Nature Communications

https://doi.org/10.1038/s41467-025-67127-3

Article in Press

Benchmarking all-atom biomolecular structure prediction with FoldBench

Received: 21 June 2025

Accepted: 19 November 2025

Cite this article as: Xu, S., Feng, Q., Qiao, L. *et al.* Benchmarking all-atom biomolecular structure prediction with FoldBench. *Nat Commun* (2025). https://doi.org/10.1038/s41467-025-67127-3

Sheng Xu, Qiantai Feng, Lifeng Qiao, Hao Wu, Tao Shen, Yu Cheng, Shuangjia Zheng & Siqi Sun

Benchmarking all-atom biomolecular structure prediction with FoldBench

Sheng Xu[1,*]
Qiantai Feng[1,*]
Lifeng Qiao[2]
Hao Wu[1]
Tao Shen[1]
Yu Cheng[3,4,†]
Shuangjia Zheng[2,5,6,†]
Siqi Sun[1,4,†]

[1] Research Institute of Intelligent Complex Systems, Fudan University, Shanghai, China
[2] School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, China
[3] Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
[4] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[5] Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai, China
[6] Lingang Laboratory, Shanghai, China
[*] Contributed equally.
[†] Correspondence: siqisun@fudan.edu.cn; shuangjia.zheng@sjtu.edu.cn; chengyu@cse.cuhk.edu.hk

Abstract

Accurate prediction of biomolecular complex structures is fundamental for understanding biological processes and rational therapeutic design. Recent advances in deep learning methods, particularly all-atom structure prediction models, have significantly expanded their capabilities to include diverse biomolecular entities, such as proteins, nucleic acids, ligands, and ions. However, comprehensive benchmarks covering multiple interaction types and molecular diversity remain scarce, limiting fair and rigorous assessment of model performance and generalizability. To address this gap, we introduce FoldBench, an extensive benchmark dataset consisting of 1,522 biological assemblies categorized into nine distinct prediction tasks. Our evaluations reveal critical performance dependencies, showing that ligand docking accuracy notably diminishes as ligand similarity to the training set decreases, a pattern similarly observed in protein-protein interaction modeling. Furthermore, antibody-antigen predictions remain particularly challenging, with current methods exhibiting failure rates exceeding 50%. Among evaluated models, AlphaFold 3 consistently demonstrates superior accuracy across the majority of tasks. In summary, our results highlight significant advancements yet reveal persistent limitations within the field, providing crucial insights and benchmarks to inform future model development and refinement.

Introduction

Accurate modeling of biomolecular complexes at atomic resolution is central to understanding molecular mechanisms and guiding rational drug discovery [1–4]. Advances in deep learning-based structural prediction methods, such as AlphaFold 2 [5], AlphaFold-Multimer [6], ESMFold [7], and RoseTTAFold2 [8], have demonstrated remarkable success in predicting the structures of protein monomers and protein complexes. Building upon these achievements, AlphaFold 3 [9] has recently been introduced to extend structural prediction capabilities beyond proteins alone, covering a broader range of biomolecules, including ligands, ions, nucleic acids, and chemically modified residues.

Despite these advancements, the training code and datasets of AlphaFold 3 are not publicly available, limiting community efforts to iterate on and extend the frameworks. In response, several open-source reproductions such as Boltz-1 [10], Protenix [11], Chai-1 [12] and HelixFold 3 [13] have emerged. However, their accuracy across different biomolecule classes has not been systematically compared because a cross-domain benchmark is still missing.

The difficulty is further increased by marked diversity among target types. Antibody-antigen interfaces, protein-ligand complexes, and nucleic acids are widely considered hard cases because of their conformational flexibility and the limited number of experimental structures available. Therefore, a broad evaluation that spans these diverse interactions is essential for mapping current capabilities and identifying priorities for future improvement.

Existing benchmarking studies [14–20] have generally addressed a single task, covered a limited range of methods and targets, and often overlooked how prediction accuracy changes with similarity between test cases and training data. Without unified dataset construction and evaluation criteria, published performances are hard to compare and may not reflect true generalization to targets.

In this work, we present FoldBench, a unified, low-homology benchmark of 1,522 biological assemblies spanning nine prediction tasks across proteins, nucleic acids, ligands, and diverse interfaces (Fig. 1). We systematically evaluate five recent all-atom predictors (AlphaFold 3, Boltz-1, Chai-1, HelixFold 3, and Protenix), quantify performance dependence on training set and ligand similarity, and map strengths and failure modes across tasks. We find that AlphaFold 3 leads overall, yet antibody-antigen complexes, allosteric protein-ligand systems, and nucleic acids remain challenging. We release all data, code, and reference results to support fair comparison and future method development.

Results

Overview of the benchmark

This benchmark aims to establish a comprehensive dataset to make fair comparisons of current all-atom structure prediction models. To prevent potential data leakage, we

collected the bioassemblies from PDB entries after 2023-01-13 (validation set cutoff of AlphaFold 3) and before 2024-11-01. Targets with high sequence or structural similarity to entries in the training set were removed (see Methods), resulting in a low-homology benchmark dataset. The curation process of the benchmark dataset is illustrated in Fig. 2A.

As shown in Fig. 2B-C, the benchmark dataset includes monomers and various interfaces, containing 334 protein monomers, 15 RNA monomers, and 14 DNA monomers, as well as 6 interface types involving 279 protein-protein, 172 antibody-antigen, 558 protein-ligand, 70 protein-RNA, 330 protein-DNA, and 51 protein-peptide interfaces. Fig. 2D-E provide the statistics of the dataset, the majority of complexes have two or fewer polymer chains, accounting for 69.8%, and most complexes have less than 1,000 tokens under AlphaFold 3's tokenization scheme.

We used OpenStructure [21] as our primary evaluation tool. For most protein-related interfaces, we applied DockQ [22] to compute the docking accuracy. The DockQ success rate is the ratio of predictions achieving a DockQ score ≥ 0.23, the threshold indicating a successful docking pose [22].

