Melo-1: Accurate All-Atom Complex Prediction with Near-Native Interfaces and Geometric Validity

Atomelody Team
https://www.atomelody.com

October 12, 2026

Abstract

All-atom biomolecular structure prediction requires accurate modeling of complexes containing proteins, peptides, nucleic acids, and small molecules while maintaining physically plausible local geometry. However, highly flexible peptide complexes and small-molecule binding requiring rigorous pocket geometry remain significant challenges for current models. We present Melo-1, an AF3-like all-atom diffusion model designed to provide a reliable structural foundation for molecular discovery. Melo-1 establishes a new capability benchmark on these highly challenging tasks: it attains 47.9% ranked DockQ high on FoldBench peptide–protein interfaces and 41.0% on the independent PXM-22to25 set, while demonstrating superior geometric validity with 91.7% of predictions free of hard atomic overlaps. For protein–ligand complexes, it achieves 66.7% ranked success on FoldBench and 50.6% on Run-N-Pose, outperforming other contemporary open models. We additionally report performance across protein–protein, nucleic-acid, antibody–antigen, and monomer settings to provide a comprehensive capability profile, highlighting both current strengths and areas for future improvement.

1 Introduction

Interactions among biomolecules—proteins, nucleic acids, small molecules, and other cellular components—are fundamental to understanding biological mechanisms and modulating disease processes. Computational models capable of predicting these interactions at experimental-level accuracy can provide a scalable and reusable structural foundation for drug discovery.

AlphaFold 2 advanced the structure prediction of many protein monomers to near-experimental accuracy [1]. AlphaFold 3 (AF3) further extended this paradigm to joint, all-atom modeling of proteins, nucleic acids, small molecules, ions, and modified residues [2]. Consequently, structure prediction is no longer merely about producing static coordinates; it is increasingly serving as a reusable engine for epitope inference, binder design, virtual screening, and the acceleration and interpretation of experimental workflows.

Following AF3, an initial wave of AF3-like systems rapidly emerged, including Boltz-1 [3], HelixFold3 [4], Chai-1 [5], and Protenix [6]. These models largely adopted similar Pairformer and all-atom diffusion backbones and maintained the narrower pair representation width of AF3 (e.g., $c_z = 128$), primarily completing the paradigm shift from protein folding to all-atom complex prediction. Subsequently, models such as Protenix-v2 [7], OpenDDE [8], ESMFold2 [9], and SeedFold [10] pushed this frontier further by scaling up the pair representation width (increasing c_z to 256–512) and training on larger datasets. However, performance remains uneven across different interaction types. While conventional protein folding and certain protein–protein interfaces are now relatively mature, highly flexible peptide complexes and small-molecule binding requiring rigorous pocket geometry still frequently suffer from insufficient accuracy, poor generalization, and physically implausible local geometry.

Peptides represent one of the most challenging regimes. Occupying an intermediate conformational scale between small molecules and folded domains, peptides possess a large conformational space and typically exist as ensembles rather than single static structures. Whether a peptide can form a viable interaction depends on atomic-level interface complementarity and non-covalent interactions. Protein–ligand tasks similarly demand chemically defensible pocket geometry. Furthermore, high interface metrics do not guarantee structural usability, as predictions with high scores can still contain severe atomic overlaps.

In this context, we release Melo-1, the formal release of our joint all-atom complex prediction model. Figure 1 shows that on FoldBench [11] near-native peptide interfaces (DockQ ≥ 0.80 [12]), Melo-1 occupies a higher success regime. Under identical inputs and sampling budgets, Melo-1 achieves the highest ranked DockQ high rates on FoldBench (47.9%) and the independent PXM-22to25 set [7] (41.0%), while demonstrating markedly higher geometric stability (91.7% of FoldBench

predictions are free of interface and peptide-internal hard overlaps). For ligands, it attains 66.7% ranked ligand success on FoldBench and 50.6% on Run-N-Pose [13], generally outperforming other contemporary open models.

