
HELIXFOLD-MULTIMER: ELEVATING PROTEIN COMPLEX STRUCTURE PREDICTION TO NEW HEIGHTS

Xiaomin Fang, Jie Gao, Jing Hu, Lihang Liu, Yang Xue, Xiaonan Zhang, Kunrui Zhu

PaddleHelix Team*

Baidu Inc.

<https://paddlehelix.baidu.com/>

ABSTRACT

While monomer protein structure prediction tools boast impressive accuracy, the prediction of protein complex structures remains a daunting challenge in the field. This challenge is particularly pronounced in scenarios involving complexes with protein chains from different species, such as antigen-antibody interactions, where accuracy often falls short. Limited by the accuracy of complex prediction, tasks based on precise protein-protein interaction analysis also face obstacles. In this report, we highlight the ongoing advancements of our protein complex structure prediction model, HelixFold-Multimer, underscoring its enhanced performance. HelixFold-Multimer provides precise predictions for diverse protein complex structures, especially in therapeutic protein interactions. Notably, HelixFold-Multimer achieves remarkable success in antigen-antibody and peptide-protein structure prediction, surpassing AlphaFold-Multimer by several folds. HelixFold-Multimer is now available for public use on the PaddleHelix platform, offering both a general version and an antigen-antibody version. Researchers can conveniently access and utilize this service for their development needs.

Keywords Protein complex structure · PaddleHelix · Antigen-antibody

1 Introduction

Understanding protein structures is vital for decoding their diverse functions, aiding drug discovery, and revealing evolutionary relationships. These structures elucidate enzymatic activity, signal transduction, and regulatory mechanisms. Accurately predicting protein structures enables protein engineering, driving biotechnological innovations for biomedical and industrial applications. Though current tools information.[1, 2, 3] can predict protein monomer structures with considerable accuracy, accurately predicting the structures of protein complexes containing multiple chains remains a formidable challenge. This difficulty arises from the complexity of capturing the interactions between chains within protein complexes, which cannot be adequately addressed solely through co-evolutionary.

Computational protein docking tools, such as ZDock [4], HDock [5], ClusPro [6], and HADDOCK [7], offer a cost-effective alternative, utilizing various sampling techniques to explore the conformational space and employing scoring functions to evaluate and identify optimal conformations. Despite their invaluable contributions to structural bioinformatics, these tools encounter limitations such as constraints in scoring function accuracy, conformational sampling complexities, and the treatment of protein flexibility, potentially compromising the precision of their predictions.

Recent developments in end-to-end deep learning-based protein complex approaches [8, 9, 10, 11, 12], exemplified by AlphaFold-Multimer [8], aim to simultaneously fold and dock proteins within a complex, directly predicting complex structures from protein sequences. This integrated Fold and Dock methodology represents a novel problem-solving approach that notably improves the precision of structure prediction for various protein complex types. While AlphaFold-Multimer has demonstrated commendable precision compared to preceding methods for predicting protein complex structures, its performance still falls short in numerous scenarios. Nonetheless, previous studies [13, 14, 15]

*Corresponding author: Xiaomin Fang (fangxiaomin01@baidu.com). The authors are listed in alphabetical order according to their last names.

