Balancing Speed and Precision in Protein Folding: A Comparison of AlphaFold2, ESMFold, and OmegaFold

Anna Hýskova [1 2] **Eva Maršálková** [1 2] **Petr Šimeček** [1]

Abstract

We present a systematic benchmark of AlphaFold2, ESMFold, and OmegaFold on 1,336 protein chains deposited in the PDB between July 2022 and July 2024, ensuring no overlap with the training data of any tool. As expected, AlphaFold2 achieves the highest median TM-score (0.96) and lowest median RMSD (1.30Å), outperforming ESMFold (TM-score 0.95, RMSD 1.74Å) and OmegaFold (TM-score 0.93, RMSD 1.98Å). Crucially, however, many cases exist in which the performance gap among these methods is negligible, suggesting that the faster, alignment-free predictors (10-30 times faster) can be sufficient. We identify the sequence length, structural family, and experimental context features that drive substantial discrepancies in accuracy, and—leveraging ProtBert embeddings and per-residue confidence scores—train LightGBM classifiers that accurately predict when AlphaFold2's added investment is warranted. Our framework thus provides actionable guidance for practitioners deciding between speed and precision in large-scale structural pipelines.

1. Introduction

All living organisms—from simple bacteria and algae to plants, fungi, animals, and humans—contain a multitude of proteins that participate in virtually every cellular process (Alberts, 2017; Cooper, 2000). These molecular machines must fold into specific three-dimensional structures, organized hierarchically at four distinct levels (Figure 1): from the linear sequence of amino acids (primary structure), through local folding patterns of α-helices and β-sheets (secondary structure), to the complete three-dimensional arrangement of these elements (tertiary structure), and finally to the assembly of multiple chains into functional complexes (quaternary structure). While the amino acid sequence alone determines the final structure, protein misfolding often leads to disease (Selkoe, 2003). Experimental structure determination through X-ray crystallography, cryo-EM, or NMR spectroscopy remains the gold standard (Smyth & Martin, 2000; Milne et al., 2013; Hu et al., 2021), but these methods are time-consuming, expensive, and not always feasible. This creates an urgent need for reliable computational prediction methods, particularly as the gap between known protein sequences and solved structures continues to widen—with over 254 million sequences known (UniProtKB) but only about 230,444 experimentally determined structures available in the Protein Data Bank (as in January, 2025).

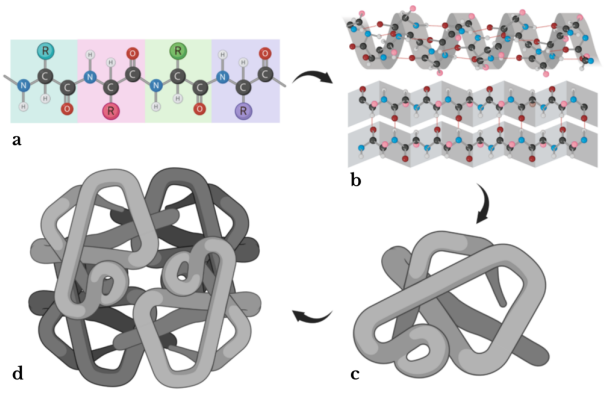

Figure 1. Hierarchical organization of protein structure. (a) Primary structure: the linear sequence of amino acids. (b) Secondary structure: local conformations including α-helices and β-sheets stabilized by hydrogen bonds. (c) Tertiary structure: the complete three-dimensional fold. (d) Quaternary structure: assembly of multiple chains into functional complexes.

The field of protein structure prediction has been transformed by artificial intelligence approaches. The introduction of AlphaFold2 in 2020 marked a watershed moment,

[1]Central European Institute of Technology (CEITEC), Faculty of Science, Masaryk University, Brno, Czechia [2]National Centre for Biomolecular Research (NCBR), Faculty of Science, Masaryk University, Brno, Czechia. Correspondence to: Petr Simecek <simecek@mail.muni.cz>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

achieving near-experimental accuracy (Jumper et al., 2021). This success has spurred the development of alternative approaches, particularly language model-based predictors like ESMFold and OmegaFold that can generate predictions without requiring multiple sequence alignments (Lin et al., 2023; Wu et al., 2022). These newer methods promise faster predictions and potentially better performance on challenging targets like designed or rapidly evolving proteins.

Despite these advances, the field lacks a comprehensive comparison of these tools' performance on truly novel proteins—structures solved after the tools' training cutoff dates (Kovalevskiy et al., 2024). Such evaluation is crucial for understanding each method's strengths and limitations, particularly as these tools become increasingly integrated into structural biology workflows. While the Critical Assessment of Structure Prediction (CASP) (Moult et al., 1995) and Continuous Automated Model EvaluatiOn (CAMEO) (Robin et al., 2021) provide valuable benchmarks, they are limited to participating methods and may not reflect real-world usage patterns.

Here, we present a systematic comparison of AlphaFold2, ESMFold, and OmegaFold using a dataset of over 1,300 protein structures deposited in the PDB between 2022 and 2024. Using multiple evaluation metrics including RMSD (Kufareva & Abagyan, 2012), TM-score (Zhang & Skolnick, 2004), and pLDDT (Tunyasuvunakool et al., 2021), we assess both overall performance and specific challenging cases. Our analysis reveals that while AlphaFold2 achieves the highest average accuracy, ESMFold and OmegaFold excel in particular niches, especially for proteins with limited homology information. Given 10-30 fold speed difference between alignment-free methods and AlphaFold2, our findings help researchers assess when the faster tools may provide sufficient accuracy for large-scale structural analyses.