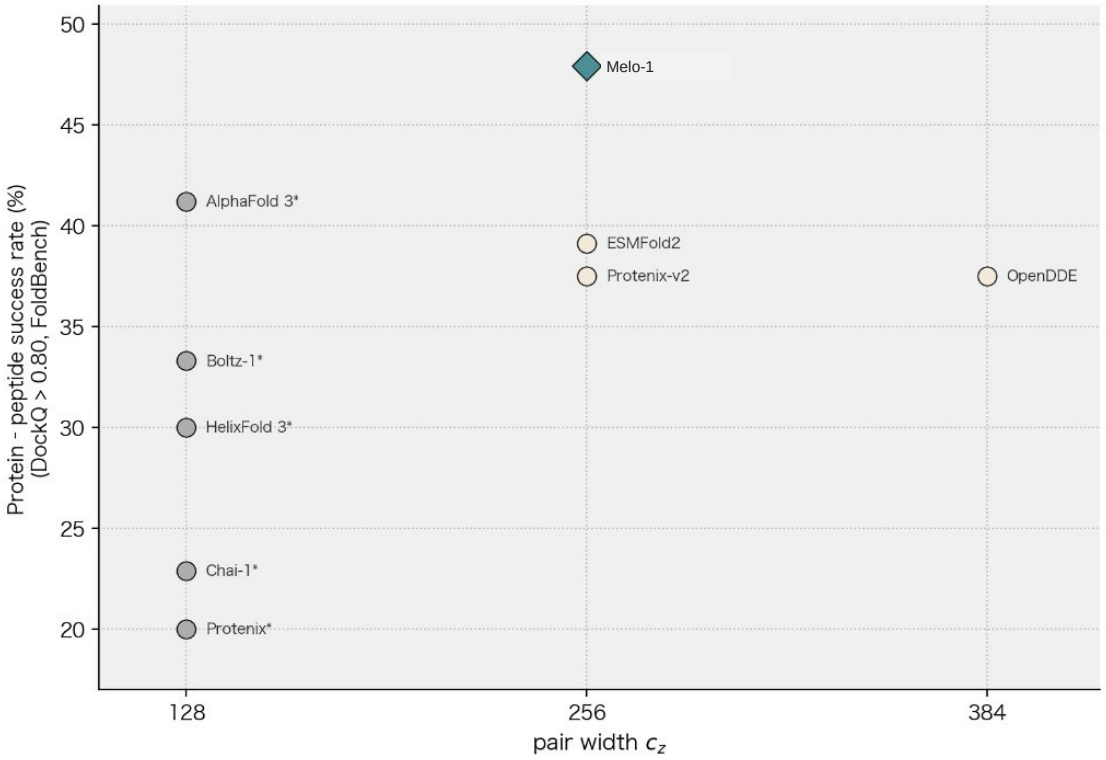

Figure 1: Protein–peptide near-native interface success rate (DockQ $\geq$ 0.80 [12]) on FoldBench [11] versus pair representation width c_z. Gray circles denote the first wave of AF3-like systems and related public references ($c_z = 128$, names marked with $*$); beige circles denote subsequent models evaluated locally; the teal diamond denotes Melo-1. Asterisks ($*$) indicate values reported in public literature or official releases; other values are from our local evaluation.

The primary contributions of this report are as follows:

1. **Establishing a new capability benchmark on highly challenging tasks such as peptides and ligands.** Melo-1 achieves a significant lead in near-native peptide interface recovery (DockQ $\geq$ 0.80), validating this advantage on the independent peptide set (PXM-22to25) [7] and ligand benchmarks (FoldBench / Run-N-Pose) [11, 13].
2. **Evaluating interface accuracy alongside geometric validity.** In addition to conventional DockQ and ligand success metrics, we systematically audit hard atomic clashes at peptide–protein interfaces and within the peptide chain. These complementary assessments characterize both interface recovery and local steric validity for model evaluation in discovery scenarios.
3. **Providing an aligned and multi-dimensional all-atom capability profile.** Under identical inputs and sampling budgets, we systematically report results for protein–protein, nucleic-acid, antibody–antigen, and monomer targets alongside peptides and ligands, fully characterizing the strengths and weaknesses of current all-atom predictions.

2 Results

2.1 Near-native peptide interfaces and ligand success

Protein–peptide and protein–ligand complexes are central scenarios for molecular discovery. Peptides require near-native pose and atomic interface complementarity beyond coarse pocket occupancy; ligands likewise need chemically defensible pocket geometry. We report primary accuracy on FoldBench [11], a low-homology all-atom benchmark spanning multiple molecular and interface types.

For peptides, current methods generally place the peptide near the pocket; the true challenge lies in precisely recovering the near-native pose and atomic-level interface details, which are critical for downstream affinity prediction or design. We therefore take the ranked high success rate (DockQ $\geq$ 0.80 [12]) as the primary peptide endpoint. Figure 2a,b summarizes this result under identical inputs and sampling budgets. Melo-1 attains a 47.9% high success rate on FoldBench and 41.0% on PXM-22to25 [7], an independent set of recent protein–peptide complexes (2022–2025) aggregated by cluster. On FoldBench this is 8.8 percentage points above ESMFold2 and 10.4 points above Protenix-v2 and OpenDDE, and it also exceeds the published AlphaFold 3 reference (41.2%). On PXM-22to25 the corresponding margins are 8.9 points above ESMFold2, 3.7 above Protenix-v2, and 5.1 above OpenDDE, indicating that the near-native peptide modeling advantage transfers to an independent recent collection.

For ligands, we evaluate on two distinct benchmarks. On FoldBench (Figure 2c), Melo-1 reaches 66.7% ranked ligand success, a limited but clear lead of 2.6–8.5 points over the same baselines, and slightly above the published AlphaFold 3 reference (64.9%). On the novelty-oriented Run-N-Pose set [13] (Figure 2d) it reaches 50.6%, essentially tied with Protenix-v2 (50.0%) and ahead of ESMFold2 (48.3%) and OpenDDE (41.4%).

2.2 Geometric validity

A high success rate does not guarantee a usable structure: even when DockQ $\geq$ 0.80, a pose can still contain severe atomic overlaps that obstruct interpretation or downstream design. We therefore treat geometric validity as a first-class result alongside Section 2.1 accuracy, auditing hard clashes—non-bonded heavy-atom pairs closer than 1.6 Å, excluding bonded and two-hop neighbours for peptide-internal contacts.

Figure 3 reports neither type of clash in panel **a**, no interface clash in panel **b**, and no peptide-internal clash in panel **c**, on FoldBench and PXM-22to25. Melo-1 has neither overlap type in 91.7% of FoldBench cases, compared with 79.2%, 68.8%, and 35.4% for Protenix-v2, ESMFold2, and OpenDDE, with interface and peptide-internal clash-free rates of 97.9% and 93.8%, respectively. The baselines fail differently. ESMFold2 is largely free of interface clashes (93.8%) but much weaker on peptide-internal geometry (70.8%), so the joint rate collapses; OpenDDE fails on both axes. Melo's joint lead therefore reflects balanced interface and peptide-internal validity rather than a single-sided metric. The independent PXM-22to25 audit (Figure 3) reproduces the same pattern: Melo leads on the joint rate (92.2%) and ties Protenix-v2 on peptide-internal clash-free predictions (97.6%), while ESMFold2 again pairs a clean interface (97.6%) with more peptide-internal clashes (84.4%). This audit uses unclustered interfaces and is a geometry diagnostic rather than a one-to-one companion to the cluster-aggregated accuracy endpoint in Figure 2b. These results establish a strong clash-free rate alongside the interface accuracy assessment in Section 2.1.