Benchmark evaluation was performed on either individual monomer chains or distinct interfaces extracted from predicted full biological assemblies. Predictions were generated using a 5×5 sampling strategy (5 seeds × 5 samples) with 10 recycles to ensure thorough conformational space exploration [9] for each model. For more details, please refer to the Methods section.

Benchmarking protein-ligand co-folding performance

Protein-ligand interactions are crucial for most pharmaceutical interventions [4,23]. Therefore, accurate prediction and characterization of these interactions are vital for modern drug discovery, facilitating the identification of small molecules that can specifically alter protein function [24]. The emergence of AlphaFold 3 has been reported to be better than classical docking tools in specified tasks[9,25]. For such co-folding methods, it is essential to assess the accuracy of both the ligand and protein structures [14]. Fig. 3A illustrates the ligand success rate for five models, based on the overall dataset (558 targets), the scores for the "unseen protein" subset (76 targets), and the scores for the "unseen ligand" set (482 targets). Here, "unseen protein" refers to proteins in the protein-ligand pairs that exhibit less than 40% sequence identity to any protein sequence in the training set. Meanwhile, "unseen ligand" indicates that, although the protein sequence from the interface has homologs in the training set, the ligand must have less than 0.50 Tanimoto similarity to any ligands in complexes containing the homologous protein.

For the entire protein-ligand set, AlphaFold 3 achieves a 64.9% success rate, surpassing the runner-up, Boltz-1, by nearly 10%. Surprisingly, success rates rise across most predictors when we focus on the "unseen protein" subset, where AlphaFold 3 reaches 69.0%. By contrast, for the "unseen ligands" set, AlphaFold 3's performance (64.3% success rate) is close to the performance on the overall dataset.

We first look into the protein modeling quality among the two sets to further investigate the performance gap between the unseen protein and the unseen ligand subset. Fig. 3D presents the ligand pocket modeling performance for both low-homology and non-low-homology proteins, evaluated using LDDT-LP (ligand pocket LDDT). All methods perform well, with scores generally above 0.8, although low-homology proteins slightly impact the scores, which aligns with the good performance on protein monomers (Supplementary Note 1). We also show that protein modeling quality (LDDT-LP) has a positive correlation with ligand quality (Fig. 3E) for both the unseen protein subset (orange) and unseen ligand subset (blue). For targets having LDDT-LP greater than 0.8, their LRMSDs typically range below 2 Å for the unseen protein subset, as we did not impose a ligand similarity limitation on this subset. The same trend has been observed in the unseen ligand subset, but the overall LRMSD is higher. It suggests that protein modeling difficulty is unlikely to be the primary bottleneck in protein-ligand interface modeling; instead, ligand similarity appears to be the more decisive limiting factor.

Fig. 3F examines the impact of ligand similarity to the training set on the modeling accuracy. The x-axis represents different ligand similarity ranges, while the y-axis shows the success rate of ligand docking. As similarity increases, the model's ability to accurately model ligands also improves, with AlphaFold 3 showing a slight performance advantage. This trend indicates that models rely substantially on memorized ligand binding modes encountered during training. In practice, this means that current all-atom predictors are effective at recapitulating known protein-ligand complexes but exhibit limited generalization to unseen ligands.

We analysed the proportion of targets with RMSD values below different thresholds, as shown in Fig. 3G, with the traditional ligand RMSD threshold of 2.0 indicated by a dashed line. It is clear that AlphaFold 3 consistently outperforms other models across various thresholds, while the performance of the other models shows no significant distinction.

Fig. 3B-C report, for each method, the fraction of targets whose lowest ligand-RMSD (LRMSD) among the 25 predictions attains the overall best value. In panel B, predictions are ordered by each model's ranking score; in panel C, they are ordered by the LRMSD against the ground truth. Under both ranking schemes, AlphaFold 3 takes the lead, providing the best prediction for roughly 40% of the targets.

Allosteric regulation occurs when a ligand binds to a site topologically distinct from the orthosteric active site, inducing conformational or dynamic changes that modulate protein function without directly competing with endogenous substrates [26]. This modulation mode often affords greater receptor subtype selectivity and reduced liability for off-target effects. Computational identification of allosteric sites is therefore critical for structure-based drug design.

As shown in Fig. 3H, PDB entry 8FNY reveals ADP bound within a distinct pocket situated on the face opposite to the kinase's orthosteric ATP-binding site [27]. In this case, most prediction results were incorrect, including both inaccurate pocket predictions and correct pocket predictions with incorrect conformations. Only two predictions from AlphaFold 3 were accurate. However, while there are successful predictions among

AlphaFold 3's 25 samples, it is hard to distinguish them by the ranking scores of the model. This ranking bias likely reflects the overrepresentation of orthosteric complexes in the training data and the high chemical similarity between ATP and ADP. Moreover, the steric constraints of this distal site—whose geometry is incompatible with an ATP [27]—underscore the necessity of incorporating non-orthosteric binding examples and refining ranking algorithms to improve recognition of allosteric interactions.

Since 8FNY has a small molecule at the orthosteric site, we aim to further investigate the model's performance on allosteric sites when no ligand is present at the orthosteric site. Consequently, we selected the popular target CDK2 (PDB: 7RWF [28]) for our study.

As illustrated in Supplementary Fig. 3, although the protein structures are correctly predicted, the ligand docking quality remains incorrect, with LRMSD mostly exceeding 10 Å. All models incorrectly placed the ligand in the active site where ATP should be positioned.

Modeling protein-protein interactions

Protein-protein interactions (PPIs) mediate almost every cellular process, from signal transduction to metabolic regulation. Structurally characterizing these interfaces is essential for understanding biological mechanisms and developing therapeutic PPI modulators [29,30].

Fig. 4A illustrates the relationship between complex LDDT and structural similarity against the training set for the initial 501 targets, before structure redundancy filtering. Prediction accuracy rises steadily as structural similarity increases, suggesting a potential structural-memory effect in the models.