2. Methods

2.1. Dataset

We compiled a benchmark dataset of 1,336 protein structures deposited in the Protein Data Bank (PDB) between July 2022 and July 2024. This temporal restriction ensures no overlap with training data used by AlphaFold2 (cutoff April 2020), ESMFold (June 2020), or OmegaFold (2021). The dataset contains three distinct groups: (1) single-chain monomers (980 structures), (2) small multi-chain complexes (245 structures with 2-6 chains), and (3) de novo designed proteins whose sequence does not naturally occur in any living organism (102 structures). De novo proteins were identified through PDB annotations marking them as "designed" or "synthetic construct" in the source organism field.

Structures were selected using the RCSB PDB Search API (Rose et al., 2021; Bittrich et al., 2023) with the following criteria: (i) deposition date between July 2022 and July 2024, (ii) protein-only structures without nucleic acids or oligosaccharides, (iii) chain lengths between 20 and 400 amino acids to ensure compatibility with all prediction tools, and (iv) availability of structural information in PDB format. To ensure diversity, structures within monomer and de novo protein groups were filtered to have at most 70% pairwise sequence identity.

We developed a custom PDB file parsing pipeline to extract complete amino acid sequences and experimental C_α coordinates. The pipeline addresses common challenges in PDB files, including non-standard residue numbering, insertion codes, and post-translational modifications. For modified residues, we reconstructed the original amino acid sequence using BioPython's extended residue dictionary and MODRES records. Structures containing non-standard residues without clear mapping to canonical amino acids (26 cases) were excluded from the analysis.

Each structure was annotated with protein family classifications using UniProt and PDBe APIs to map PDB identifiers to Pfam and InterPro database entries. These annotations enable analysis of prediction tools' performance across different protein families and structural motifs. The numbers of protein structures the dataset contained in various stages of the experiment are stated in Table 1. The final curated dataset, including all protein sequences, is available at Hugging Face Hub repository.

Table 1. Size of the dataset in various stages of the experiment.

GROUP	PDB HITS	SELECTED STRUCTURES	EVALUATED CHAINS
MONOMERS	3830	1000	979
SMALL COMPLEXES	3988	250	255
DE NOVO PROTEINS	139	103	102
IN TOTAL	7957	1353	1336

2.2. Structure Prediction Tools

Three tools were selected for protein structure prediction: AlphaFold2, ESMFold, and OmegaFold. While alignment-based AlphaFold2 is an obvious choice, considering how widely used it is (Kovalevskiy et al., 2024), language model-based ESMFold and OmegaFold were chosen because they provide promising results with much lower requirements on time and computational power, making them more suitable for large-scale applications (Lin et al., 2023; Wu et al., 2022).

AlphaFold2. We used AlphaFold v2.1.1 running on the in-

stitute's infrastructure with its monomer model and reduced database settings to optimize computational resources. The model architecture consists of two main components: (i) an Evoformer module, which processes multiple sequence alignments (MSAs) and pairwise representations through 48 transformer blocks, and (ii) a structure module that converts the refined representations into 3D coordinates through 8 equivariant transformer blocks with Invariant Point Attention. MSAs were generated using Uniref90, BFD, and MGnify databases. For each sequence, five model predictions were generated and ranked by predicted confidence, with the highest-confidence model (ranked_0.pdb) selected for evaluation.

ESMFold. Predictions were obtained via REST API calls to the ESM Metagenomic Atlas. ESMFold combines two components: (i) the ESM-2 protein language model with 15B parameters, pre-trained on masked sequence prediction, and (ii) a folding head consisting of 48 folding blocks that process sequence and pairwise representations. Unlike AlphaFold2, ESMFold predicts structures directly from single sequences without requiring MSA generation.

OmegaFold. Predictions were performed using OmegaFold v1.0 running on university computational cluster with NVIDIA A40 GPU. OmegaFold employs: (i) OmegaPLM, a 670M parameter language model trained on masked protein sequences, and (ii) a Geoformer architecture that refines the language model representations to be geometrically consistent before structure prediction. Like ESMFold, OmegaFold operates on single sequences without MSA requirements.

All predictions were made for individual protein chains, as both ESMFold and OmegaFold do not support prediction of protein complexes. While AlphaFold2 offers a multimer model, we used its monomer model to ensure fair comparison. The original dataset together with prediction outputs is available at HuggingFace Hub repository.

2.3. Evaluation Metrics

We employed three complementary metrics to assess prediction quality: RMSD measuring atomic distance deviation, TM-score evaluating topological similarity, and pLDDT reflecting model confidence.

Root Mean Square Deviation (RMSD). RMSD quantifies the average distance between corresponding C_α atoms in superimposed structures:

$$\text{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\delta_i^2} \tag{1}$$

where n is the number of aligned C_α atom pairs and δ_i is the distance between atoms in the i-th pair. To compute RMSD, we first extract C_α coordinates from both experimental and predicted structures, then determine the optimal superposition using the Bio.SVDSuperimposer module from BioPython (Cock et al., 2009a), which finds the rotation and translation matrices minimizing the RMSD value. While RMSD is widely used, it is sensitive to protein size and can be disproportionately affected by local structural deviations.

