Technical Report of HelixFold3 for Biomolecular Structure Prediction

PaddleHelix Team
Baidu Inc.
Shenzhen, China
Official website: https://paddlehelix.baidu.com/
GitHub: https://github.com/PaddlePaddle/PaddleHelix

August 14, 2024

ABSTRACT

The AlphaFold series has transformed protein structure prediction with remarkable accuracy, often matching experimental methods. AlphaFold2, AlphaFold-Multimer, and the latest AlphaFold3 represent significant strides in predicting single protein chains, protein complexes, and biomolecular structures. While AlphaFold2 and AlphaFold-Multimer are open-sourced, facilitating rapid and reliable predictions, AlphaFold3 remains partially accessible through a limited online server and has not been open-sourced, restricting further development.

To address these challenges, the PaddleHelix team is developing HelixFold3, aiming to replicate AlphaFold3's capabilities. Leveraging insights from previous models and extensive datasets, HelixFold3 achieves accuracy comparable to AlphaFold3 in predicting ligands, nucleic acids, and proteins. The initial release of HelixFold3 is available as open source on GitHub for academic research, promising to advance biomolecular research and accelerate discoveries.

1 Introduction

AlphaFold series [1, 2, 3] revolutionizes protein structure prediction with unprecedented accuracy, often rivaling experimental methods. AlphaFold2 [1], AlphaFold-Multimer [2], and AlphaFold3 [3], achieve breakthrough progress in the prediction of single protein chains, protein complexes, and biomolecular structures. Both AlphaFold2 and AlphaFold-Multimer have fully open-sourced their codes, significantly accelerating protein-related research by providing rapid and reliable predictions. These tools not only enhance our understanding of protein functions and interactions, but also exemplify the transformative potential of artificial intelligence in solving complex scientific challenges.

AlphaFold3, the latest in the series, supports biomolecular interaction predictions and offers an online server [1] for limited structural prediction services. This server allows researchers to utilize its advanced capabilities, although it cannot support arbitrary biomolecular structure predictions and imposes a daily limit on the number of predictions. Furthermore, AlphaFold3 has not yet been open-sourced, limiting its accessibility for widespread use and further development by the scientific community.

Replicating AlphaFold3's capabilities presents significant opportunities for advancing the life sciences but involves substantial challenges due to the model's complexity, data requirements, and the extensive computational resources needed for training.

The PaddleHelix team is working on HelixFold3 with the objective of replicating the advanced capabilities of AlphaFold3. Our approach is informed by insights from the AlphaFold3 paper and builds on our prior work with HelixFold [4], HelixFold-Single [5], HelixFold-Multimer [6], and HelixDock [7]. For training, we utilized targets from the Protein Data Bank (PDB) [8] released before September 30, 2021, along with self-distillation datasets. Currently, HelixFold3's accuracy in predicting the structures of small molecule ligands, nucleic acids (including DNA and RNA), and proteins is comparable to that of AlphaFold3. We are committed to continuously enhancing the model's performance and rigorously evaluating it across a broader range of biological molecules.

[1] https://alphafoldserver.com/

The initial release of HelixFold3, which includes the inference code and the current version of model parameters, is now available as open source on PaddleHelix's GitHub repository. You can access it at `https://github.com/PaddlePaddle/PaddleHelix/blob/dev/apps/protein_folding/helixfold3` for academic research. This release is intended for non-commercial use and provides researchers with the tools needed to explore and leverage HelixFold3's advanced capabilities in protein structure prediction. We believe that the open-source availability of HelixFold3 will significantly contribute to the advancement of research in biomolecular interactions. Researchers can now build upon this foundation, conduct further studies, and apply HelixFold3 to a broader range of biological questions, thereby advancing our understanding of complex biomolecular systems and accelerating the development of new applications in structural biology and related areas.

2 Results

We begin by evaluating the performance of HelixFold3 in multiple datasets. For ligands, we utilize the PoseBusters benchmark [9] to assess precision and physical plausibility. For nucleic acids, HelixFold3 is evaluated on CASP15 [10] RNA targets and recent RNA and DNA structures from the RCSB Protein Data Bank (RCSB PDB) [8]. We also assessed the model's precision for protein structure predictions on CASP15 protein targets. Each sample is processed with 5 different random seeds, and diffusion inference is performed 5 times per seed with 200 sampling steps. The most accurate predictions are selected based on their confidence scores.