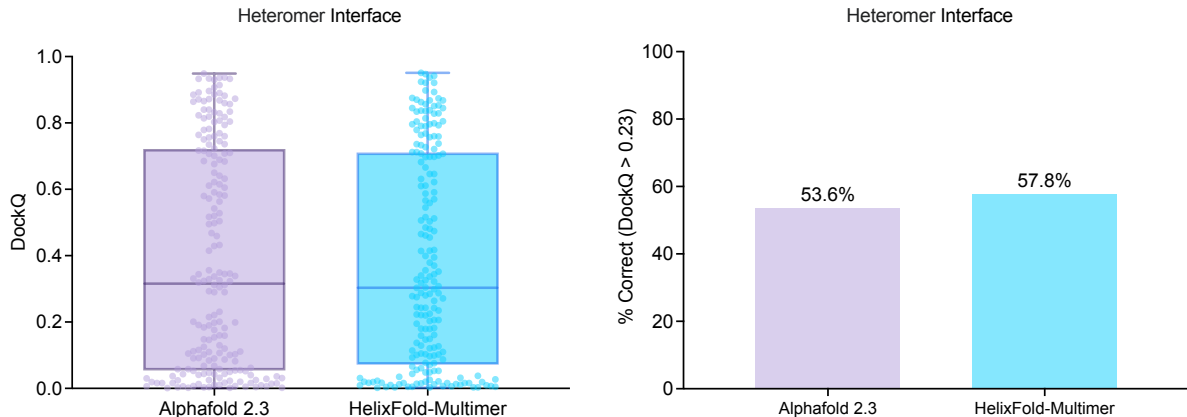


Figure 1: Comparison between AlphaFold and HelixFold-Multimer for heteromeric protein complexes. The box plots on the left show DockQ score distributions, while the bar graph on the right reflects the percentage of accurate (DockQ > 0.23).

highlighted the limitations of AlphaFold-Multimer, particularly evident when tackling protein complexes that lack extensive paired Multiple Sequence Alignments (MSAs). This deficiency becomes especially apparent in modeling antigen-antibody complexes, alongside other adaptive immune recognition mechanisms, and in complexes comprising chains from diverse species. Due to the insufficient accuracy in predicting the structures of these protein complexes, researchers encounter challenges in conducting thorough analyses of protein interactions. This limitation not only hinders the understanding of the protein-protein interactions but also restricts the potential for innovative applications in protein engineering and design.

We introduce HelixFold-Multimer, a novel approach that significantly enhances the accuracy and widens the application scope of protein complex structure prediction. Building upon the groundwork laid by our prior work, HelixFold [16] and HelixFold-Single [1], HelixFold-Multimer improves cross-chain interaction modeling by integrating domain expertise into model architecture, input features, and training strategies. HelixFold-Multimer is designed to provide highly accurate predictions for a diverse array of protein complex structures, particularly those involved in therapeutic protein interactions. We find that HelixFold-Multimer exhibits impressive effectiveness in accurately predicting a wide range of complex protein structures, notably excelling in predicting antigen-antibody, nanobody-antigen interfaces, and peptide-protein interfaces. Furthermore, HelixFold-Multimer demonstrates exceptional performance with commonly studied protein targets in drug development, suggesting its potential to optimize the protein design process for therapeutic development. HelixFold-Multimer currently consists of two versions: a general version designed for predicting common protein complex structures, particularly interfaces between peptides and proteins, and a specialized version developed for predicting antigen-antibody structures. Both the general version and the antigen-antibody version are publicly available for use on the PaddleHelix platform.

2 Results

2.1 Results of General Version

We first present the results of the general version of HelixFold-Multimer. We compare its performance on heteromeric protein complexes and peptide-protein complexes.

2.1.1 Overall Performance of General Proteins

Heteromeric complexes are formed by the interaction of multiple distinct protein chains. The comparison of HelixFold-Multimer with AlphaFold and RosettaFold on heteromeric protein complexes is summarized in Figure 1. The median DockQ score of HelixFold-Multimer (0.304) is comparable to that of AlphaFold (0.316). When considering percentage of the correct samples with DockQ > 0.23, the performance of HelixFold-Multimer (57.8%) exceeds that of AlphaFold (53.6%) by 4.2%. This highlights the superior capabilities of HelixFold-Multimer in the task of predicting the structures of heteromeric protein complexes.