Template Modeling Score (TM-score). TM-score evaluates the topological similarity of protein structures while accounting for protein length:

$$\text{TM-score} = \max\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+(\frac{d_i}{d_0})^2}\right] \tag{2}$$

where L_N is the length of the reference structure, L_T is the number of aligned residues, d_i is the distance between the i-th pair of aligned residues after superposition, and $d_0 = 1.24\sqrt[3]{L_N - 15} - 1.8$ is a length-dependent scaling factor. TM-score ranges from 0 to 1, with values above 0.5 indicating proteins share the same fold and 1 representing perfect structural alignment. Unlike RMSD, TM-score is length-normalized and less sensitive to local structural variations.

Predicted LDDT (pLDDT). The predicted local distance difference test (pLDDT) is a confidence metric provided by each prediction tool. For each residue, it estimates the expected agreement between predicted and experimental structures on 0 to 100 scale. Scores above 90 indicate high prediction confidence. Scores above 70 suggest at least reliable backbone prediction.

For our analysis, we used the mean pLDDT across all residues in each protein chain. While pLDDT correlates with prediction accuracy, high confidence scores do not guarantee correct structure prediction, particularly for challenging targets like intrinsically disordered regions or proteins with limited homology information.

2.4. Statistical Analysis and Annotation

We compared these metrics across our dataset using Kruskal-Wallis tests followed by Dunn's method with Bonferroni correction for multiple comparisons. The correlation between metrics was assessed using Spearman's rank correlation coefficient.

Protein chains were mapped to functional annotations using UniProt and PDBe APIs. For family-specific analysis, we focused on Pfam and InterPro families with at least 10 member proteins in our dataset. The experimental method of structure determination (X-ray crystallography, cryo-EM, or NMR) was recorded for each chain to assess potential biases in prediction accuracy.

Predictions were classified as "poor" if they met any of the following criteria: average pLDDT < 70, TM-score

< 0.5, or RMSD > 9 Å. The 9 Å RMSD threshold was chosen to match the resolution cutoff used in training AlphaFold2. Statistical significance of family-specific enrichment in poor predictions was assessed using Fisher's exact test with Benjamini-Hochberg correction for multiple comparisons.

2.5. Implementation and Availability

All preprocessing was implemented in Python using BioPython (Cock et al., 2009b) for structure manipulation and tmtools for TM-score calculation (Xu & Zhang, 2010). Statistical analysis and visualization were performed in R (R Core Team et al., 2020). The complete dataset, including protein sequences, experimental structures, predictions, and evaluation results is available at HuggingFace Hub, https://huggingface.co/datasets/hyskova-anna/proteins. Source code and documentation are provided at GitHub, https://github.com/ML-Bioinfo-CEITEC/CAoPSPT.

3. Results

Structure predictions were attempted for 1,337 protein chains using AlphaFold2, ESMFold, and OmegaFold. Our AlphaFold2 pipeline failed to generate a prediction for one chain (8B2M:A), which was subsequently excluded from all analyses. The remaining 1,336 chains (shown in Table 1) were successfully predicted by all three tools and form the basis of our evaluation. Selected examples of predictions aligned with their experimental structures are visualized in Figure 2.

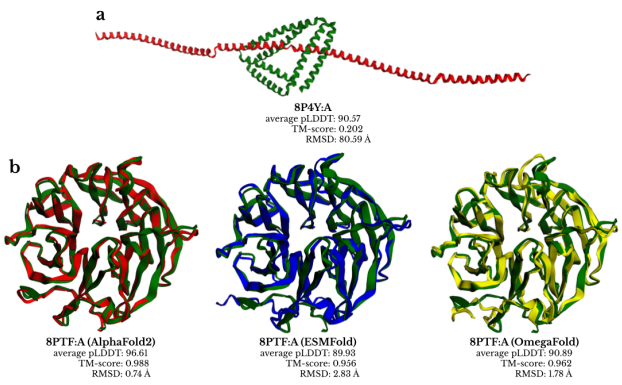

Figure 2. Examples of structure predictions from AlphaFold2 (red), ESMFold (blue) and OmegaFold (yellow) aligned with corresponding experimentally determined structures (green). (a) An example of a poorly predicted structure (8P4Y:A) by AlphaFold2. (b) Structure of protein 8PTF:A showing varying prediction quality across tools.

3.1. Comparative Performance Analysis

All three tools demonstrated generally satisfactory performance, with AlphaFold2 achieving the highest accuracy across all metrics (Figure 3). AlphaFold2 predictions showed the highest median TM-score (0.96) and lowest median RMSD (1.30 Å), followed by ESMFold (TM-score: 0.95, RMSD: 1.74 Å) and OmegaFold (TM-score: 0.93, RMSD: 1.98 Å). Consistently, AlphaFold2 displayed the highest confidence in its predictions with median pLDDT of 92.65, compared to 87.40 for ESMFold and 89.00 for OmegaFold.