In addition, we investigate the effectiveness of confidence metrics in evaluating prediction quality. This involves analyzing how well the confidence scores correlate with the actual accuracy of predictions across various datasets. We also explore how different factors impact prediction quality, including the number of random seeds, diffusion inference iterations, and the number of sampling steps in the diffusion process. This comprehensive analysis aims to refine our understanding of how these parameters influence the reliability of the prediction.

2.1 Ligands

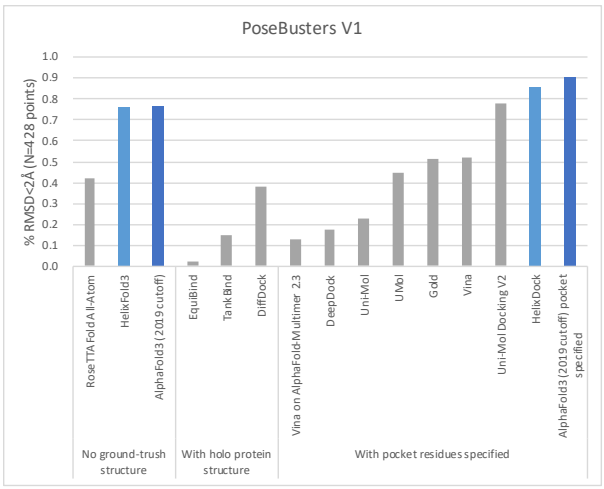
We first compare HelixFold3 and the baseline methods in PoseBusters [9] to evaluate the quality of the protein-ligand structure predictions. The PoseBusters dataset, a benchmark for ligand docking algorithms, initially had 428 structures (PoseBusters V1). After excluding data points with ligands within 5.0Å of multiple biological units, it was refined to 308 structures (PoseBusters V2). The baseline methods can be classified into three groups: methods with no ground-truth protein structure specific, methods with holo protein structure specified, and methods with pocket residues specified. The comparison of the success rate for PoseBuseters V1 and PoseBuseters V2 is shown in Figure 1a and Figure 1b. Even though no ground-truth protein structure is specified, HelixFold3 achieves a high success rate, surpassing methods that rely on given homo protein structures. Its prediction accuracy is comparable to that of AlphaFold3, highlighting its strong performance in predicting protein-ligand interactions. Due to the overlap between the training data for HelixFold3 and PoseBusters, there is a risk of overestimating HelixFold3's performance. To address this, we evaluated the average success rate on 290 samples that do not overlap with the training data. This analysis showed only a 2% reduction in performance, suggesting that HelixFold3 maintains strong accuracy in predicting ligand-protein interface structures.

To evaluate the stereochemistry and physical plausibility of the predicted ligand structures, including intra- and intermolecular measurements, we used the PoseBusters test suite [9]. As shown in Figure 1c, HelixFold3, AlphaFold3, and HelixDock all achieve pass rates above 90% for nearly all metrics, with the exception of tetrahedral chirality.

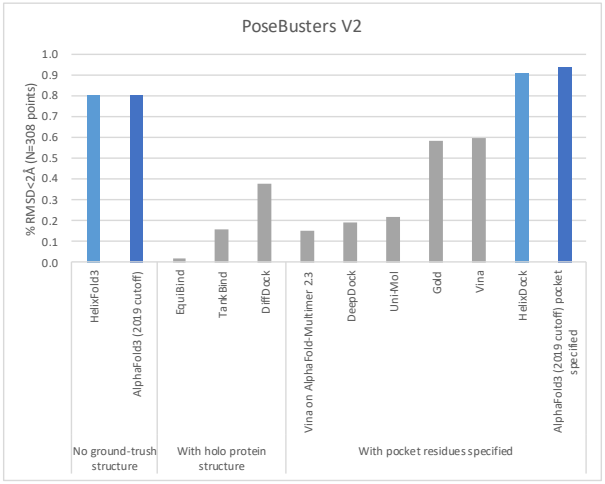
2.2 Nucleic Acids

Accurately predicting the structure of nucleic acid targets in a fully automated manner, without human input, remains a formidable challenge, primarily due to the limited availability of crystallized nucleic acid structures. To evaluate HelixFold3's capabilities in this domain, we conducted a comparative analysis against several baseline methods, starting with RNA targets from CASP15 [10], using the evaluation framework established by AlphaFold3. Figure 2a illustrate a comparison between HelixFold3 and the baseline methods, showcasing the average RNA LDDT across 12 targets as well as the RNA LDDT for each individual target. While HelixFold3's accuracy on CASP15 RNA samples does not yet match that of AIchemy_RNA2 [11], which benefits from human intervention, it achieves a level of accuracy on par with AlphaFold3 among models that operate in a fully automated manner.