To alleviate potential model memorization and to incorporate more challenging targets, we therefore constituted our protein-protein target subset exclusively from targets exhibiting low structural similarity, defined as a TM-score < 0.5 relative to complexes in the training set. After filtering out these targets, our final protein-protein subset comprises 279 protein-protein interfaces.

Fig. 4B presents the cumulative DockQ-score distributions for these 279 targets as the success rate threshold ranges from 0 to 1. AlphaFold 3 consistently leads all other models, achieving an AUC of 0.59 compared to 0.53 (Boltz-1), 0.53 (Chai-1), 0.52 (Protenix) and 0.50 (HelixFold 3). At the conventional success threshold of 0.23, AlphaFold 3 attains a 72.9% DockQ success rate—over 4 percentage points higher than the runner-up, Chai-1 (68.5%) (Supplementary Table 3).

After splitting by oligomeric state, heteromeric interfaces show higher mean DockQ scores than homomeric ones (Fig. 4C), opposite to the trend reported for AlphaFold-Multimer[6]. A likely reason is that, in heteromers, high global structural similarity to the training set does not necessarily preserve the correct subunit orientation, so docking can fail even when the complex TM-score is high (Supplementary Fig. 4).

Fig. 4D highlights AlphaFold 3's potential to model conformational changes in complexes. For the domain-swapped homodimer 8DPA [31], it not only predicts the

swapped conformation accurately but also ranks the correct dimer highest among 25 candidates. However, the other models fail to reproduce the dimeric conformation and even misidentify the docking interface. Future work should focus on developing learning strategies that extract maximal information from the limited number of experimentally determined structures exhibiting conformational changes, thereby improving the models' ability to generalize to flexible structural rearrangements.

Results on antibody-antigen complexes

Antibody-antigen interactions form the molecular foundation of adaptive immunity and represent critical targets for structural prediction due to their therapeutic implications [32,33]. Structural characterization of these complexes provides essential insights for epitope mapping, rational antibody engineering, and accelerated therapeutic development. Despite computational advances, accurate prediction remains challenging due to CDR (complementarity-determining region) variability, binding-induced conformational changes, and diverse recognition modes [19,34]. These interfaces constitute particularly difficult targets in our benchmark, highlighting a frontier where improved modeling could significantly impact drug discovery timelines and success rates.

To analyse the model performance on antibody-antigen tasks, we benchmarked the models on 172 antibody-antigen pairs, including 123 antibodies, 46 nanobodies and 3 single-chain variable fragments (scFvs). Fig. 5A shows the cumulative distribution of DockQ scores for each model. Compared to the performance on protein-protein interface, the success rate on antibody-antigen dropped to 47.9% for AlphaFold 3, while other methods exhibited over 60% failure rate. Despite the modest overall performance, AlphaFold 3 still takes the lead at most DockQ score cutoffs (AUC=0.36), significantly outperforming the second-best model (Protenix) by 0.13.

Given that AlphaFold 3 achieves enhanced performance for antibodies through extensive sampling (1,000 seeds)[9], we aimed to assess the impact of a deeper sampling space. As shown in Fig. 5B, we progressively increased the number of samples in the order of seeds, selecting the top-ranked sample by ranking score at each increment for evaluation. The success rate of AlphaFold 3 gradually increases with more samples, demonstrating the advantages of increased sampling. However, other models exhibit greater fluctuations and even declines, indicating that increasing the number of samples without robust ranking capabilities may result in ineffective sampling and lower-quality conformations. To fully leverage the benefits of increased sampling and thereby obtain the best possible sampled conformations, the ranking methodology is required to approximate "oracle" (i.e., identifying the best-scored prediction) performance, as depicted in Supplementary Fig. 5A.

Additionally, we explored the effect of increasing the number of generated samples through two primary methods. The first approach involved using different random seeds, where each seed initiated a complete, independent run of the entire pipeline, encompassing both backbone generation and the diffusion process. In contrast, the second approach consisted of a single execution of the initial backbone extractor, with its output then serving as the condition for multiple diffusion samples. The results of the

two strategies are shown in Supplementary Fig. 5B. When a ranking score was employed to identify potential successes, the two strategies did not exhibit a significant difference in performance. However, if an “oracle” selection was applied, the former strategy (full re-execution with diverse seeds) yielded markedly superior results. This suggests that while the full re-execution approach is more computationally expensive, it facilitates a broader exploration of diverse, high-quality structure predictions; yet, these better predictions are not guaranteed to be selected by the model’s ranking procedure.

The case study in Fig. 5C further illustrates the critical influence of CDR H3 loop modeling quality on antibody-antigen docking accuracy. We select the nanobody-phospholipase A2 complex (PDB: 8PIH) as an example, with the experimental structures of the antigen and nanobody shown in grey and green, respectively. Two predictions from AlphaFold 3 are highlighted: one (Sample 1, blue) exhibits near-native H3 conformation with an H3 RMSD of 0.17 Å and achieves a high DockQ score of 0.96, indicating a highly accurate binding. In contrast, another sample (Sample 2, orange) shows an incorrect H3 conformation (H3 RMSD: 3.07 Å), resulting in a poor DockQ score of 0.03. The scatter plot further quantifies this relationship: for 25 samples, DockQ scores decrease as the H3 RMSD increases. This strong anti-correlation underscores the role of precise CDR H3 loop modeling in contributing to overall docking success.

To further analyse the performance of different antibody types, the antibody-antigen interactions are split into three groups: standard antibodies (n=123), nanobodies (n=46) and scFvs (n=3). Fig. 5D shows the overall model docking performance on antibodies, measured by the DockQ success rate. Most models achieve ≤40% success; only AlphaFold 3 exceeds 40% (45.4%). The other models average around 30%, highlighting the challenges of the antibody-antigen prediction task. In terms of high-quality docking, AlphaFold 3 also has the highest fraction (13.4%), indicating its ability to make accurate predictions while ensuring high quality. In contrast, the proportion of high-quality docking in other models is lower, with Boltz-1 scoring just 0.8% for high-quality docking.