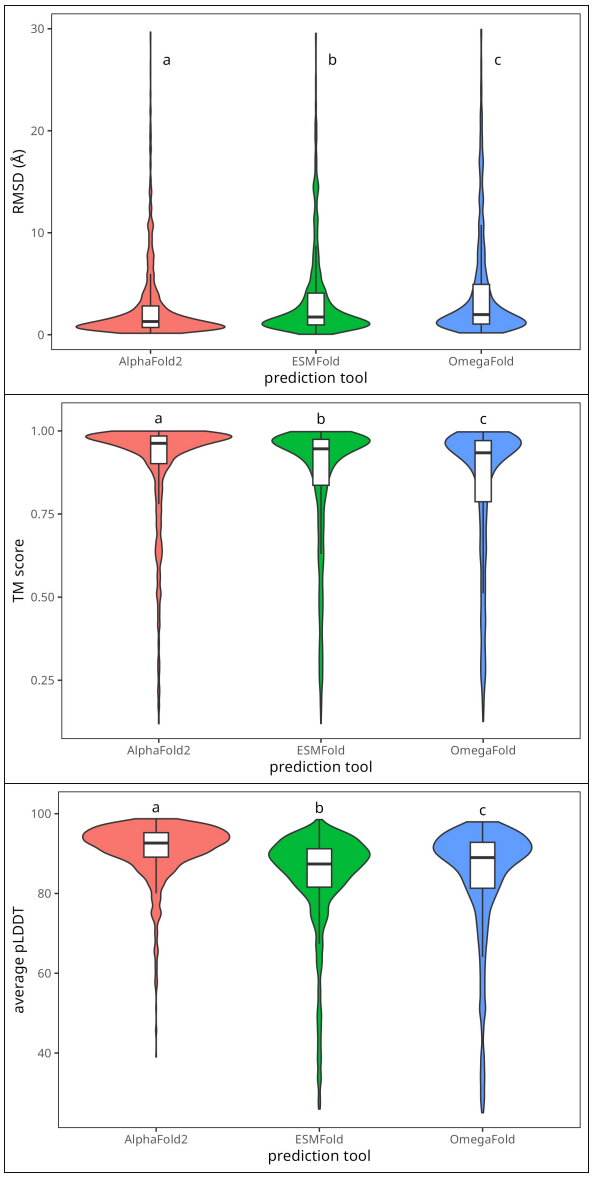

Figure 3. Performance comparison across prediction tools. Distribution of (a) RMSD values, (b) TM-scores, and (c) pLDDT scores. Box plots show median, quartiles, and outliers. All pair comparisons have been statistically significant ($p < 0.01$).

3.2. Metric Correlations and Their Dependencies on Sequence Length and Other Factors

We observed significant correlations between prediction confidence (pLDDT) and accuracy metrics (Figure 4). Most notably, there was a negative correlation between average pLDDT and RMSD (Spearman's $\rho = -0.87$, -0.87, and -0.88 for AlphaFold2, ESMFold, and OmegaFold, respectively) and a positive correlation between average pLDDT and TM-score ($\rho = 0.60$, 0.66, and 0.71). The correlation was strongest for ESMFold and OmegaFold, suggesting that their confidence scores more accurately reflect prediction quality than those of AlphaFold2.

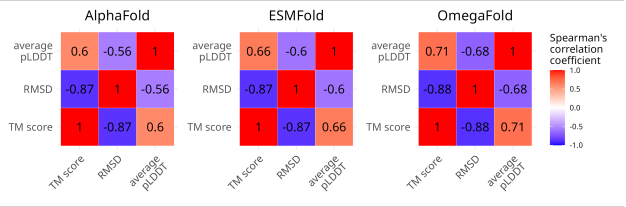

Figure 4. Correlation analysis between prediction metrics. Heatmaps show Spearman's correlation coefficients between average pLDDT, RMSD, and TM-score for each prediction tool. All correlations are statistically significant ($p < 0.001$).

While low-confidence predictions rarely achieved good accuracy metrics, we found numerous cases of incorrect structures with high pLDDT scores across all tools (Figure 5)

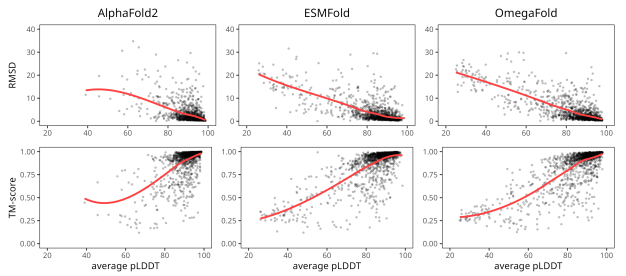

Figure 5. Dependency of RMSD and TM-score on average pLDDT of structures generated by different tools. The LOESS curve (red) was obtained by locally estimated scatterplot smoothing. Sample points with RMSD greater than 40 Å are omitted from the visualization for better clarity.

Analysis of sequence length dependency also revealed interesting patterns. While RMSD showed weak correlation with sequence length, TM-scores displayed stronger positive associations, particularly for AlphaFold2 ($\rho = 0.41$, $p < 0.001$). This suggests that predictions for shorter proteins (< 100 amino acids) tend to achieve lower TM-scores across all tools, though this trend is less pronounced in RMSD values due to the metric's inherent length dependency. ESMFold and OmegaFold showed weaker but still significant correlations with sequence length ($\rho = 0.29$ and $\rho = 0.28$, respectively, for TM-score).

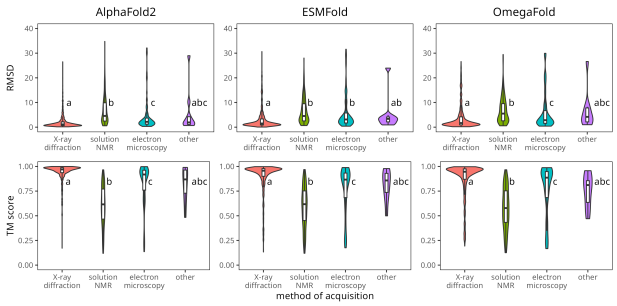

Figure 6. Dependency of RMSD and TM-score on the experimental method of acquisition of the protein chain structure. The differences between groups were tested by Kruskal-Wallis test, post-hoc comparisons were done using Dunn's method with a Bonferroni correction for multiple tests. Statistical significance visualized by difference in letter codes. Sample points with RMSD greater than 40 Å are omitted from the visualization for better clarity.