Figure 4 provides a fixed qualitative view of target 8v3s, marking hard-overlap atoms on predictions; it illustrates failure appearance rather than replacing the aggregate rates above.

2.3 Capability profile across interfaces and monomers

Beyond peptides and ligands, all-atom models must handle a diverse range of biomolecular interactions. Figure 5 broadens the evaluation scope, presenting the DockQ quality profiles (stratified into exclusive accept, medium, and high bins) of Melo-1 and Melo-1-preview across other major FoldBench interface types (protein, nucleic acid, and antibody), together with results on the PXM-22to25 antibody–antigen subset (PXMeter-AB in Protenix-v2 [7]). In the text, we primarily discuss the high success rate (DockQ $\geq$ 0.80) to remain consistent with the peptide endpoint.

On protein–protein interfaces (Figure 5a), Melo-1 reaches a 47.0% high success rate, comparable to OpenDDE (46.7%) and Protenix-v2 (44.1%) and close to the published AlphaFold 3 reference (45.9%; Boltz-1* is 33.7%). The peptide lead is therefore not an artefact of a peptide-only evaluation: conventional protein interfaces remain competitive with contemporary open models. Nucleic-acid interfaces split. On protein–RNA (Figure 5b), Melo has a lower high success rate than Protenix-v2 and OpenDDE (3.8% versus 5.8%), and absolute success remains low on this still-hard task. On protein–DNA (Figure 5c), Melo leads at 49.4% among the compared methods, including above the published AlphaFold 3 reference (45.4%). Antibody–antigen complexes (Figure 5d) are the clear gap: Melo's 24.1% high success rate trails OpenDDE (30.6%) and Protenix-v2 (30.0%), despite a substantial overall accept-or-better stack. The capability profile is therefore asymmetric—stronger on peptides, ligands, and several polymer interfaces; weaker on antibody near-native recovery. Relative to Melo-1-preview, protein–protein and antibody high success improve from 45.6% and 20.0%, respectively, DNA high success remains unchanged, and RNA high success declines from 11.5%.

On the PXM-22to25 antibody–antigen subset, Melo-1 achieves 49.4% acceptable-or-better, 40.5% medium-or-better, and 16.1% high-quality success after cluster aggregation. Published PXM total-success references are 19.6% for Boltz-1, 27.9%

