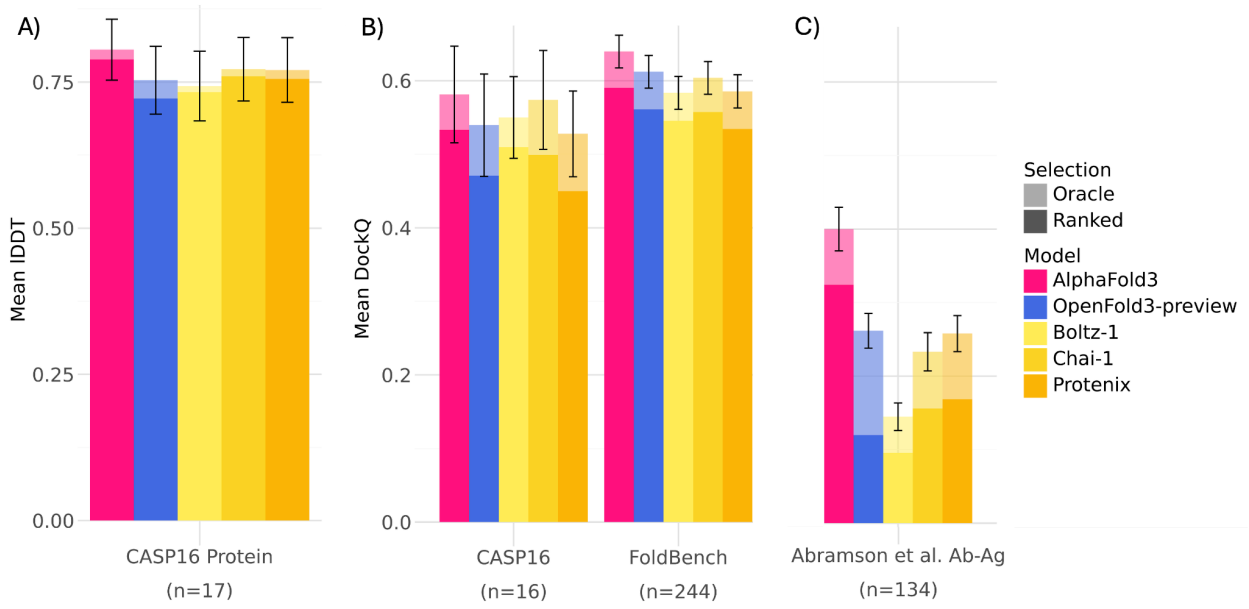

Figure 1. Performance of OF3p and other models on protein monomers and complexes as summarized by bar plots of (A) lDDTs of predictions of the CASP16 protein monomer set and (B) DockQ scores of predictions of protein-protein complexes sourced from CASP16, FoldBench [7], and the antibody-antigen test set of AF3 (Abramson et al.). For FoldBench and Abramson et al., DockQ scores are computed at the interface level, whereas for CASP16, results are aggregated over all interfaces in a given structure. For antibody-antigen complexes, we predicted 50 seeds for each structure and emulated the 5 seed performance by randomly subsampling a set of 5 seeds.

RNA. We evaluated OF3p on two RNA test sets: “CASP16 RNA”, derived from all publicly released monomeric RNA structures in the CASP16 experiment, and “Ludaic & Elofsson RNA”, derived from RNA monomers described in [8]. On both test sets, OF3p achieves performance comparable to AF3 and outperforms Protenix, Chai, and Boltz-1 (Fig. 2A). To illustrate these performance gains, we include predictions for CASP16 target R1241, a group II intron (Fig. 2B-C).

Both OF3p and AF3 use RNA multiple sequence alignments (MSAs) during structure prediction – a feature absent in other reproductions – which in some instances leads to large gains in accuracy (Fig. 3A). In the most dramatic case of target 8TJU (Fig. 3B-C), incorporating RNA MSAs improves the predicted model by ~23Å RMSD.