Relative to conventional antibodies, nanobodies present a more compact, single-domain binding interface. In this setting, AlphaFold 3, Boltz-1 and Protenix achieve substantially higher docking success (Fig. 5E): AlphaFold 3 reaches a 53.3% overall success rate with 33.3% of predictions classified as high quality. By contrast, Chai-1 and HelixFold 3 attain only 20.9% and 26.2%, respectively. This performance gap likely reflects the advantage of reduced inter-chain complexity for methods with accurate loop modeling and ranking routines.

Single-chain variable fragment (scFv) is a type of engineered antibody that consists of the variable regions of the heavy and light chains of an antibody connected by a short linker. Fig. 5F presents the performance on scFv in our benchmark, including 7UA2, 8PMZ, and 8R4U. In this case, AlphaFold 3 outperformed the others, successfully predicting 2 out of 3 instances, while the other models failed to make any correct predictions. This suggests that the limited training data for scFv may limit the performance of the models.

Assessment of nucleic acids

Resolving structures of nucleic acids is crucial for understanding biological processes [35]. A significant advancement made by these all-atom structure prediction models is their capability to model nucleic acids, whereas earlier approaches often required developing dedicated architectures for RNA or DNA tasks [36–38]. Our results demonstrate that nucleic acid structure prediction remains a major challenge for all current models. As shown in Fig. 6A, C, the average LDDT scores for RNA and DNA monomers typically fall within the range of 0.2 to 0.6, markedly lower than those for protein monomers (up to 0.88, Supplementary Table 3). For both tasks, AlphaFold 3 achieves the highest performance (LDDT 0.53 for DNA and 0.61 for RNA) among the five models, followed by Protenix (0.44 for DNA and 0.59 for RNA) and Chai-1 (0.46 for DNA and 0.49 for RNA). Although HelixFold 3 has a relatively high LDDT in RNA monomers (0.55), its accuracy on DNA is reduced (0.29), primarily due to the poor performance of G-quadruplexes (e.g., PDB: 8D78 and 8P6B). Boltz-1 exhibits similar performance on these structures as well.

The prediction of large or highly flexible nucleic acids poses particular difficulties. For instance, in the case of an RNA aptamer (PDB: 7ZJ4, length=374) and a non-coding RNA (PDB: 9G7C, length=224), all models fail to reconstruct the global architecture (Fig. 6B). This is evidenced by high LDDT scores (e.g., 0.68 for 7ZJ4 by AlphaFold 3) alongside large RMSD values (20 Å). This discrepancy indicates that while local elements like hairpins or regular helices can be accurately captured, the modeling of tertiary folds and long-range interactions remains largely unsolved.

Another challenging target type is G/C-rich nucleic acids (Fig. 6D). For example, PDB entry 8UTG is an RNA from the NS5 gene in the West Nile Virus genome, with a short length of 21 nucleotides. Although 8UTG is not very long, its G-Quadruplex structure still poses certain challenges for the model. The visualized cases illustrate that the model struggles to learn G/C-rich complexes, frequently with low LDDT and high RMSD.

Fig. 6E-F shows the performance on protein-DNA and protein-RNA interactions. AlphaFold 3 continues to perform the best, achieving a success rate of 79.2% in the protein-DNA task. However, the scores for protein-RNA interactions are lower than those for protein-DNA, with AlphaFold 3's score dropping to 62.3%, and all models showing a significant decrease in the proportion of high-quality predictions. Fig. 6G illustrates both successful (Protein-dsDNA) and unsuccessful (Protein-tRNA) cases of AlphaFold 3. Compared to the more structured system in protein-dsDNA complexes, the inherent complexity and flexibility of nucleic acid structures might make predictions more challenging for these structure prediction models.

The underperformance of current models on nucleic acid structure prediction primarily stems from data scarcity. As of May 2025, nucleic acid structures account for only about 8% of the 237,000 entries in the PDB, severely limiting the diversity and generalizability that models can achieve compared to protein-rich datasets. Beyond data constraints, nucleic acids—particularly RNA—display remarkable structural heterogeneity and conformational flexibility, including non-canonical base pairs, complex tertiary motifs,

and dynamic long-range interactions, all of which compound the intrinsic modeling difficulty. Furthermore, nucleic acid folding and stability depend highly on environmental factors such as ion concentration and pH [39], variables that most current prediction algorithms do not explicitly encode. The relatively higher performance on protein-DNA complexes can be attributed to the greater structural regularity of DNA [40], especially in canonical double helices, as well as to the richer training data available for DNA-binding proteins. By contrast, protein-RNA interactions remain particularly challenging due to the pronounced structural diversity [41] and limited structural data available for RNA.

Discussion

To comprehensively evaluate existing models capable of general all-atom structure prediction across various biomolecular systems, we established FoldBench. This systematic benchmark comprises 1,522 biological assemblies spanning 9 target types. The assessment of AlphaFold 3, Boltz-1, Chai-1, HelixFold 3, and Protenix across these diverse systems reveals significant advancements alongside persistent challenges within the field of structural prediction.

AlphaFold 3 consistently outperforms other models across all evaluation metrics and structural categories. Its superior abilities in monomer and interaction prediction, conformational change modeling, and ranking underscore its remarkable generalization and robustness, positioning it as the leading model. Beyond AlphaFold 3's overall lead, the remaining systems exhibit task-specific strengths. For instance, in the protein-protein task, the performance gap between AlphaFold 3 and its closest competitors, Boltz-1 and Chai-1, was not statistically significant (Supplementary Fig. 7). Furthermore, HelixFold 3 emerged as the strongest performer on protein-peptide interfaces with an 89.5% success rate (Supplementary Note 2). These patterns suggest that the choice of prediction method should be guided by the specific biological task rather than by aggregate performance rankings alone. Meanwhile, more recent systems such as Boltz-2[42] and Chai-2[43] are being developed rapidly, with incremental advances and tighter integration with design-oriented workflows (e.g., structure-sequence/ligand co-design). FoldBench serves as a consistent benchmark for tracking this progress and its implications for structure prediction.