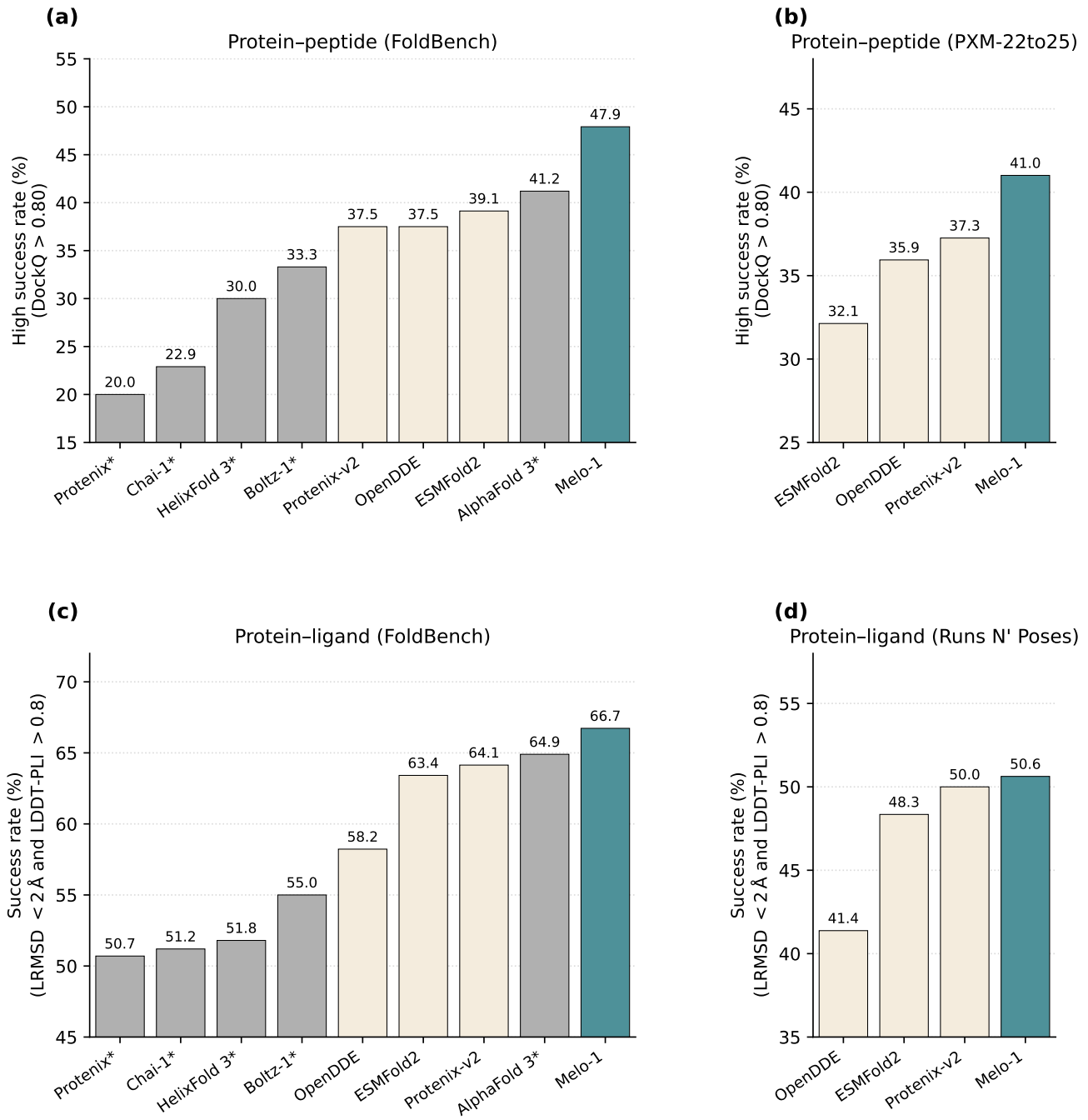

Figure 2: Primary ranked results. (**a**) Protein–peptide on FoldBench [11] (n = 48; ESMFold2 n = 46); (**b**) protein–peptide on PXM-22to25 [7] (n = 99 clusters; ESMFold2 n = 97); (**c**) protein–ligand on FoldBench (n = 541; ESMFold2 n = 533); (**d**) protein–ligand on Run-N-Pose [13] (n = 638; ESMFold2 n = 635). Gray bars marked $*$ are reference values transcribed from the FoldBench paper; beige bars are baselines evaluated under identical inputs and sampling budgets; teal denotes Melo-1. High success rate (DockQ $\geq$ 0.80 [12]) is the primary peptide endpoint; scores use PXMeter [14].

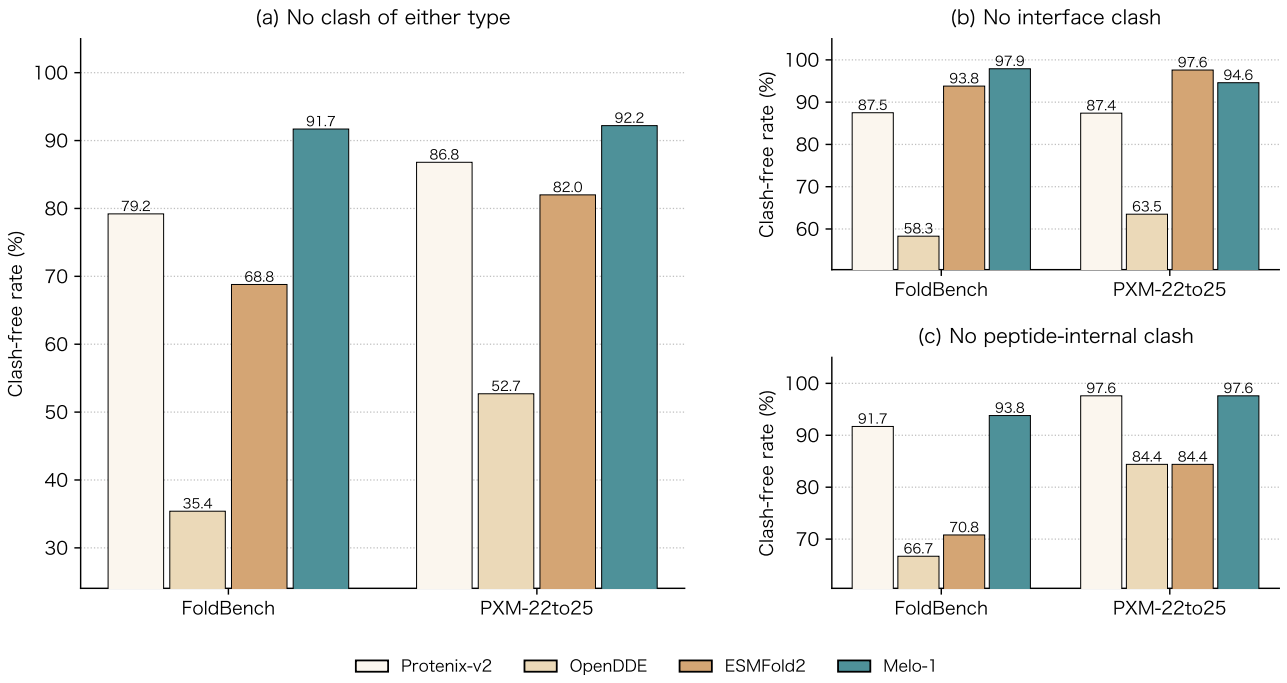

Figure 3: Geometric validity on FoldBench [11] (n = 48) and PXM-22to25 [7] (n = 167 unclustered interfaces). The main panel (**a**) reports predictions free of both clash types; the smaller panels show no interface clash (**b**) and no peptide-internal clash (**c**). Each panel compares Melo-1 and three baselines on both datasets. Melo-1 is teal; cream, tan, and brown denote Protenix-v2, OpenDDE, and ESMFold2. The PXM-22to25 geometry audit counts individual interfaces, whereas Figure 2b averages accuracy over n = 99 clusters.