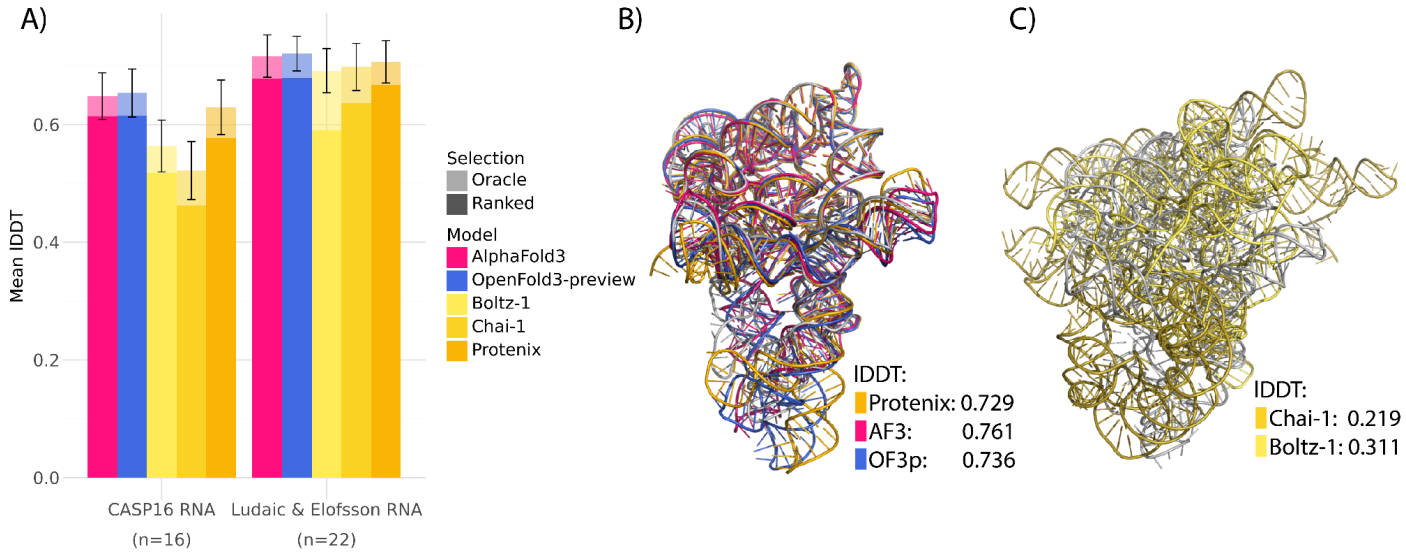

Figure 2.Performance of OF3p and other models on RNA. A) Bar plots of lDDTs of predictions of the CASP16 RNA monomer set [6] (left) and the Ludaic & Elofsson RNA set [8] (right). B-C) Examples of predicted structures (bright colors) aligned to the experimental structure (grey) for CASP16 entry R1241, a group II intron.