The experimental method used for structure determination significantly influenced prediction accuracy (Figure 6). All tools performed best on X-ray crystallography structures (median RMSD: 1.24 Å, 1.65 Å, and 1.89 Å for AlphaFold2, ESMFold, and OmegaFold, respectively) but struggled with NMR-determined structures (median RMSD: 2.31 Å, 2.89 Å, and 3.12 Å). This pattern likely reflects both the inherent flexibility of proteins amenable to NMR analysis and the predominance of X-ray structures in training data.

When comparing performance across different protein types (monomers, complexes, and de novo proteins), we observed an interesting pattern. While all tools generally performed similarly across these categories, there are two notable exceptions. First, ESMFold and OmegaFold achieved significantly lower RMSD values for de novo proteins compared to natural proteins. Second, AlphaFold2 showed a unique weakness with de novo proteins, achieving significantly lower TM-scores for these proteins compared to monomers and complexes. This suggests that language model-based tools may have an advantage in predicting structures of artificial proteins where evolutionary information is limited.

3.3. Analysis of Prediction Failures

We classified predictions as incorrect if they met any of the following criteria: average pLDDT < 70, TM-score < 0.5, or $RMSD > 9$ Å. AlphaFold2 produced the fewest incorrect predictions (8.9% of total), followed by ESMFold (13.0%) and OmegaFold (16.8%). The overlap of prediction failures between tools was limited, suggesting complementary strengths (Figure 7).

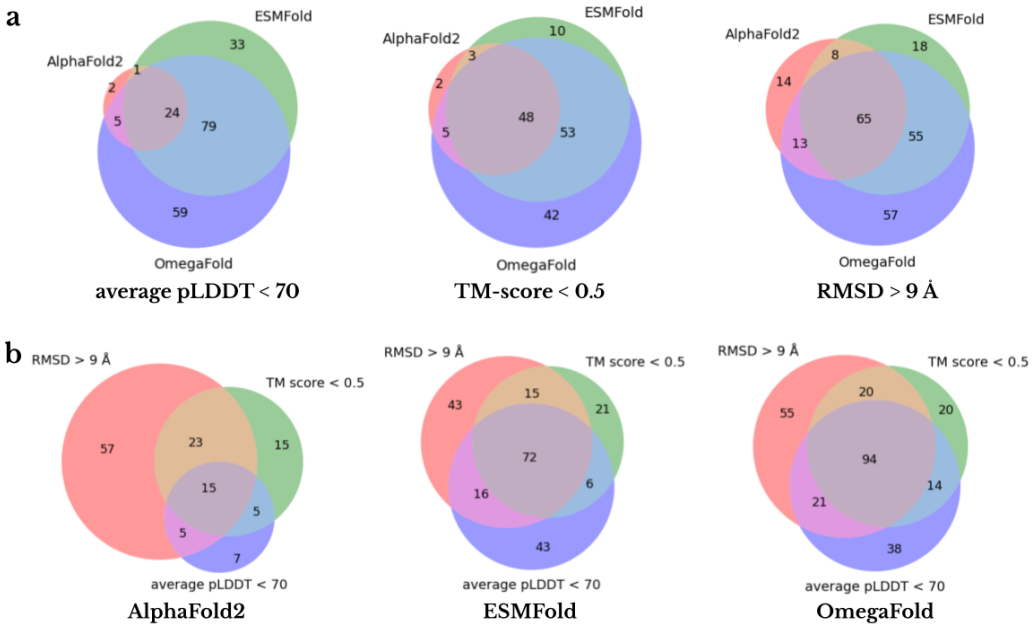

Figure 7. Venn diagrams comparing the overlap of poorly predicted protein chains based on three evaluation criteria: (a) average pLDDT < 70, TM-score < 0.5, and RMSD > 9 Å for AlphaFold2, ESMFold, and OmegaFold, and (b) the overlap of predictions that fail across the three metrics for each tool individually.

Analysis of protein families revealed that proteins lacking Pfam annotations were particularly challenging for AlphaFold2 but not for ESMFold or OmegaFold, highlighting the importance of evolutionary information in AlphaFold2's predictions. Conversely, viral proteins, especially from coronavirus, were better predicted by AlphaFold2 than by the language model-based tools. All tools showed reduced accuracy for proteins containing leucine-rich repeats or von Willebrand factor A-like domains, suggesting these structural motifs pose particular challenges for current prediction methods.

The analysis of protein family associations revealed distinctive patterns in prediction accuracy. Notably, AlphaFold2 showed significantly reduced performance for proteins lacking Pfam family annotations (odds ratio = 0.67, $p < 0.01$), while ESMFold and OmegaFold maintained consistent performance regardless of family assignments. This pattern was also observed with InterPro annotations, highlighting AlphaFold2's dependence on evolutionary information.

Certain protein families were consistently well-predicted across all tools. These included protein kinase domains (PF00069, IPR000719), the SH2 domain (IPR000980), and the NAD(P)-binding domain superfamily (IPR036291). Conversely, all tools struggled with leucine-rich repeats (IPR001611, IPR003591) and von Willebrand factor A-like domains (IPR036465), suggesting these structural motifs remain challenging for current prediction methods.