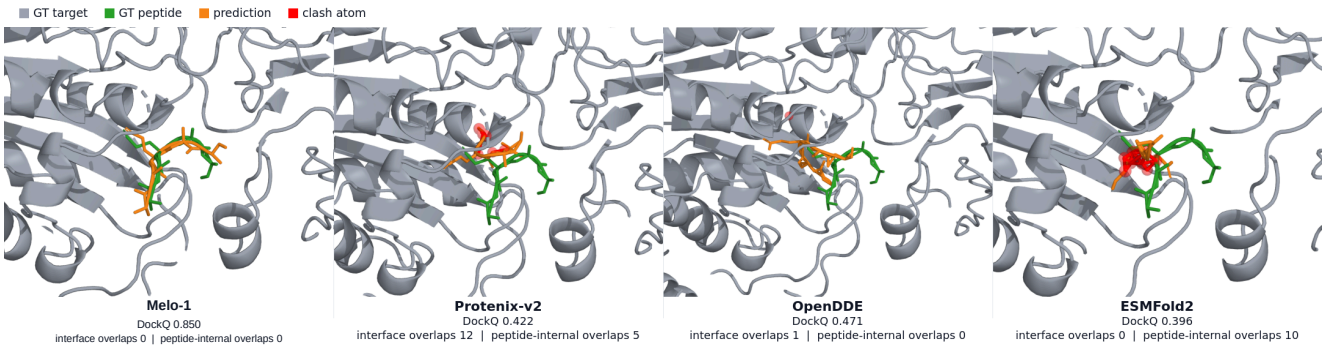

Figure 4: 8v3s diagnostic. Ground-truth target is gray, ground-truth peptide is green, prediction is orange; red spheres mark hard-overlap atoms.

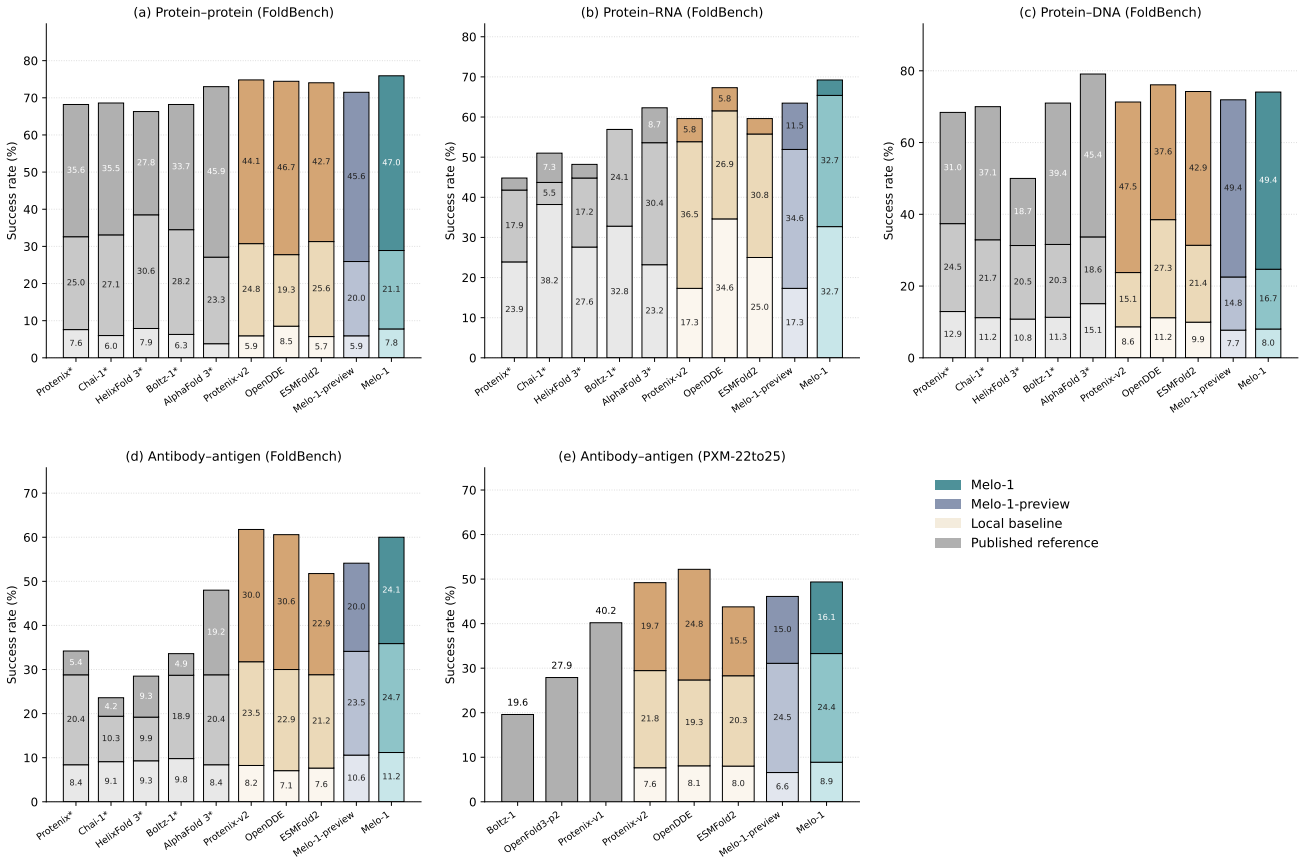

Figure 5: DockQ quality profiles across interface types on FoldBench [11] (**a–d**) and PXM-22to25 [7] (**e**). (**a**) Protein–protein (n = 270); (**b**) protein–RNA (n = 52); (**c**) protein–DNA (n = 324); (**d**) antibody–antigen (n = 170). (**e**) PXM-22to25 antibody–antigen (n = 358 clusters; ESMFold2 n = 355). Bars show exclusive percentages for accept (DockQ ≥ 0.23), medium (≥ 0.49), and high (≥ 0.80) quality bins. Gray bars with ∗ denote reference values from the FoldBench paper [11]; Melo-1 is teal and Melo-1-preview is blue-gray. Other methods were evaluated locally using PXMeter [14] under identical inputs and sampling budgets. Solid gray bars in panel **e** show total success at DockQ > 0.23 reported in Protenix-v2 Figure 2 [7], on a larger set of 516 structures in 376 clusters.

for OpenFold3-preview2, and 40.2% for Protenix-v1 [7]. These literature results cover a larger set of targets, as detailed in the figure caption.