However, inherent weaknesses persist across these models. All models exhibit a degree of memorization, an issue common in data-driven approaches. While most models perform well on tasks with sufficient data or relatively simple structures (e.g., protein monomers), their performance declines in data-scarce domains. This is particularly apparent for nucleic acids and for predicting specific sites or conformers crucial in drug discovery, such as cyclic peptides (Supplementary Note 2) and allosteric sites in protein-ligand interactions. Allosteric sites remain a major challenge; frequent pocket misidentification and pose errors may be exacerbated by the over-representation of orthosteric complexes in training data and the shallow, plastic geometry of many allosteric pockets.

Compared with protein-protein interfaces, antibody-antigen interactions remain challenging. As shown in Supplementary Fig. 8, a near-native CDR-H3 is often necessary but not sufficient; success also requires correct paratope-epitope registration and contributions from other CDRs. Accurate CDR modeling is hard because these loops are highly flexible and there are few close sequence templates, so MSA-based methods offer little guidance, and errors in CDR geometry carry over to the interface. Crucially, this flexibility undermines confidence calibration: per-sample confidence scores do not reliably identify near-native conformations. Accurate antibody-antigen prediction requires both diverse conformational sampling and reliable ranking: broader sampling increases coverage of plausible states, and calibrated confidence scores prioritize near-native conformations.

These findings also suggest potential future research directions. For the research community, expanding the scale and diversity of foundational training datasets, coupled with more active data sharing, is essential. Furthermore, exploring data synthesis techniques, such as self-distillation, could offer valuable strategies to augment available data. Regarding ranking, the development of more effective scoring functions or ranking algorithms appears crucial, with potential exploration of contemporary approaches such as reinforcement learning.

Methods

Data curation procedure

This benchmark aims to create a low-homology and comprehensive dataset for fair comparison of current all-atom biomolecular structure prediction models. To remove homology targets in the benchmark set, we first reproduced the training set of AlphaFold 3 (2021-09-30 cutoff), and constructed the benchmark dataset as follows (see also Fig. 2A).

Filtering of targets:

- All entries must be deposited in the PDB (Protein Data Bank) after 2023-01-13 (AlphaFold 3 validation cutoff), and before 2024-11-01.
- Entry must be non-NMR and with a resolution of less than 4.5 Å.
- Number of polymer chains should be less than 300.
- Number of tokens should be less than 2,560 and larger than 10 under AlphaFold 3's tokenization scheme.
- The first bioassembly of each entry is selected as the prediction target.

Filtering of bioassemblies:

- Water is removed.
- Polymer chains with less than 4 resolved residues are removed.
- Polymer chains with all unknown residues are filtered out.
- Protein chains with continuous C_α distance greater than 10 Å are removed.
- Clashing chains are filtered out. For any pair of chains sharing >30% of all non-hydrogen atoms closer than 1.7 Å, the chain with the higher clash fraction (or, if equal, the smaller atom count) is removed.

Filtering to the low-homology set. We processed targets into two types, monomer and interface (pairs of chains with minimum heavy atom (i.e., non-hydrogen) separation less than 5 Å):

- Monomers: Monomer targets are bioassemblies that contain a single polymer (protein/DNA/RNA) chain, and have less than 40% sequence identity to the training set.
- Polymer-polymer interface: If both polymers have greater than 40% sequence identity to two chains in the same complex in the training set, then this interface is filtered out. For protein-protein interfaces, we further applied Foldseek-

Multimer [44] to filter the interfaces, using a threshold of TM-score < 0.5 with the training set to exclude structurally similar ones.

- Protein-ligand interface: If the protein has greater than 40% sequence identity with a chain in a training complex and the ligand has greater than 50% Tanimoto similarity to a ligand from the same complex, then the interface is filtered out. For further filtering, we randomly selected a single interface as the target for each ligand.
- Protein-peptide interface: For interfaces between a protein and a peptide (less than 16 residues), the protein chain should have less than 40% sequence identity to the training set.

Final clustering and resolution cut-offs. The remaining low-homology assemblies were clustered (40 % identity for proteins; 100 % identity for nucleic acids and peptides). Within each cluster, the structure of the highest resolution was retained (excluding protein-ligand interfaces), subject to task-specific resolution limits:

- Protein monomers: 334 targets (7 *de novo* designed proteins were removed).
- DNA monomers: 14 targets.
- RNA monomers: 15 targets.
- Protein-ligand interfaces: 558 targets after additional ligand-quality filters (Supplementary Note 3).
- Antibody-antigen interfaces: 172 targets; resolution < 2.5 Å.
- General protein-protein interfaces: 279 targets (antibody-antigen complexes excluded); resolution < 2.0 Å.
- Protein-peptide interfaces: 51 targets; no further resolution constraint owing to limited data.
- Protein-RNA interfaces: 70 targets; resolution < 2.5 Å.
- Protein-DNA interfaces: 330 targets; resolution < 2.5 Å.

Unless noted otherwise, performance was assessed on individual monomer chains or on the specified interfaces extracted from full-assembly predictions.

Model Inference

First, we generated the JSON inputs of AlphaFold 3 by parsing the mmCIF files of the bioassemblies using the *folding_input.Input.from_mmcif* function from the AlphaFold 3 repository (https://github.com/google-deepmind/alphafold3) and removed bond information. Next, these JSON inputs were converted to other formats compatible with the different models, according to their respective input requirements and instructions.

After input generation, each model performed predictions using a 5×5 sampling strategy (5 seeds × 5 samples) with 10 recycles to ensure thorough conformational space exploration. Unless otherwise specified, the results are based on the structure with the highest ranking score for each model. Supplementary Table 2 shows the commit ID and MSA source of each method. Inference was made on NVIDIA H800 80GB GPUs.

To predict more targets in the benchmark as much as possible, some modifications were made to run each prediction model:

For AlphaFold 3, input files with sequences that have no template hits will trigger a *StopIteration* error. We fixed this problem according to the related issue #364 (https://github.com/google-deepmind/alphafold3/issues/364). Note that AlphaFold 3 and HelixFold 3 include RNA MSA searching and template searching in their local searching pipeline, which differs slightly from the other three methods.