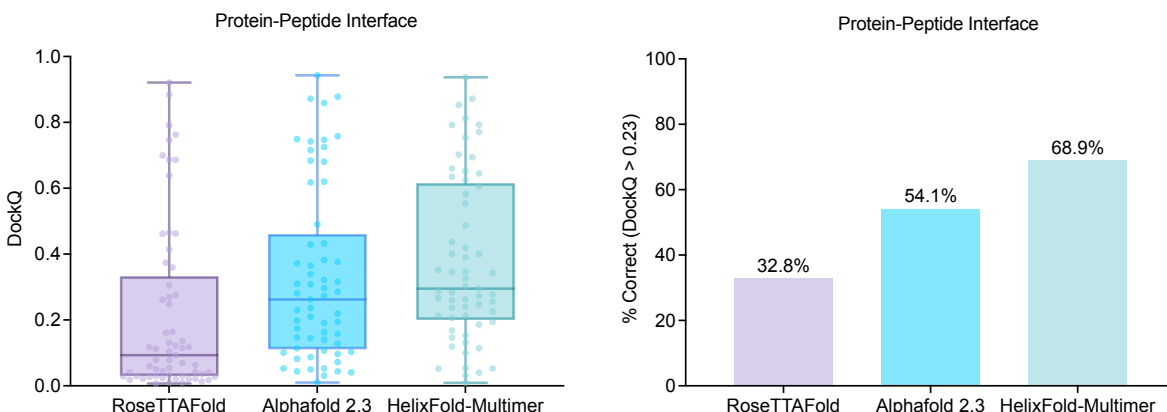


Figure 2: Comparison of AlphaFold and HelixFold-Multimer in protein-peptide docking accuracy. The box plots on the left show DockQ score distributions, while the bar graph on the right reflects the percentage of accurate (DockQ > 0.23).

Peptides, being shorter chains of amino acids compared to proteins, can display a wide range of flexible structures. The challenge in predicting the precise structures of protein-peptide complexes lies in accurately modeling the flexibility of the peptides. The results of protein-peptide interfaces is shown in Figure 2. HelixFold-Multimer achieved a median DockQ score of 0.295, marking a pronounced improvement compared to the median DockQ score of 0.262 obtained by AlphaFold and 0.093 by RoseTTAFold. This enhancement in predictive accuracy is further evidenced by the success rate metric (DockQ > 0.23), where HelixFold-Multimer attained a success rate of 68.9%. In contrast, AlphaFold exhibited a success rate of 54.1% for the same threshold. The structural analysis of protein-peptide complexes profoundly influences the research and development of peptide therapeutics. Understanding the intricate interplay between proteins and peptides not only aids in the design of more effective drugs but also sheds light on fundamental biological processes, paving the way for innovative therapeutic strategies.

2.2 Results of Antigen-Antibody Version

Antigen-antibody structure prediction is essential for unraveling immune responses and informing drug design, vaccine development, and diagnostic assays. It serves as a cornerstone in advancing biomedical research and enhancing therapeutic interventions. When assessing HelixFold-Multimer for antigen-antibody, we evaluate its effectiveness across three key dimensions: overall performance, model confidence, and efficacy on various antigen-antibody categories.

2.2.1 Overall Comparison

We first assess HelixFold-Multimer on antibody-related complex structure predictions across two categories: Antibody-Antigen Interfaces (consisting of heavy and light chains) and Nanobody-Antigen Interfaces, as illustrated in Figure 3. HelixFold-Multimer exhibits significant advantages in predicting both antibody-antigen and nanobody-antigen interfaces. In the evaluation of antibody-antigen interfaces, HelixFold-Multimer's predictions surpassed those of AlphaFold, achieving a mean DockQ score of 0.390, marking a substantial 5-fold improvement. HelixFold-Multimer achieves an impressive success rate (DockQ > 0.23) of 52.7% in accurately predicting antibody-antigen structures. This contrasts sharply with AlphaFold's success rate of 7.6% and RoseTTAFold's 4.6%. The HelixFold-Multimer outperformed both AlphaFold and RoseTTAFold in predicting the nanobody-antigen interface. It achieved a median DockQ score of 0.703 and a mean DockQ of 0.538, representing a significant improvement over previous baselines. With a success rate of 69.2%, HelixFold-Multimer demonstrated superior predictive capabilities compared to AlphaFold (15.4%) and RoseTTAFold (7.7%). This indicates its effectiveness in accurately predicting interactions, even for smaller and more variable nanobody structures.

We subsequently evaluated the predictive accuracy for antibody VH-VL interfaces, as shown in Figure 4. While existing protein structure prediction tools have achieved decent accuracy in predicting the conformation of antibodies (including heavy chain H and light chain L), higher-precision prediction results remain highly meaningful for antibody understanding. The prediction of the VH-VL chain interface by HelixFold-Multimer yielded a median DockQ score of 0.823, surpassing AlphaFold's score of 0.774 and RoseTTAFold's score of 0.653. HelixFold-Multimer achieved a very high accuracy rate (DockQ > 0.8) of 59.5%, representing a significant improvement over AlphaFold's 37.4% and