FoldBench monomer targets complete the profile (Figure 6a–c). Protein monomers are near ceiling across methods (Melo 0.896 versus 0.894 / 0.893 for Protenix-v2 / OpenDDE) and do not differentiate systems. Separation appears for nucleic-acid monomers, where Melo reaches DNA and RNA LDDT of 0.506 and 0.695, compared with Protenix-v2 (0.511 / 0.642), OpenDDE (0.504 / 0.644), and ESMFold2 (0.429 / 0.562); sample sizes are small and reported in the figure caption, so we treat these as supporting signals rather than primary claims. Overall, the Results section supports a bounded capability picture: near-native peptide pose with usable geometry as the core gain, stable but bounded ligand improvement, competitive performance on nucleic-acid interfaces, and a remaining antibody–antigen deficit.

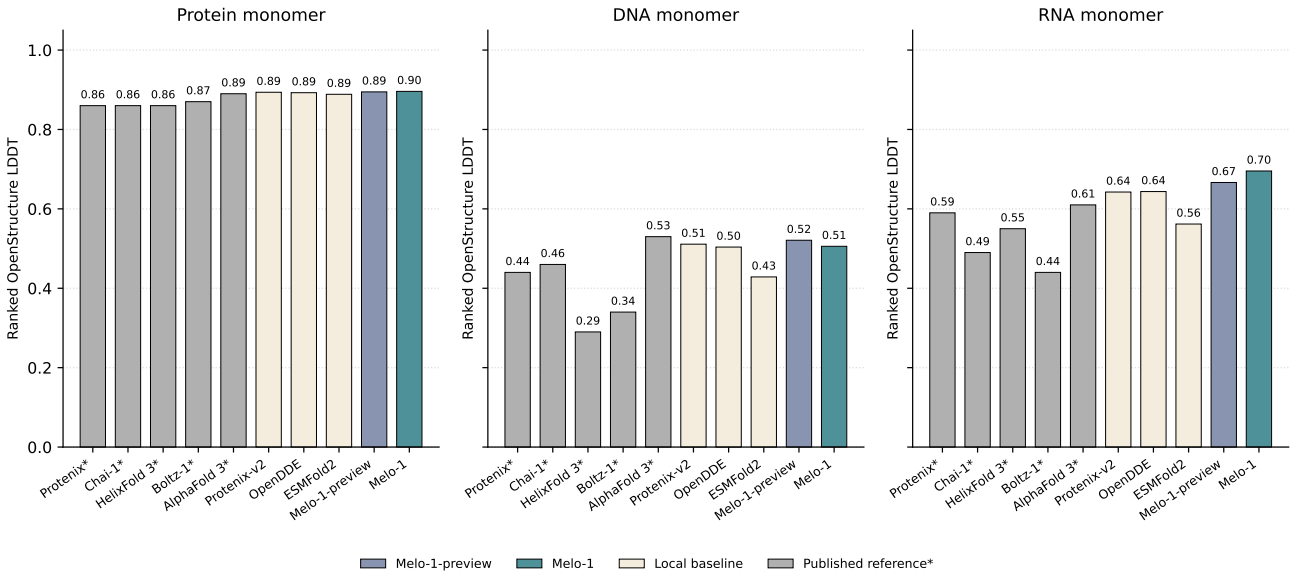

Figure 6: FoldBench [11] monomer ranked OpenStructure [15] LDDT for (**a**) protein (n = 334), (**b**) DNA (n = 14), and (**c**) RNA (n = 15). Gray bars with a trailing $*$ are transcribed from the FoldBench paper Table S3; cream bars are baselines under identical inputs and sampling budgets and teal denotes Melo-1, and blue-gray denotes Melo-1-preview. Method order matches Figure 5.

2.4 Virtual screening of designed binders

We evaluate whether predicted complex confidence can prioritize experimentally validated binders on Claude-1230 [16] and Overath [17]. Claude-1230 is derived from the Anthropic Claude Protein Binder Design Dataset [16] and contains 1,230 designs across 14 targets after excluding inconclusive MatureGDF8 measurements, designs without a final binding label, and MBP, which has no positive designs. Binding labels follow the source dataset's final assessment (`binder_final`). The Overath dataset compiles experimentally characterized de novo binder designs [17]. Our processed set contains 3,669 designs across 15 targets after removing seven incomplete rows lacking target identifiers or sequences; binding labels follow the source dataset's `binder` field. Candidates are ranked within each target separately by minimum ipSAE and binder–target ipTM. We report average precision (AP) and the enrichment factor among the top 10% of candidates (EF@10%), averaged equally over targets (Figure 7). AP summarizes ranking quality over the full score range (higher is better), whereas EF@10% measures how many times more experimental binders appear in the top 10% of the ranked list than expected by chance (EF = 1 corresponds to random screening).

On Claude-1230 and Overath, Melo-1 achieves AP of 0.558 and 0.564, respectively, with corresponding EF@10% of 2.695 and 5.212 when ranked by minimum ipSAE. On Claude-1230 and Overath, Melo-1 achieves AP of 0.538 and 0.522, respectively, with corresponding EF@10% of 2.477 and 4.896 when ranked by binder–target ipTM. These comparisons use the same predictions and experimental labels; only the ranking score changes. Point estimates do not establish statistical significance.