For Boltz-1, glycans, which were formatted as multiple Chemical Component Dictionary (CCD) entries corresponding to a single chain ID—were removed from the bioassemblies, as Boltz-1 cannot process this input structure.

Note that in our assessment version of the Chai-1 model, it cannot infer biological assemblies with a token count exceeding 2048, which will hinder the evaluation of those targets.

For HelixFold 3, there are nested *ProcessPoolExecutors* in the MSA searching pipeline, which easily get stuck in our practice, potentially due to memory limitations. We replaced *ProcessPoolExecutors* into *ThreadPoolExecutors* limited by *max_workers*, which can help alleviate this issue. Additionally, since HelixFold 3 cannot handle modifications, we removed them from the input.

For Protenix, the output CIF (Crystallographic Information File) doesn't have *_entity.type* information, which OpenStructure uses to distinguish between polymer and non-polymer entities in the file, so we added this *_entity.type* value.

Evaluation and Metrics

We used the widely adopted OpenStructure v2.8 as the main assessment tool to calculate the scores between the predicted results and the ground truth. For common monomer metrics such as LDDT, TM-score, and GDT-TS, we also used OpenStructure for calculations.

For most of the protein interface measurements, we used DockQ to compare the prediction with the ground truth complexes. The DockQ success rate is the ratio of predictions achieving a DockQ score greater than or equal to 0.23, the threshold indicating a successful docking pose [22]. This score reflects the quality of the predicted interaction, with higher values suggesting a more reliable model. To provide a clearer assessment of docking results, DockQ scores can be divided into 4 bins [45]:

- Incorrect: DockQ < 0.23
- Acceptable: $0.23 \leq$ DockQ < 0.49

- Medium: $0.49 \leq \text{DockQ} < 0.80$
- High: $\text{DockQ} \geq 0.80$

The DockQ score for each interface was calculated using OpenStructure's *compare-structures*, selecting based on the two native chain IDs (*label_asym_id*). However, OpenStructure does not support the DockQ score of the protein-nucleic acid interface, so we used the DockQ v2 program[45] as our assessment tool instead in this case. After completing the above adjustments and processing, Supplementary Table 1 shows the final counts of predictable and assessable targets across various tasks and models, and the detailed performance results of various tasks and models are shown in Supplementary Table 3.

The protein-ligand interface is measured using Binding-Site Superposed, Symmetry-Corrected Pose Root Mean Square Deviation, referred to as LRMSD (ligand RMSD) [46]. Following [14], the ligand docking success rate is defined as LRMSD < 2 Å and LDDT-PLI > 0.8. The LRMSD and LDDT-PLI scores for each protein-ligand interface were calculated using OpenStructure's *compare-ligand-structures*, selecting based on the two native label chain IDs.

For antibody CDR loop assessment, we followed the evaluation pipeline outlined in [34] to calculate the CDR H3 loop RMSD. We renumbered the antibody structures using the AbNum webserver [47] according to the Chothia scheme [48], and then computed the H3 RMSD score using PyRosetta4 [49].

Data Availability

All benchmark data and reference results can be accessed at https://github.com/BEAM-Labs/FoldBench and Zenodo [50]. The PDB entries cited in this paper are: 8FNY [https://doi.org/10.2210/pdb8fny/pdb], 8DPA [https://doi.org/10.2210/pdb8dpa/pdb], 8PIH [https://doi.org/10.2210/pdb8pih/pdb], 7ZJ4 [https://doi.org/10.2210/pdb7zj4/pdb], 9G7C [https://doi.org/10.2210/pdb9g7c/pdb], 8UTG [https://doi.org/10.2210/pdb8utg/pdb], 8AYG [https://doi.org/10.2210/pdb8ayg/pdb], 8TAA [https://doi.org/10.2210/pdb8taa/pdb], 8E3R [https://doi.org/10.2210/pdb8e3r/pdb], 8JOZ [https://doi.org/10.2210/pdb8joz/pdb], 7RWF [https://doi.org/10.2210/pdb7rwf/pdb], 8D78 [https://doi.org/10.2210/pdb8d78/pdb], 8P6B [https://doi.org/10.2210/pdb8p6b/pdb], 7UA2 [https://doi.org/10.2210/pdb7ua2/pdb], 8PMZ [https://doi.org/10.2210/pdb8pmz/pdb], and 8R4U [https://doi.org/10.2210/pdb8r4u/pdb]. The full list of assessed PDB entries is available at the GitHub repository above. Source Data are provided with this paper.

Code Availability

The FoldBench evaluation code can be found at https://github.com/BEAM-Labs/FoldBench and Zenodo [50].

References

1. Kreitz, J. *et al.* Programmable protein delivery with a bacterial contractile injection system. *Nature* **616**, 357–364 (2023).

2. Lim, Y. *et al.* In silico protein interaction screening uncovers DONSON's role in replication initiation. *Science* **381**, eadi3448 (2023).

3. Mosalaganti, S. *et al.* AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science* **376**, eabm9506 (2022).

4. Durrant, J. D. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol.* **9**, 1–9 (2011).

5. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

6. Evans, R. *et al.* Protein complex prediction with AlphaFold-multimer. *bioRxiv* 2021–10 (2021).

7. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

8. Baek, M. *et al.* Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv* 2023–05 (2023).

9. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

10. Wohlwend, J. *et al.* Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv* 2024–11 (2024).

11. ByteDance AML AI4Science Team *et al.* Protenix-advancing structure prediction through a comprehensive alphafold3 reproduction. *bioRxiv* 2025–01 (2025).

12. Chai Discovery team *et al.* Chai-1: Decoding the molecular interactions of life. *bioRxiv* 2024–10 (2024).

13. Liu, L. *et al.* Technical report of HelixFold3 for biomolecular structure prediction. *arXiv Preprint arXiv:2408.16975* (2024).

14. Škrinjar, P., Eberhardt, J., Durairaj, J. & Schwede, T. Have protein-ligand co-folding methods moved beyond memorisation? *bioRxiv* 2025–02 (2025).