Interestingly, several protein families showed tool-specific prediction patterns. AlphaFold2 excelled at predicting viral protein families, particularly the viral RNA-dependent RNA polymerase (PF00680, IPR001205) and coronavirus-specific proteins (PF05409, IPR043503), achieving significantly better accuracy than ESMFold or OmegaFold ($p < 0.001$). Conversely, the S-adenosyl-L-methionine-dependent methyltransferase superfamily (IPR029063) showed markedly different prediction quality between AlphaFold2 (odds ratio = 1.83, $p < 0.05$) and the language model-based tools (odds ratio = 0.64 and 0.51 for ESMFold and OmegaFold respectively, $p < 0.001$).

3.4. Prediction of Structure Determination Success Using Machine Learning

To help identify potential failures in structure prediction, we trained gradient boosting LightGBM (Ke et al., 2017) models for AlphaFold2, ESMFold, and OmegaFold, respectively, using ProtBert BFD embeddings (Brandes et al., 2022) calculated from protein sequences and pLDDT scores. The models were trained to predict whether a structure prediction would likely be unsuccessful, allowing early identification of challenging cases. The trained models and source code are available on GitHub repository, enabling to assess potential challenges early in structure prediction pipelines.

As you can see in Figure 8, it is typically pLDDT, the length of the sequence, and a few selected embedding elements

that have the greatest influence on prediction.

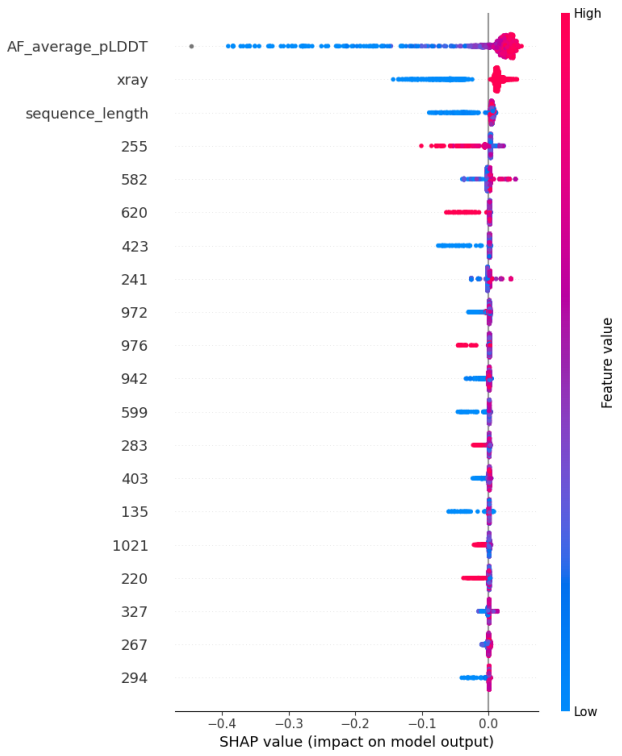

Figure 8. SHAP values of LightGBM model for AlphaFold2.

4. Discussion

Since the beginning of this decade, structural biology and protein structure prediction fields have undergone a significant transition. Currently, there are two large projects dealing with this issue: CASP (Moult et al., 1995) and CAMEO (Robin et al., 2021). While AlphaFold2 has participated in both CASP14 and CAMEO, ESMFold has entered only CASP15, and OmegaFold has not been included in either. However, both ESMFold and OmegaFold have been subsequently evaluated on CAMEO and CASP15 datasets by independent research groups (Moussad et al., 2023; Huang et al., 2023). There are also a few publications dealing with the comparison of protein structure prediction tools, but they usually focus mainly on AlphaFold2 and similar tools (e.g. ColabFold) (Kalogeropoulos et al., 2024) or perform the evaluation on a particular set of proteins, namely human proteins (Manfredi et al., 2024; 2025), snake venom toxins (Kalogeropoulos et al., 2024), and nanobodies (Valdés-Tresanco et al., 2023). This paper tries to increase our understanding by creating an inclusive dataset of protein structures recently added to PDB.

The key finding of this work is that AlphaFold2 outperforms ESMFold and OmegaFold on a majority of proteins in the dataset, measured by both RMSD and TM-score. When comparing the two protein language-based models, ESMFold seems to be a slightly better choice, as it produced fewer incorrect structures than OmegaFold and achieved significantly better median RMSD and TM-score. Still, the difference in performance between ESMFold and OmegaFold is much smaller compared to the gap between both of these tools and AlphaFold2.

While all three tools rarely produce a good prediction with low confidence, wrong structures with a high average pLDDT are outputted quite frequently. Our analysis revealed that prediction accuracy is influenced by various factors. All three tools performed best when predicting proteins whose experimental structure was determined by X-ray crystallography, while structures determined by NMR proved to be the most challenging. Because NMR is typically used to determine the structures of small proteins, a corresponding decrease in prediction accuracy is observed for shorter sequences.

Interestingly, proteins without family annotations proved particularly difficult for AlphaFold2 but did not change the performance of ESMFold and OmegaFold. A possible explanation is that proteins belonging to no family lack homologs with a known structure, which AlphaFold2 could use as a template during the prediction. In contrast, ESMFold and OmegaFold do not rely on MSAs and modelling templates, so their performance remained largely unaffected.

Our analysis shows several key insights, yet certain constraints of our study must be noted. First, the dataset does not contain only proteins whose experimental structure was previously unknown but also proteins that were just recently analyzed again, usually in different conditions. This might be an advantage for AlphaFold2, which uses a reduced PDB database for template searching during the prediction process. Moreover, the whole analysis focuses only on single protein chains without the context of their interacting partners, which might be crucial for structure formation, especially in protein complexes. Additionally, speed comparisons should be interpreted with caution, as pipelines for OmegaFold and AlphaFold2 predictions with different hardware configurations, potentially affecting relative performance metrics. Last but not least, all the protein chains in the dataset have a maximum length of 400 amino acids due to using ESMAtlas API.