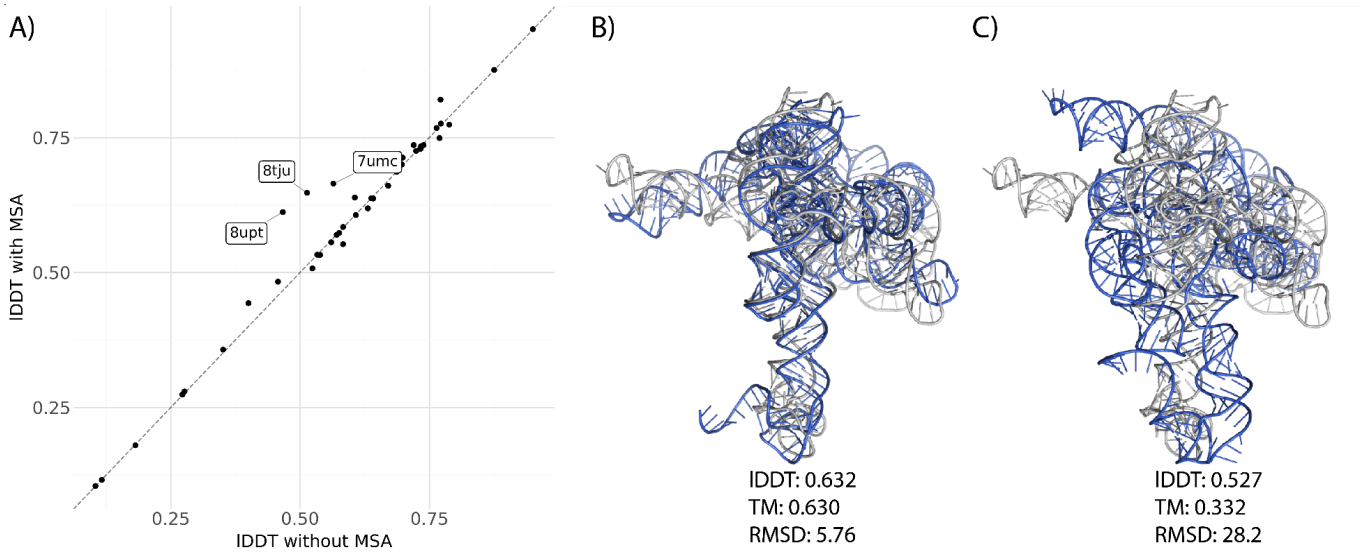

Figure 3. Impact of RNA MSAs on RNA structure predictions. A) Scatterplot of lDDTs of individual predictions of the Ludaic & Eloffson RNA set with and without RNA MSAs. Target 8TJU shows considerable improvement when using an RNA MSA (B) versus when not (C).

Protein-ligand complexes. For assessing protein-small molecule performance, we evaluated OF3p on the Runs N' Poses test set [9]. Runs N' Poses stratifies complexes based on binding pocket similarity to the training set, determined by the product of binding pocket coverage of the protein with Combined Overlap Score (SuCOS) of the ligand (Fig. 4). We find that AF3

consistently outperforms all reproductions tested, with considerable margins on the most difficult similarity bins. OF3p oracle performance is generally the best among reproductions, while ranked performance is more middle-of-the-pack.