15. Zhai, S. *et al.* PepPCBench is a comprehensive benchmark for protein-peptide complex structure prediction with AlphaFold3. *bioRxiv* 2025–04 (2025).

16. Sachdev, S. *et al.* Evaluation of alphafold modeling for elucidation of nanobody-peptide epitope interactions. *J. Biol. Chem.* 110268 (2025).

17. Zheng, H. *et al.* AlphaFold3 in drug discovery: A comprehensive assessment of capabilities, limitations, and applications. *bioRxiv* 2025–04 (2025).

18. Bernard, C., Postic, G., Ghannay, S. & Tahi, F. Has AlphaFold 3 achieved success for RNA? *Acta Cryst. D* **81**, 49–62 (2025).

19. Yin, R., Feng, B. Y., Varshney, A. & Pierce, B. G. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci.* **31**, e4379 (2022).

20. Zhou, F. *et al.* Benchmarking AlphaFold3-like methods for protein-peptide complex prediction. *bioRxiv* (2025) doi:10.1101/2025.03.09.642277.

21. Biasini, M. *et al.* OpenStructure: An integrated software framework for computational structural biology. *Acta Cryst. D* **69**, 701–709 (2013).

22. Basu, S. & Wallner, B. DockQ: A quality measure for protein-protein docking models. *PLOS One* **11**, e0161879 (2016).

23. Schreyer, A. & Blundell, T. CREDO: A protein–ligand interaction database for drug discovery. *Chem. Biol. Drug Des.* **73**, 157–167 (2009).

24. Sousa, S. F., Fernandes, P. A. & Ramos, M. J. Protein–ligand docking: Current status and future challenges. *Proteins: Struct., Funct., Bioinf.* **65**, 15–26 (2006).

25. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).

26. Wenthur, C. J., Gentry, P. R., Mathews, T. P. & Lindsley, C. W. Drugs for allosteric sites on receptors. *Annu. Rev. Pharmacol. Toxicol.* **54**, 165–184 (2014).

27. Piserchio, A. *et al.* ADP enhances the allosteric activation of eukaryotic elongation factor 2 kinase by calmodulin. *Proc. Natl. Acad. Sci.* **120**, e2300902120 (2023).

28. Faber, E. B. *et al.* Development of allosteric and selective CDK2 inhibitors for contraception with negative cooperativity to cyclin binding. *Nat. Commun.* **14**, 3213 (2023).

29. Westermarck, J., Ivaska, J. & Corthals, G. L. Identification of protein interactions involved in cellular signaling. *Mol. Cell. Proteomics* **12**, 1752–1763 (2013).

30. Peng, X., Wang, J., Peng, W., Wu, F.-X. & Pan, Y. Protein–protein interactions: Detection, reliability assessment and applications. *Brief. Bioinform.* **18**, 798–819 (2017).

31. McCombe, C. L. *et al.* A rust-fungus Nudix hydrolase effector decaps mRNA in vitro and interferes with plant immune pathways. *New Phytol.* **239**, 222–239 (2023).

32. Chungyoun, M. F. & Gray, J. J. AI models for protein design are driving antibody engineering. *Curr. Opin. Biomed. Eng.* **28**, 100473 (2023).

33. Sela-Culang, I., Kunik, V. & Ofran, Y. The structural basis of antibody-antigen recognition. *Front. Immunol.* **4**, 302 (2013).

34. Hitawala, F. N. & Gray, J. J. What does AlphaFold3 learn about antigen and nanobody docking, and what remains unsolved? *bioRxiv* 2024–09 (2025).

35. Zhu, Y., Zhu, L., Wang, X. & Jin, H. RNA-based therapeutics: An overview and prospectus. *Cell Death Dis.* **13**, 644 (2022).

36. Pearce, R., Omenn, G. S. & Zhang, Y. De novo RNA tertiary structure prediction at atomic resolution using geometric potentials from deep learning. *bioRxiv* 2022–05 (2022).

37. Shen, T. *et al.* E2Efold-3D: End-to-end deep learning method for accurate de novo RNA 3D structure prediction. *arXiv Preprint arXiv:2207.01586* (2022).

38. Wang, W. *et al.* trRosettaRNA: Automated prediction of RNA 3D structure with transformer network. *Nat. Commun.* **14**, 7266 (2023).

39. Holbrook, S. R. RNA structure: The long and the short of it. *Curr. Opin. Struct. Biol.* **15**, 302–308 (2005).

40. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome Biol.* **1**, reviews001.1 (2000).

41. Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M. & Thornton, J. M. Protein–RNA interactions: A structural analysis. *Nucleic Acids Res.* **29**, 943–954 (2001).

42. Passaro, S. *et al.* Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv* 2025–06 (2025).

43. Chai Discovery Team *et al.* Zero-shot antibody design in a 24-well plate. *bioRxiv* (2025).

44. Kim, W. *et al.* Rapid and sensitive protein complex alignment with Foldseek-Multimer. *Nat. Methods* **22**, 469–472 (2025).

45. Mirabello, C. & Wallner, B. DockQ v2: Improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinform.* **40**, btae586 (2024).

46. Robin, X. *et al.* Assessment of protein–ligand complexes in CASP15. *Proteins: Struct., Funct., Bioinf.* **91**, 1811–1821 (2023).

47. Abhinandan, K. & Martin, A. C. Analysis and improvements to kabat and structurally correct numbering of antibody variable domains. *Mol. Immunol.* **45**, 3832–3839 (2008).

48. Chothia, C. & Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–917 (1987).

49. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: A script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinform.* **26**, 689–691 (2010).

50. Feng, Q. & XU, S. BEAM-labs/FoldBench: FoldBench v1.0. Zenodo https://doi.org/10.5281/zenodo.17180806 (2025).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62041209, to S.Z.), the Natural Science Foundation of Shanghai (24ZR1440600, to S.Z.), the Lingang Lab Fund (LGL-8888, to S.S. and S.Z.), the Key Project of the Shanghai Municipal Education Commission (2024AIZD013, to S.Z.), and the Shanghai Artificial Intelligence Laboratory (to S.S.). Some figures were created using resources from (Flaticon.com).