The performance patterns we observed reflect fundamental architectural differences between these approaches. AlphaFold2's superior accuracy stems from leveraging evolutionary information through MSAs, but this becomes a limitation for de novo proteins where we observed reduced TM-scores. In contrast, language models learn protein grammar from sequence patterns alone, potentially capturing more general folding principles. The limited overlap in

prediction failures between tools suggests complementary error modes that could be exploited through ensemble approaches, though computational costs may be prohibitive for large-scale applications.

The recent proliferation of AlphaFold3 (Abramson et al., 2024; Callaway, 2024) and its alternatives, including Chai-1 (Chai Discovery, 2024), Boltz-1 (Wohlwend et al., 2024), and HelixFold3 (Liu et al., 2024), demonstrates the community's commitment to structure prediction. Independent benchmarks have begun evaluating these tools: FoldBench (Xu et al., 2025), evaluating 1,522 biological assemblies across nine tasks, found AlphaFold3 consistently outperforming alternatives across most categories, though all methods showed concerning failure rates exceeding 50% for antibody-antigen predictions. For protein-peptide interactions, newer models achieve dramatic improvements, with success rates of 70-80% under stringent criteria compared to 53% for AlphaFold2-multimer, and Protenix reaching 80.8% accuracy (Zhou et al., 2025). However, as shown in (Škrinjar et al., 2025), protein-ligand predictions reveal a critical limitation: current methods largely memorize poses from training data rather than genuinely predicting novel interactions, particularly struggling with ligands not seen in their training sets. Practical deployment is being facilitated by tools like ABCFold (Elliott et al., 2025), which standardizes inputs and outputs across different methods. This proliferation of capable yet specialized tools, each with distinct strengths and limitations, reinforces our findings: optimal structure prediction requires matching tools to specific tasks based on target type, available computational resources, and accuracy requirements rather than relying on any single universal solution.

Acknowledgments

The project was supported by the OPUS LAP program of the Czech Science Foundation, project no. 23-04260L ("Biological code of knots – identification of knotted patterns in biomolecules via AI approach"). Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning and its applications to structural biology. While improved protein structure prediction methods have many potential societal benefits, including accelerating drug discovery and understanding disease mechanisms, we acknowledge that these tools could potentially be misused for harmful purposes. However, we believe the benefits of openly comparing and understanding the strengths and limitations of these tools far outweigh the risks, particularly as this knowledge helps the scientific community make more informed decisions about which tools to use for different applications.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Alberts, B. *Molecular Biology of the Cell*. W.W. Norton, 2017. ISBN 978-1-317-56375-4. URL https://books.google.cz/books?id=jK6UBQAAQBAJ.

Bittrich, S., Bhikadiya, C., Bi, C., Chao, H., Duarte, J., Dutta, S., Fayazi, M., Henry, J., Khokhriakov, I., Lowe, R., Piehl, D., Segura, J., Vallat, B., Voigt, M., Westbrook, J., Burley, S., and Rose, Y. RCSB Protein Data Bank: Efficient Searching and Simultaneous Access to One Million Computed Structure Models Alongside the PDB Structures Enabled by Architectural Advances. *Journal of Molecular Biology*, 435(14):167994, July 2023. ISSN 0022-2836. doi: 10.1016/j.jmb.2023.167994. URL https://www.sciencedirect.com/science/article/pii/S0022283623000505.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Callaway, E. Ai protein-prediction tool alphafold3 is now more open. *Nature*, 635(8039):531–532, 2024.

Chai Discovery. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955. URL https://www.biorxiv.org/content/early/2024/10/11/2024.10.10.615955.

Cock, P., Antao, T., Chang, J., Chapman, B., Cox, C., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, June 2009a. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163. URL https://doi.org/10.1093/bioinformatics/btp163.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009b.

Cooper, G. *The Cell: A Molecular Approach*. Sunderland (MA): Sinauer Associates, 2000. ISBN 0-87893-106-6.

Elliott, L. G., Simpkin, A. J., and Rigden, D. J. Abcfold: easier running and comparison of alphafold 3, boltz-1 and chai-1. *Bioinformatics Advances*, pp. vbaf153, 2025.

Hu, Y., Cheng, K., He, L., Zhang, X., Jiang, B., Jiang, L., Li, C., Wang, G., Yang, Y., and Liu, M. NMR-Based Methods for Protein Analysis. *Analytical Chemistry*, 93 (4):1866–1879, February 2021. ISSN 0003-2700. doi: 10.1021/acs.analchem.0c03830. URL https://doi.org/10.1021/acs.analchem.0c03830. Publisher: American Chemical Society.

Huang, B., Kong, L., Wang, C., Ju, F., Zhang, Q., Zhu, J., Gong, T., Zhang, H., Yu, C., Zheng, W.-M., et al. Protein structure prediction: challenges, advances, and the shift of research paradigms. *Genomics, proteomics & bioinformatics*, 21(5):913–925, 2023.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2.