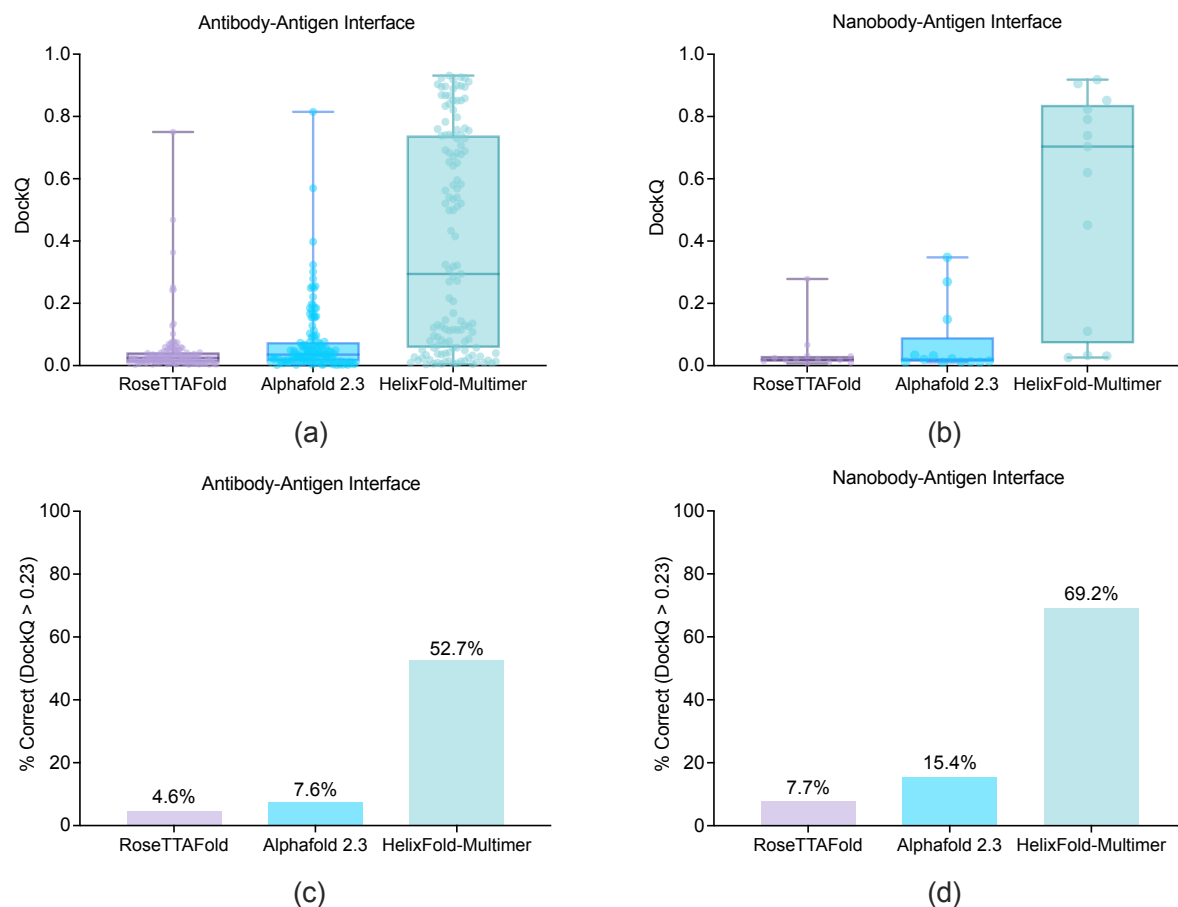


Figure 3: Overall comparison for antibody-related complex structure predictions. Panels (a), and (b) show the distribution of DockQ scores, which indicate the quality of the predicted structures, for the antibody-antigen interface and nanobody-antigen interface, respectively. Meanwhile, panels (c), and (d), display the percentage of correct predictions for these interfaces. The data underscores a notable improvement in prediction accuracy with the HelixFold-Multimer model compared to both AlphaFold and RoseTTAFold across all interfaces evaluated.