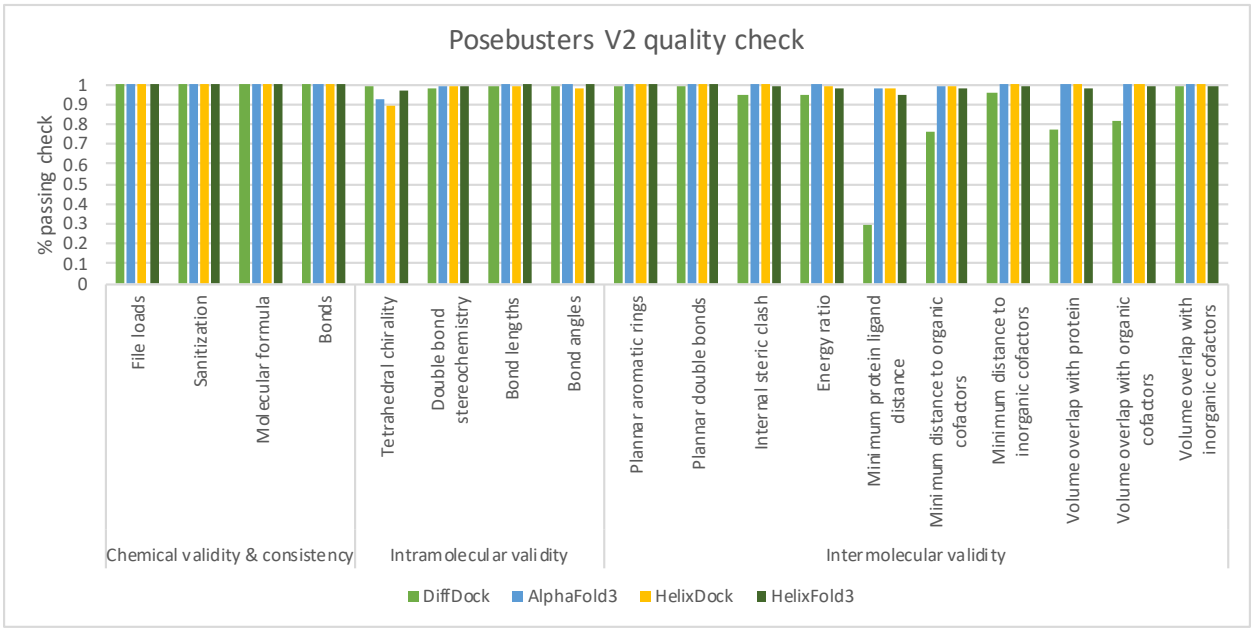
Given the limited number of RNA targets in CASP15, we further expanded our evaluation by collecting 41 RNA-only and 41 DNA-only complexes released between May 1, 2022, and June 30, 2024, from the Protein Data Bank (PDB) to more thoroughly assess HelixFold3's performance in nucleic acid structure prediction. The results, depicted in Figure 2b, demonstrate that HelixFold3 significantly outperforms RoseTTAFold2NA [12], a model specifically designed



(a) Comparison of the success rate of HelixFold3 and baselines on the PoseBusters(V1) with 428 targets.



(b) Comparison of the success rate of HelixFold3 and baselines on the PoseBusters(V2) with 308 targets.



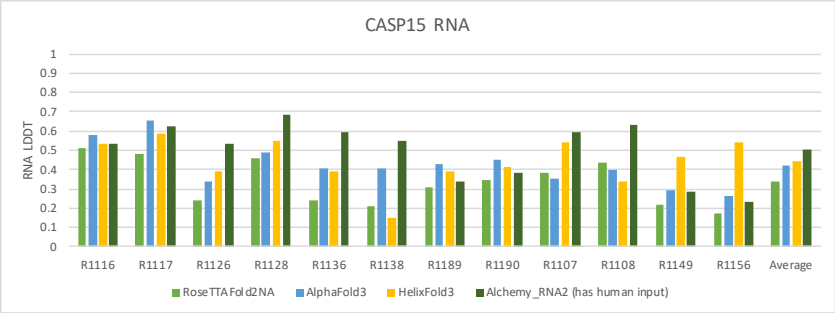
(c) PoseBusters Quality check.

Figure 1: Results for ligands.

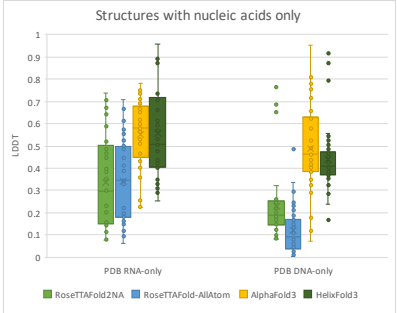
for nucleic acid target structure prediction, as well as RoseTTAFold-AllAtom, an all-atom biomolecular structure prediction model.