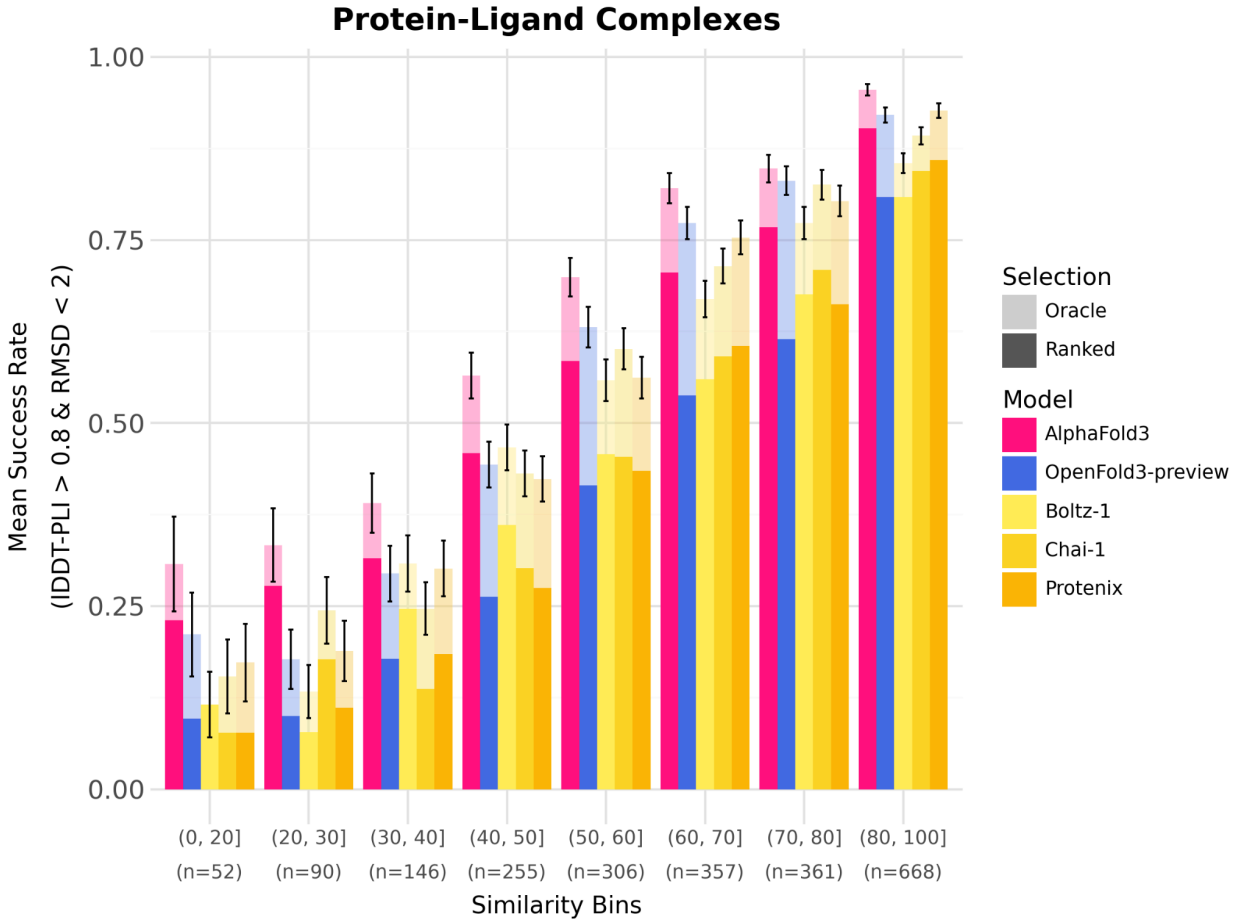

Figure 4. Performance of OF3p and other models on the Runs N' Poses set of protein-ligand complexes, as summarized by bar plots of the success rates of predicting complexes with less than 2Å RMSD and greater than 0.8 lDDT-PLI. Predictions are stratified by the same similarity bins to the training set defined in [9]. In this evaluation, we make use of structural templates for AF3 predictions.

Limitations and Next Steps

OF3p does not currently match AF3 performance on all modalities, however, early benchmarking indicates that OF3p is among the most performant AF3-based reproductions. One particular current limitation is ranking performance, which substantially trails AF3 relative to oracle performance. Our team is currently working on a finalized, AF3-bitwise-compatible (modulo literal model weights parity) OF3 that reaches the full performance of AF3.

This preview release lacks finalized documentation for training and does not include the training dataset, which will be released with the final version of OF3. We anticipate that further

performance gains can be had from training on newer versions of the PDB and plan to release new checkpoints with later training set cutoff dates.

While AF3 parity is our immediate objective, as AF3 remains the state-of-the-art system for generalized biomolecular modeling, we anticipate that modification of the AF3 architecture and training procedure can yield substantial performance gains. Following release of the full OF3 model, our strategic priorities will be improved generalization performance across underperforming modalities (lDDT < 0.8), faster inference, new capabilities beyond structure prediction, and improved tooling for training and inference.

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

[3] 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.

[4] 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.

[5] 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.

[6] Zhang, J., et al. “Assessment of Protein Complex Predictions in CASP16.” *bioRxiv*, version v1, 29 May 2025, doi:10.1101/2025.05.29.656875.

[7] 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.

[8] 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.

[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.