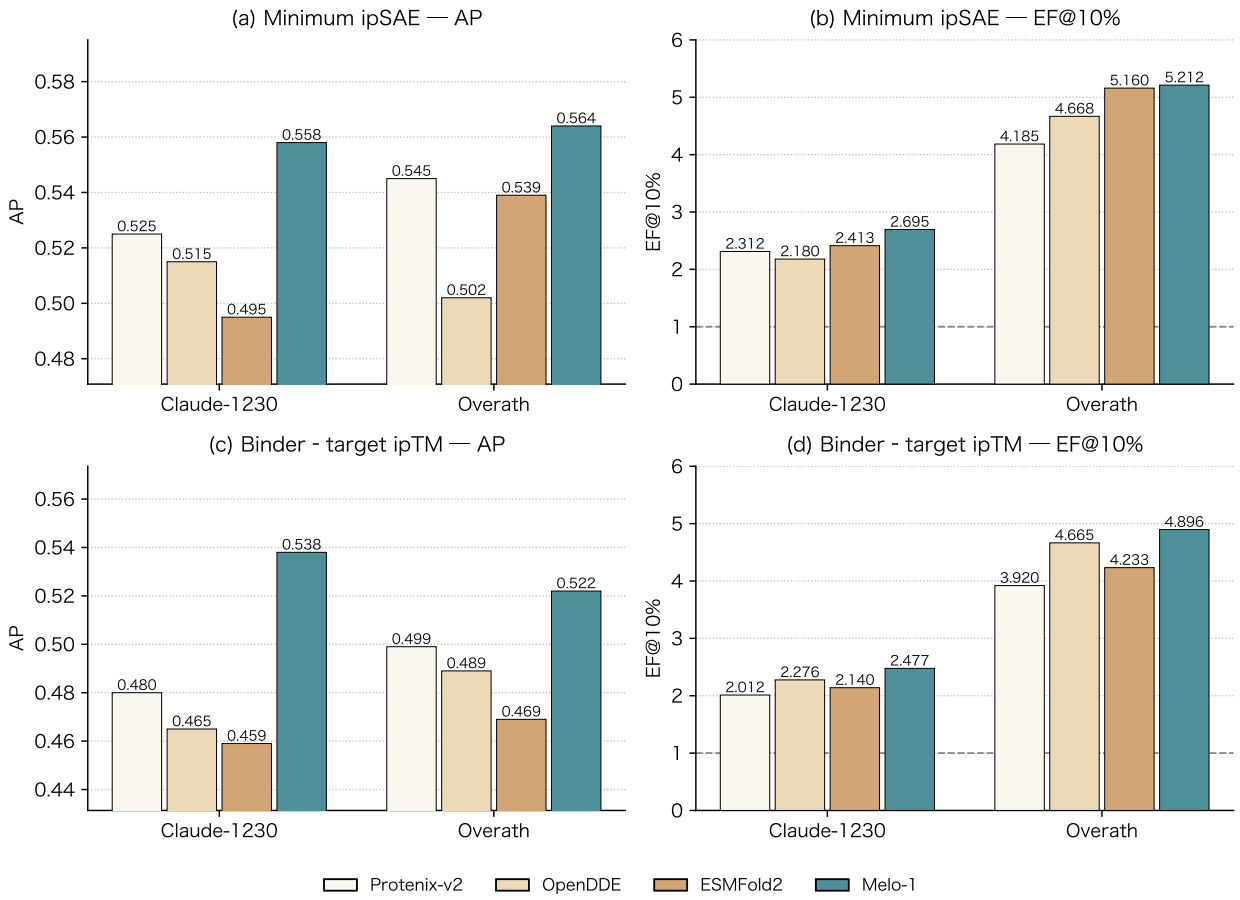

Figure 7: Virtual screening on Claude-1230 [16] and Overath [17]. Target-macro average precision (AP; **a,c**) and enrichment factor at 10% (EF@10%; **b,d**). AP summarizes ranking quality over all score thresholds; EF@10% is the fold enrichment of experimental binders in the top 10% of the ranked list relative to random selection (dashed line, EF = 1). The top row ranks candidates by minimum ipSAE; the bottom row uses binder–target ipTM. Both scores use the same single prediction per design. Melo-1 is teal; cream, tan, and orange denote Protenix-v2, OpenDDE, and ESMFold2.

3 Model and evaluation setup

3.1 Evaluation scope

Figures 1–2 evaluate peptide and ligand accuracy, and Figures 3–4 assess geometric validity; Figures 5–6 additionally include Melo-1-preview for comparisons across other molecular interfaces and monomer structures.

For the structural benchmarks, both versions use five random seeds and five samples per seed, selecting predictions by model confidence. Comparisons use the same source benchmarks; endpoint-specific sample counts reflect completed targets after availability filtering.

For virtual screening, all four models generate exactly one prediction per design, using one seed and one sample. Melo-1 and all baselines share the same target-side MSA and template assets, with binder-side MSA and templates disabled; ESMFold2 uses the shared MSAs but does not consume templates. The shared minimum-ipSAE score uses a 10 Å PAE cutoff, takes the minimum of the two directional interface predicted aligned-error scores, and does not add a coordinate-distance filter. The binder–target ipTM score averages the binder-to-target chain-pair ipTM entries over target protein chains. AP is computed from the precision–recall curve using score thresholds to handle ties; EF@10% selects $\lceil 0.1N \rceil$ candidates per target and uses the expected hit count when a tied score crosses the cutoff.

3.2 Model overview

Melo-1 follows the broad design pattern of modern all-atom diffusion predictors: sequence and multiple-sequence-alignment inputs condition single-residue and pairwise representations, which condition an all-atom coordinate denoiser and confidence estimation. Templates and RNA MSAs are used when available. Molecular components are represented jointly in one assembly rather than treating peptide or ligand docking as post-processing.