2.3 Proteins

For protein-protein complex structure prediction, AlphaFold-Multimer represents a significant advancement over earlier models, though its success rate and accuracy still have room for improvement. AlphaFold3 further enhances these capabilities, delivering superior predictive performance. We analyzed 186 protein complexes released in the PDB from January 19, 2022, to November 30, 2022 to assess HelixFold3 and the competitive methods. As depicted in Figure 3, HelixFold3 has already outperformed AlphaFold-Multimer in predicting protein-protein interfaces, yet a gap remains between HelixFold3 and AlphaFold3. To address this, ongoing research will concentrate on targeted optimizations and iterative refinements of HelixFold3, with the aim of achieving greater accuracy and reliability in protein-protein complex predictions.

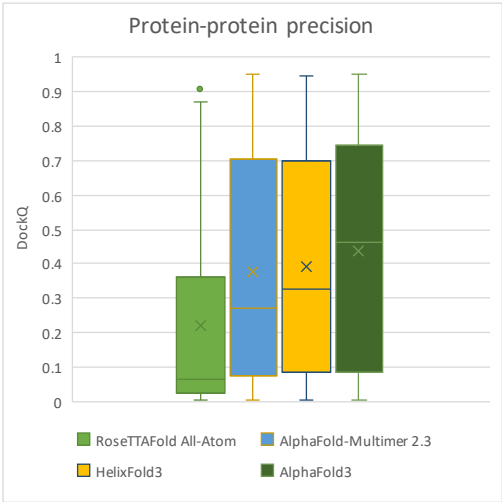


(a) Comparison of RNA LDDT on the 12 RNA targets in CASP15 with available ground-truth structures in PDB.

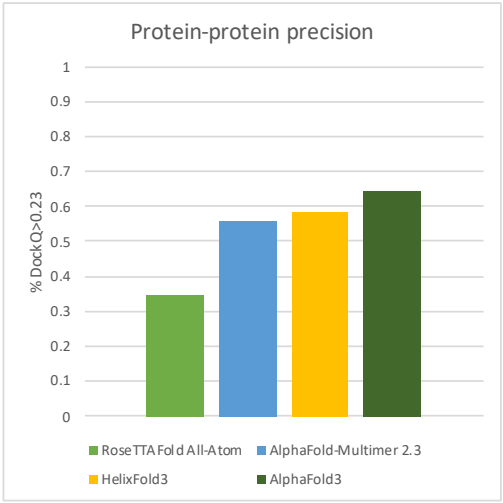


(b) Comparison on RNA-only or DNA-only complexes from PDB.

Figure 2: Results for nucleic acid targets.



(a) DockQ of protein-protein interface targets collected from PDB.



(b) Success rate of protein-protein interface targets collected from PDB

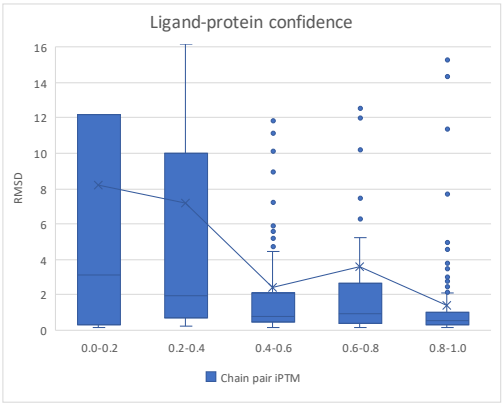
Figure 3: Results for protein targets.

2.4 Model Confidence

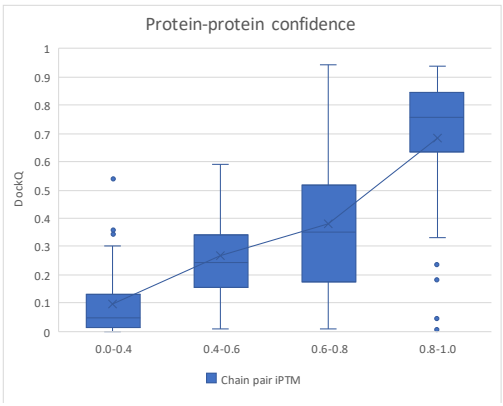
Confidence scores from structure prediction models are essential for assessing the accuracy of their predictions. HelixFold3 employs several confidence metrics, including pLDDT, pAE, and pTM, to evaluate its predictions. We performed a correlation analysis between these confidence scores and the actual accuracy of predicted structures using data from protein complexes. HelixFold3 generated confidence scores for datasets, including small molecule ligand-protein interactions from PoseBusters, as well as protein-protein complexes, RNA molecules, and DNA molecules collected from the PDB. Across all these datasets, we observed a strong correlation between the confidence scores and the structural accuracy (Figure 4), indicating the reliability of these metrics in evaluating prediction quality.