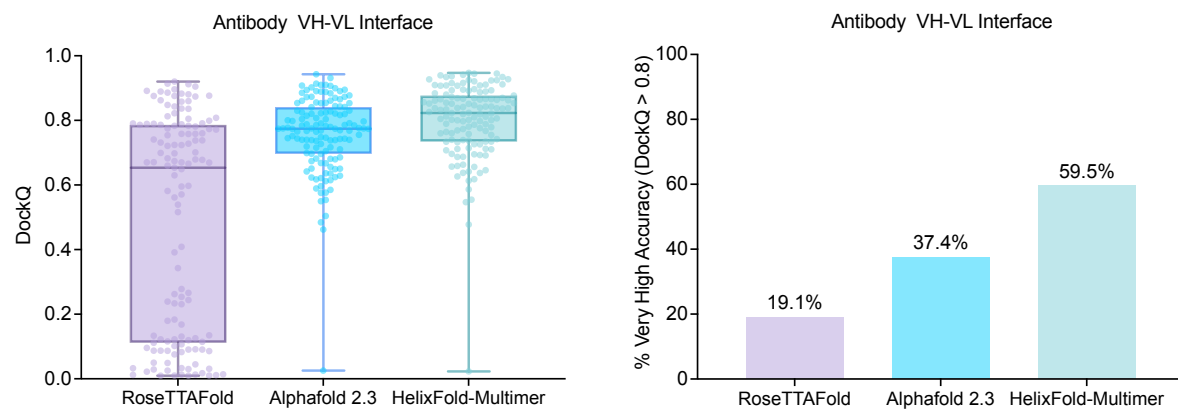


Figure 4: Evaluation of antibody VH-VL interfaces: Box plots on the left illustrate the performance comparison among RoseTTAFold, AlphaFold, and HelixFold-Multimer. The bar graph on the right quantifies the percentage of predictions with very high accuracy (DockQ > 0.8).

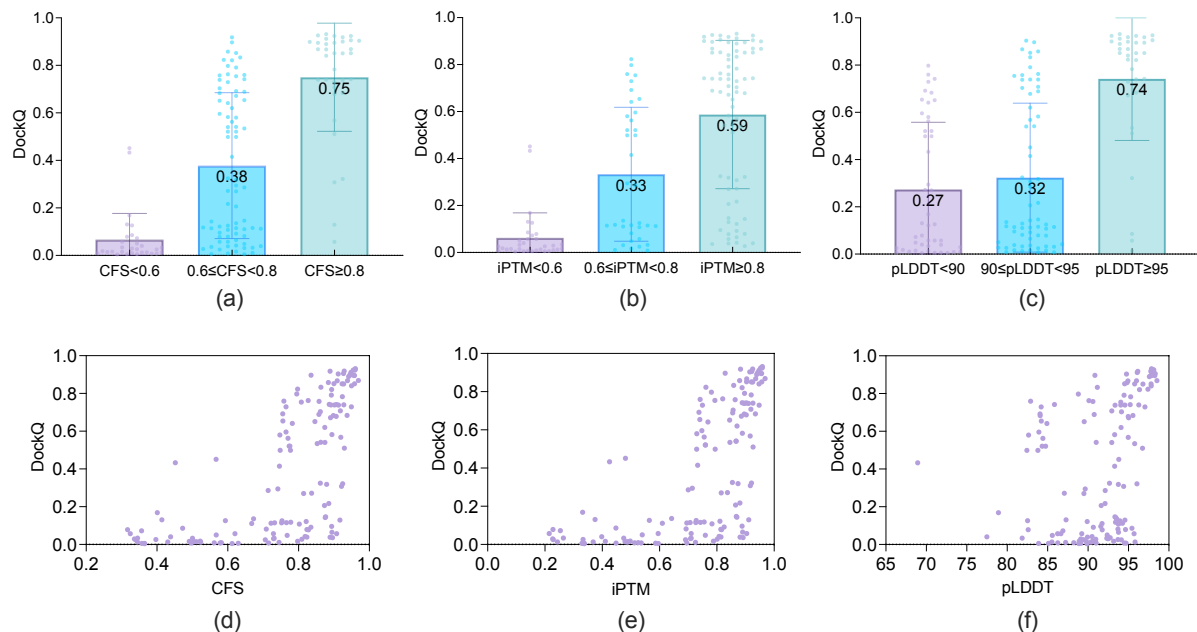


Figure 5: Correlational analysis between the DockQ scores and the predictive indicators outputted by HelixFold-Multimer for antigen-antibody docking. Subfigures (a) to (c) present box plots illustrating the distribution of DockQ scores across different levels of predictive confidence score, iPTM, and pLDDT scores. Meanwhile, subfigures (d) to (f) depict scatter plots illustrating the associations between DockQ scores and the corresponding predictive metrics. Note that "CFS" is an abbreviation for "confidence score".