Kalogeropoulos, K., Bohn, M., Jenkins, D., Ledergerber, J., Sørensen, C., Hofmann, N., Wade, J., Fryer, T., Thi Tuyet Nguyen, G., auf dem Keller, U., Laustsen, A., and Jenkins, T. A comparative study of protein structure prediction tools for challenging targets: Snake venom toxins. *Toxicon*, 238:107559, February 2024. ISSN 0041-0101. doi: 10.1016/j.toxicon.2023.107559. URL https://www.sciencedirect.com/science/article/pii/S0041010123003707.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Kovalevskiy, O., Mateos-Garcia, J., and Tunyasuvunakool, K. AlphaFold two years on: Validation and impact. *Proceedings of the National Academy of Sciences*, 121 (34):e2315002121, August 2024. doi: 10.1073/pnas.2315002121. URL https://www.pnas.org/doi/10.1073/pnas.2315002121.

Kufareva, I. and Abagyan, R. Methods of protein structure comparison. *Methods in molecular biology (Clifton, N.J.)*, 857:231–257, 2012. ISSN 1064-3745. doi: 10.1007/978-1-61779-588-6_10. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4321859/.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

Liu, L., Zhang, S., Xue, Y., Ye, X., Zhu, K., Li, Y., Liu, Y., Zhao, W., Yu, H., Wu, Z., et al. Technical report of helixfold3 for biomolecular structure prediction. *arXiv preprint arXiv:2408.16975*, 2024.

Manfredi, M., Savojardo, C., Iardukhin, G., Salomoni, D., Costantini, A., Martelli, P., and Casadio, R. Alpha&ESMhFolds: A Web Server for Comparing AlphaFold2 and ESMFold Models of the Human Reference Proteome. *Journal of Molecular Biology*, 436(17):168593, September 2024. ISSN 0022-2836. doi: 10.1016/j.jmb.2024.168593. URL https://www.sciencedirect.com/science/article/pii/S0022283624001888.

Manfredi, M., Savojardo, C., Martelli, P. L., and Casadio, R. Evaluation of the structural models of the human reference proteome: Alphafold2 versus esmfold. *Current Research in Structural Biology*, pp. 100167, 2025.

Milne, J., Borgnia, M., Bartesaghi, A., Tran, E., Earl, L., Schauder, D., Lengyel, J., Pierson, J., Patwardhan, A., and Subramaniam, S. Cryo-electron microscopy: A primer for the non-microscopist. *The FEBS journal*, 280(1):28–45, January 2013. ISSN 1742-464X. doi: 10.1111/febs.12078. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3537914/.

Moult, J., Pedersen, J., Judson, R., and Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3):ii–iv, 1995. ISSN 1097-0134. doi: 10.1002/prot.340230303. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340230303.

Moussad, B., Roche, R., and Bhattacharya, D. The transformative power of transformers in protein structure prediction. *Proceedings of the National Academy of Sciences*, 120(32):e2303499120, 2023.

R Core Team, R. et al. R: A language and environment for statistical computing, 2020.

Robin, X., Haas, J., Gumienny, R., Smolinski, A., Tauriello, G., and Schwede, T. Continuous Automated Model EvaluatiOn (CAMEO)—Perspectives on the future of fully automated evaluation of structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1977–1986, 2021. ISSN 1097-0134. doi: 10.1002/prot.26213. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26213.

Rose, Y., Duarte, J., Lowe, R., Segura, J., Bi, C., Bhikadiya, C., Chen, L., Rose, A., Bittrich, S., Burley, S., and Westbrook, J. RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive. *Journal of Molecular Biology*, 433(11):166704, May 2021. ISSN 0022-2836. doi: 10.1016/j.jmb.2020.11.003. URL https://www.sciencedirect.com/science/article/pii/S0022283620306227.

Selkoe, D. Folding proteins in fatal ways. *Nature*, 426 (6968):900–904, December 2003. ISSN 1476-4687. doi: 10.1038/nature02264. URL https://www.nature.com/articles/nature02264.

Škrinjar, P., Eberhardt, J., Durairaj, J., and Schwede, T. Have protein-ligand co-folding methods moved beyond memorisation? *BioRxiv*, pp. 2025–02, 2025.

Smyth, M. and Martin, J. x Ray crystallography. *Molecular Pathology*, 53(1):8–14, February 2000. ISSN 1366-8714. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1186895/.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S., Potapenko, A., Ballard, A., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J., and Hassabis, D. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03828-1. URL https://www.nature.com/articles/s41586-021-03828-1.

Valdés-Tresanco, M., Valdés-Tresanco, M., Jiménez-Gutiérrez, D., and Moreno, E. Structural Modeling of Nanobodies: A Benchmark of State-of-the-Art Artificial Intelligence Programs. *Molecules*, 28 (10):3991, January 2023. ISSN 1420-3049. doi: 10.3390/molecules28103991. URL https://www.mdpi.com/1420-3049/28/10/3991.

Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., et al. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, pp. 2024–11, 2024.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., and Peng, J. High-resolution de novo structure prediction from primary sequence. preprint, Bioinformatics, July 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.07.21.500999.

Xu, J. and Zhang, Y. How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7): 889–895, 2010.

Xu, S., Feng, Q., Qiao, L., Wu, H., Shen, T., Cheng, Y., Zheng, S., and Sun, S. Foldbench: An all-atom benchmark for biomolecular structure prediction. *bioRxiv*, pp. 2025–05, 2025.

Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4): 702–710, 2004. ISSN 1097-0134. doi: 10.1002/prot.20264. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20264.

Zhou, F., Guo, S., Peng, X., Zhang, S., Men, C., Duan, X., Zhu, G., Wang, Z., Li, W., Mu, Y., et al. Benchmarking alphafold3-like methods for protein-peptide complex prediction. *bioRxiv*, pp. 2025–03, 2025.