3 Conclusion

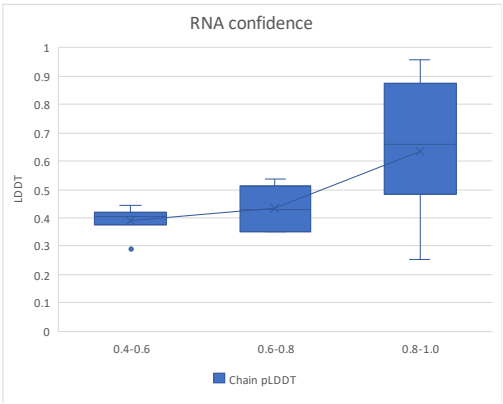
Our team is rigorously developing HelixFold3 to replicate the capabilities of AlphaFold3. We reported our current progress, which shows that HelixFold3's accuracy on conventional ligands, nucleic acids, and proteins is approaching that of AlphaFold3. The inference code and current version of model parameters of HelixFold3 are open-sourced on GitHub to facilitate its use by researchers. We will continue to refine the model and will provide updates on HelixFold3's performance with larger and more diverse datasets. We welcome you to stay updated on our progress. We invite you to follow our progress. For inquiries regarding HelixFold3 or potential commercial and research collaborations with the PaddleHelix team, please contact us at `baidubio_cooperate@baidu.com`.



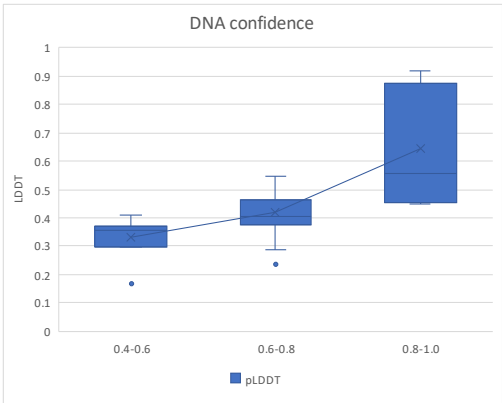
(a) Relation of iPTM and RMSD for ligand-protein interfaces.



(b) Relation of iPTM and DockQ for protein-protein interfaces.



(c) Relation between pLDDT and LDDT for RNA targets.



(d) Relation between pLDDT and LDDT for DNA targets.

Figure 4: Model confidence scores of HelixFold.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[2] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pages 2021–10, 2021.

[3] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[4] Guoxia Wang, Xiaomin Fang, Zhihua Wu, Yiqun Liu, Yang Xue, Yingfei Xiang, Dianhai Yu, Fan Wang, and Yanjun Ma. Helixfold: An efficient implementation of alphafold2 using paddlepaddle. *arXiv preprint arXiv:2207.05477*, 2022.

[5] Xiaomin Fang, Fan Wang, Lihang Liu, Jingzhou He, Dayong Lin, Yingfei Xiang, Kunrui Zhu, Xiaonan Zhang, Hua Wu, Hui Li, et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nature Machine Intelligence*, 5(10):1087–1096, 2023.

[6] Xiaomin Fang, Jie Gao, Jing Hu, Lihang Liu, Yang Xue, Xiaonan Zhang, and Kunrui Zhu. Helixfold-multimer: Elevating protein complex structure prediction to new heights. *arXiv preprint arXiv:2404.10260*, 2024.

[7] Lihang Liu, Donglong He, Xianbin Ye, Shanzhuo Zhang, Xiaonan Zhang, Jingbo Zhou, Jun Li, Hua Chai, Fan Wang, Jingzhou He, et al. Pre-training on large-scale generated docking conformations with helixdock to unlock the potential of protein-ligand structure prediction models. *arXiv preprint arXiv:2310.13913*, 2023.

[8] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[9] Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

[10] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xv. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1539–1549, 2023.

[11] Ke Chen, Yaoqi Zhou, Sheng Wang, and Peng Xiong. Rna tertiary structure modeling with briq potential in casp15. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1771–1778, 2023.

[12] Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of protein–nucleic acid complexes using rosettafoldna. *Nature methods*, 21(1):117–121, 2024.