RoseTTAFold's 19.1%. These results indicate the proficiency of HelixFold-Multimer in accurately modeling antibody structures.

HelixFold-Multimer's outstanding ability to predict antibody-related structures suggests its potential to streamline the identification and development of new antibody-based therapeutics. This proficiency has the potential to transform the landscape of drug discovery, providing more efficient and accurate techniques for designing therapeutic antibodies customized for specific targets and diseases.

2.2.2 Model Confidence

The correlation between confidence scores and model accuracy was exhaustively examined to evaluate the reliability of the HelixFold-Multimer in predicting antigen-antibody interactions.

We performed an analysis of three scoring metrics within the HelixFold-Multimer model, which assesses the confidence levels of the inference results: confidence scores (CFS), iPTM scores, and pLDDT scores. Subsequently, we organized the test samples into groups based on these confidence scores and presented the distribution of DockQ scores within each group (see Figure 5(a)-(c)). The figures suggest a clear pattern: in general, antigen-antibody complexes identified by the model as having high confidence levels tend to have significantly higher DockQ scores compared to those with lower confidence levels. Figure 5(d)-(f) elucidates the relationship between the accuracy of test samples and scoring metrics through scatter plots. All the scoring metrics show correlations with DockQ. Notably, confidence scores (Pearson correlation: 0.664) and iPTM (Pearson correlation: 0.658) exhibit stronger correlations with DockQ compared to pLDDT (Pearson correlation: 0.344). This robust correlation between scoring metrics and DockQ scores will serve as valuable guidance for leveraging HelixFold-Multimer in antibody development. For instance, analyzing these scoring metrics can help identify antibodies or antigens with higher research potential.

2.2.3 Efficacy on Various Antigen-Antibody Categories

We conducted an assessment of HelixFold-Multimer's effectiveness across various species groups and sequence identity intervals to identify its strengths in specific Antigen-Antibody categories. Species were classified into three groups: Homo sapiens, Mus musculus, and other species (including macaca mulatta, rattus norvegicus, synthetic construct,

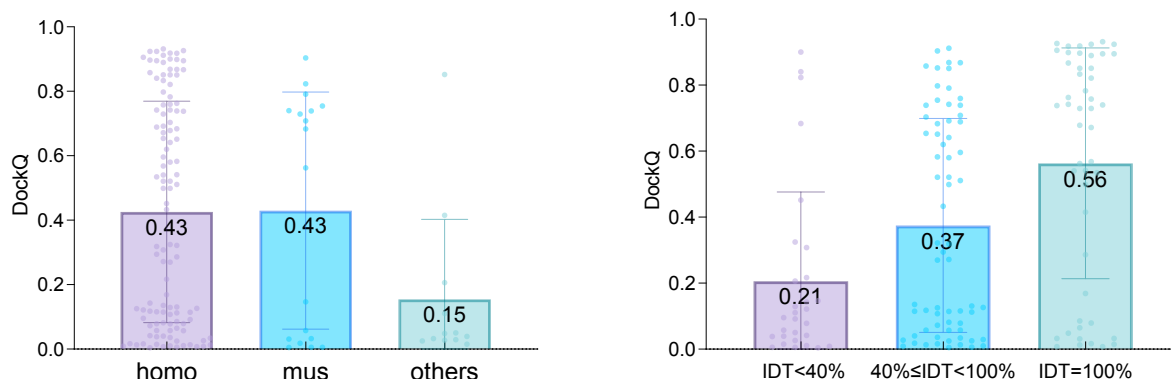


Figure 6: Performance analysis on various antigen categories. The left figure shows the performance of antigens from different species. The right figure shows the performance of antigens with different sequence identities to the training samples. "IDT" is short for "sequence identity".

gallus gallus, and oryctolagus cuniculus), based on sample counts. The classification of species was based on the origin of the antibody heavy chains. The sequence identity was calculated through MMseqs2 [17] on antigen chains.