Author Contributions Statement

S.S., S.Z. and Y.C. conceived the study and co-supervised the work. S.X. assembled the benchmark dataset, performed the analyses, generated the figures, and drafted the manuscript. Q.F. developed the prediction and benchmarking code pipelines, performed analyses, and drafted the manuscript. S.S., S.Z., Y.C., L.Q., H.W., and T.S. provided critical feedback and revised the manuscript. All authors approved the final version.

Competing Interests Statement

The authors declare no competing interests.

Figure Legends/Captions

Fig.1. Overview of the benchmark workflow for evaluating all-atom structure prediction models. We curate a low-homology dataset of 1,522 biological assemblies from the PDB, spanning monomers and six major interaction types involving proteins, nucleic acids, and ligands. These are organized into nine distinct prediction tasks to enable broad, cross-domain assessment. Finally, the performance of models is systematically evaluated against these tasks using key metrics such as DockQ, RMSD, and LDDT to quantify prediction accuracy for both monomers and interfaces.

Fig.2. Data curation pipeline and statistical summary of the benchmark dataset. **A.** Workflow illustrating the selection and processing steps applied to biological assemblies retrieved from the Protein Data Bank (PDB). Resulting datasets comprise low-homology monomer and interface targets. **B.** Number of low-homology monomer targets categorized by biomolecular types. **C.** Number of low-homology interface targets categorized by interface types. Created in BioRender. Xu, S. (2025) https://BioRender.com/oiebbja
. **D.** Distribution of polymer chain counts across biological assemblies in the benchmark dataset (n=1,522). **E.** Distribution of token counts per biological assembly.

Fig.3. Performance on protein-ligand interaction modeling. The protein-ligand dataset can be divided into two categories: unseen proteins and unseen ligands. “unseen protein” means the protein in the protein-ligand pairs is unseen (e.g. less than 40% sequence identity to the training set), and “unseen ligand” means that we allow the protein sequence in the interface can have homologs in the training set, but the ligand should have less than 0.5 Tanimoto similarity to the ligands in the same complex where the protein homolog is in. **A.** Success rate across different subsets, where the success rate is defined as the ratio of LRMSD less than 2 Å and LDDT-PLI (Local Distance Difference Test for Protein-Ligand Interactions) > 0.8. Each bar represents the mean success rate; error bars show 95% confidence intervals, and the number of targets (n) in each subset is indicated on the x-axis. **B.** The percentage of predictions made by each model is the best (lowest LRMSD among all models’ predictions). The predicted structure is selected by ranking score (left) and **C.** best scored (right). **D.** Protein (pocket) modeling quality. Box plots show the median (center line), the interquartile range (box), and whiskers extending to the most extreme values within 1.5×interquartile range of LDDT-LP (n = 558 independent ligand targets; outliers omitted). **E.** Scatter plot of protein pocket LDDT and LRMSD. **F.** Ligand similarity vs success rate in the unseen ligand subset. Lines show the mean success rate per ligand-similarity bin; shaded bands are 95% confidence intervals. **G.** Cumulative density of ligand RMSD, varying from 0 to 10. **H.** Case study of co-folding ADP with eukaryotic elongation factor 2 kinase (PDB: 8FNY). Source data are provided as a Source Data file.

Fig.4. Model performance on protein-protein task. **A.** Complex LDDT varies with structural similarity against the training set. Lines show the mean LDDT per structure

similarity bin; shaded bands are 95% confidence intervals.**B.** Cumulative distribution of DockQ scores for 279 low-homology protein-protein targets, with numbers in parentheses denoting the area under each curve (AUC). **C.** Model performance separated by oligomeric state: heteromer (n=84) and homomer (n=195). Each bar represents the mean DockQ score; error bars show 95% confidence intervals, and the number of targets (n) in each subset is indicated on the x-axis. **D.** Case study: the AvrM14-B Nudix hydrolase effector (PDB: 8DPA, homodimeric). AlphaFold 3 sampling produced both the experimentally observed domain-swapped homodimer and a monomeric state, and ranking scores can clearly separate the two basins (scatter plot). Competing methods (right) converge exclusively on the monomeric state. Source data are provided as a Source Data file.

Fig.5. Model performance on antibody-antigen (n=172) subset. **A.** Cumulative distribution of DockQ scores for the 172 targets, with numbers in parentheses denoting the area under each curve (AUC). **B.** DockQ average success rate by number of samples. **C.** Case study: The relationship between CDR H3 loop modeling quality and docking performance. **D.** Docking performance on standard antibody-antigen subset (n=123). **E.** Docking performance on nanobody-antigen subset (n=46). **F.** DockQ score on scFv subset (n=3).

Fig.6. Summary of performance of different models on nucleic acids prediction tasks. **A.** Performance on RNA monomer tasks, detailed comparison of each target and summary, respectively. Each bar represents the mean LDDT; error bars show 95% confidence intervals. **B.** Visualization of large RNA targets (grey) and predicted structures (blue), where most of the models failed. **C.** Performance on DNA monomer tasks, detailed comparison of each target and summary, respectively. Each bar represents the mean LDDT; error bars show 95% confidence intervals. **D.** Visualization of G/C-rich nucleic acid targets (grey) and predictions (blue), which are also acknowledged as hard targets. **E.** Docking performance on protein-DNA subset (n=330). **F.** Docking performance on protein-RNA subset (n=70). **G.** Visualization of successful (protein-dsDNA) and unsuccessful (protein-tRNA) cases of AlphaFold 3.

Editorial Summary:

Accurate all-atom structure prediction is essential for biology and medicine, yet systematic benchmarks remain limited. Here, authors introduce FoldBench, a dataset of 1,522 targets across nine tasks, mapping strengths and challenges across diverse biomolecular interactions.

Peer Review Information: *Nature Communications* thanks Andrew Ward, Yogesh kalakoti and the other anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.