Different species and sequence similarities exhibit significant differences. As shown in the left subfigure of Figure 6, HelixFold-Multimer achieves higher mean DockQ scores for Homo sapiens (0.43) and Mus musculus (0.43) compared to other species (0.15). This outcome aligns with expectations, given that the majority of antibody research focuses on Homo sapiens and Mus musculus. When evaluating antigens grouped by sequence similarity (the right subfigure of Figure 6, we observe a strong correlation between antigen sequence similarity and accuracy, as expected. HelixFold-Multimer achieves a mean DockQ score of up to 0.56 for samples corresponding to antigens already present in the training dataset. We believe that HelixFold-Multimer's performance on popular antigen-antibody complexes is already quite reliable and can effectively assist researchers in analyzing interactions between antigens and antibodies. In the future, we aim to further enhance the model's performance on less common antigens.

3 Conclusion

HelixFold-Multimer showcases exceptional prowess in predicting the structures of diverse protein complexes, surpassing prior methodologies by a significant margin. Particularly noteworthy is its outstanding accuracy in forecasting the structures of protein-peptide and antigen-antibody complexes, crucial components in therapeutic interventions. These findings not only highlight the model's potential in advancing drug design for macromolecular therapeutics but also underscore its pivotal role in shaping the future landscape of therapeutic development. By harnessing the predictive power of HelixFold-Multimer, researchers are poised to unlock new possibilities in precision medicine and therapeutic innovation.

Currently, HelixFold-Multimer is available on the PaddleHelix platform, a comprehensive drug development service platform. Researchers can conveniently utilize the platform for inferential analysis of protein interactions. In the future, we aim to further enhance the accuracy of antibody and peptide-related structure predictions and apply them to a broader spectrum of drug development tasks. Stay tuned for more updates.

Appendices

1 Experimental Settings

1.1 Baselines

AlphaFold and RosettaFold are currently the most representative end-to-end protein structure prediction models in the industry, serving as benchmarks. Therefore, we have chosen them as our comparison baseline.

RosettaFold [18] integrates deep learning with biochemical knowledge, showing promising results in single-chain protein structure prediction. However, its accuracy in predicting multi-chain protein complex structures is not satisfactory. On the other hand, AlphaFold [2, 8], particularly its Multimer version [8], AlphaFold-Multimer, performs exceptionally well in predicting general protein complex structures. Nonetheless, its accuracy drops significantly in scenarios lacking sufficient cross-chain evolutionary information, such as antigen-antibody complexes.

For the sake of convenience in comparison, this report contrasts the results of single-model inference rather than using ensemble methods (the original AlphaFold employs inference from 25 models followed by ensemble). We conducted inference using version 2.3 model 1 from AlphaFold’s GitHub repository. For the RoseTTAFold model, we referred to the setup described in RoseTTAFold2 [19].

1.2 Evaluation Metrics

To evaluate the accuracy of predicted protein complexes, we employed the DockQ metric [20]. Consistent with common practice, we utilized the DockQ threshold to gauge the success rates of model predictions. Predictions with a DockQ value above 0.23 were deemed accurate, while those surpassing 0.8 were categorized as exhibiting very high accuracy.

1.3 Confidence Metrics

In our study, three scoring metrics were used: pLDDT (the predicted local distance difference test) [21], PTM (the predicted TM-score) [22], and confidence score. pLDDT and PTM are metrics provided by the model and are indicators of overall structural accuracy. For interface accuracy of different chains, we measured via iPTM (the interface predicted TM-score), which is calculated based on the residues at the interface. Additionally, we used the confidence score[8] index intended to simultaneously consider the accuracy of the inter-chain interface and the overall prediction accuracy. The confidence score was calculated by assigning different weights to the PTM and iPTM values.

1.4 Sequence Identity

To comprehensively grasp the impact of sequence similarity on model performance, we employed sequence identity as the similarity metric. Sequence identity is defined as the percentage of residues in the evaluation set chain that match those in the training set chain. We utilized the default configuration settings of the MMSeqs2[17] software suite for this computation.

2 Datasets

2.1 General Version

Heteromeric protein complexes released from the period between January 16th, 2022, and December 12th, 2022, were gathered from Protein Data Bank (PDB) [23] as the evaluation set. Proteins with more than 1400 residues are excluded. These test samples are not present in the training set. The evaluation set for heteromeric proteins consists of 194 complexes.

To assess the models on protein-peptide complexes, we derived a subset of protein complexes from the heteromeric evaluation set based on the length of the shortest chain. We established a criterion where the shortest chain’s maximum length was set to 50 residues. Subsequently, we obtained a total of 61 instances of peptide-protein complexes.

2.2 Antigen-Antibody Version

We selected samples from the SAbDab [24] database with release dates between January 25, 2023, and August 9, 2023, as the evaluation set for antibody-related data, ensuring that the evaluation samples were not present in our training

set. Further, the antigen type of peptide and protein were chosen. Additionally, we excluded samples containing more than 1400 residues. The test set comprised 131 antigen-antibody complexes and 13 nanobody-antigen complexes. The antibody chains from antigen-antibody complexes were extracted to serve as a test dataset for the evaluation of antibody variable regions (VH-VL). We extracted only the fragment variable regions of the heavy and light chains of the antibodies as inputs for the models.

References

- [1] 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*, pages 1–10, 2023.
- [2] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.
- [3] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.
- [4] Rong Chen, Li Li, and Zhiping Weng. Zdock: an initial-stage protein-docking algorithm. *Proteins: Structure, Function, and Bioinformatics*, 52(1):80–87, 2003.
- [5] Yumeng Yan, Huanyu Tao, Jiahua He, and Sheng-You Huang. The hdock server for integrated protein–protein docking. *Nature protocols*, 15(5):1829–1852, 2020.
- [6] Dima Kozakov, David R Hall, Bing Xia, Kathryn A Porter, Dzmitry Padhorny, Christine Yueh, Dmitri Beglov, and Sandor Vajda. The cluspro web server for protein–protein docking. *Nature protocols*, 12(2):255–278, 2017.
- [7] Sjoerd J De Vries, Marc Van Dijk, and Alexandre MJJ Bonvin. The haddock web server for data-driven biomolecular docking. *Nature protocols*, 5(5):883–897, 2010.
- [8] Richard Evans, Michael O’Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Židek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *bioRxiv*, pages 2021–10, 2021.
- [9] Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi Jaakkola, and Andreas Krause. Independent se(3)-equivariant models for end-to-end rigid protein docking. *International Conference on Learning Representations*.
- [10] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.
- [11] Zachary C Drake, Justin T Seffernick, and Steffen Lindert. Protein complex prediction using rosetta, alphafold, and mass spectrometry covalent labeling. *Nature communications*, 13(1):7846, 2022.
- [12] Wei Zheng, Qiqige Wuyun, Yang Li, Chengxin Zhang, P Lydia Freddolino, and Yang Zhang. Improving deep learning protein monomer and complex structure prediction using deepmsa2 with huge metagenomics data. *Nature Methods*, pages 1–11, 2024.
- [13] Rui Yin, Brandon Y Feng, Amitabh Varshney, and Brian G Pierce. Benchmarking alphafold for protein complex modeling reveals accuracy determinants. *Protein Science*, 31(8):e4379, 2022.
- [14] Rui Yin and Brian G Pierce. Evaluation of alphafold antibody-antigen modeling with implications for improving predictive accuracy. *bioRxiv*, 2023.
- [15] Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using alphafold2. *Nature communications*, 13(1):1265, 2022.
- [16] 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.
- [17] 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.
- [18] Ian R Humphreys, Jimin Pei, Minkyung Baek, Aditya Krishnakumar, Ivan Anishchenko, Sergey Ovchinnikov, Jing Zhang, Travis J Ness, Sudeep Banjade, Saket R Bagde, et al. Computed structures of core eukaryotic protein complexes. *Science*, 374(6573):eabm4805, 2021.

- [19] Minkyung Baek, Ivan Anishchenko, Ian Humphreys, Qian Cong, David Baker, and Frank DiMaio. Efficient and accurate prediction of protein structure using rosettafold2. *bioRxiv*, pages 2023–05, 2023.
- [20] Sankar Basu and Björn Wallner. Dockq: a quality measure for protein-protein docking models. *PLoS one*, 11(8):e0161879, 2016.
- [21] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. Iddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.
- [22] Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.
- [23] Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. Protein data bank (pdb): the single global macromolecular structure archive. *Protein crystallography: methods and protocols*, pages 627–641, 2017.
- [24] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42(D1):D1140–D1146, 2014.