The pair representation width is $c_z = 256$, placing Melo-1 in the mid-width regime of recent AF3-like systems (AF3-era models typically use $c_z = 128$; later systems often scale to 256–512). Training uses experimental structures from the RCSB PDB [18] with a deposition cutoff of 30 September 2021, matching the AlphaFold 3 training-date alignment [2], together with antibody–antigen complexes under the same cutoff. In addition, the model is trained with long-monomer distillation structures and associated MGnify-derived MSAs from the OpenFold3 monomer distillation release [19].

3.3 Benchmarks and metrics

We use FoldBench, a low-homology all-atom benchmark [11], as the primary broad evaluation, covering protein–peptide, protein–ligand, protein–protein, protein–RNA, protein–DNA, and antibody–antigen interfaces, as well as protein, DNA, and RNA monomers. We additionally evaluate the protein–peptide and antibody–antigen subsets of PXM-22to25 [14, 7]. The peptide subset contains recent complexes from 2022–2025; accuracy is aggregated over 99 clusters, while the geometric audit counts 167 individual interfaces. The antibody–antigen subset corresponds to PXMeter-AB in the Protenix-v2 report [7]. Applying the structural evaluation scope described below yields 497 entries in 358 clusters, with 355 scored clusters for ESMFold2. Antibody–antigen success rates are aggregated by cluster; the published reference bars use the larger source cohort of 516 entries in 376 clusters. For ligand complexes, we use the clustered Run-N-Pose benchmark [13], which emphasizes pocket novelty. Following its clustered-and-distinct selection, we retain proper ligands with computable SuCOS [20] shape scores, select the representative with the lowest SuCOS shape pocket-coverage score per cluster, and require fewer than 100 training systems with similar CCDs; aggregate ranked ligand success on this filtered set is reported in Figure 2d.

Interfaces are evaluated with DockQ [12] via PXMeter [14]. We report FoldBench DockQ accept, medium, and high thresholds of ≥ 0.23, ≥ 0.49, and ≥ 0.80, respectively. The ranked high success rate (DockQ ≥ 0.80) is the primary peptide metric because current methods generally place peptides near the pocket, so near-native pose and interface detail are the quantities of interest. Protein–ligand success requires ligand RMSD below 2 Å and LDDT-PLI above 0.8. Monomer LDDT is computed with the official FoldBench evaluator and OpenStructure [15] 2.8.0. For geometric validity on peptide–protein predictions, a hard clash is a non-bonded heavy-atom pair closer than 1.6 Å at the target–peptide interface or within the peptide; directly bonded and two-hop peptide neighbours are excluded from the latter.

3.4 Baselines and inference protocol

For structural benchmarks, the local comparison includes Protenix-v2, OpenDDE, and ESMFold2. Across the local methods, we evaluate biological assemblies of up to 2,000 tokens, providing the same chain set, paired and unpaired MSA inputs when available, and relevant ligand–polymer or ligand–ligand covalent bonds. Each method performs its own featurization

and generates 25 candidates using five random seeds and five samples per seed. Reported "ranked" results select the candidate preferred by the model's native confidence score, representing normal deployment. Reported sample counts reflect the completed targets for each endpoint after availability filtering.

The baselines span two all-atom diffusion predictors and a protein-language-model-based predictor. Protenix-v2 is an AF3-style all-atom system with protein and RNA MSA and template support [7]; we used 10 recycling cycles and 200 diffusion steps. OpenDDE is an all-atom co-folding model [8]; we used its general-purpose `opendde.pt` checkpoint with 10 recycling cycles and 200 diffusion steps, disabled template-free guidance, and did not re-search MSAs. ESMFold2 derives its primary representation from the ESMC protein language model and uses iterative refinement with diffusion-ODE sampling [9]; we used the full `biohub/ESMFold2` checkpoint, 10 refinement loops, 68 sampling steps, supplied the shared MSAs, omitted templates, and ranked candidates with 0.8 ipTM + 0.2 pTM.

Table 1 summarizes public method capabilities and the configuration used for structural benchmarks.

Table 1: Baseline scope. Public capabilities describe the cited releases; the final column records the inference settings used for structural benchmarks.

Method	Public method context	Evaluation configuration
Protenix-v2	AF3-style all-atom diffusion; protein/RNA MSAs and templates supported.	10 recycles, 200 diffusion steps; native confidence ranking.
OpenDDE	All-atom co-folding; protein/RNA MSAs and templates supported.	General-purpose `opendde.pt`; 10 recycles, 200 diffusion steps; native confidence ranking.
ESMFold2	ESMC-derived protein representations; optional MSA; iterative refinement and diffusion-ODE sampling.	Full `biohub/ESMFold2`; 10 loops, 68 sampling steps; MSA on, templates off; 0.8 ipTM + 0.2 pTM ranking.

4 Discussion

Together, these results demonstrate Melo-1's broad capabilities across peptide, ligand, nucleic-acid, and protein complex prediction. In recent years, the field has increasingly relied on straightforward model scaling to drive performance improvements [7, 8, 9, 10]. While scaling is undeniably important, our results highlight that architectural efficiency and representation quality are equally critical—a philosophy consistent with prior work by members of our team in accelerating structure prediction through architectural innovation and strategic sampling [21, 22]. For instance, despite using a moderate pair width ($c_z = 256$) rather than the largest hidden dimensions among contemporary all-atom models [7, 8, 9], Melo-1 achieves strong near-native peptide recovery and highly competitive ligand success.

Crucially, we emphasize that high interface metrics alone are insufficient for practical molecular discovery. Past models have often exhibited a trade-off between maximizing accuracy scores and maintaining physically plausible local geometry. Melo-1 bridges this gap, demonstrating that near-native accuracy does not have to be purchased at the cost of severe atomic clashes. By delivering both high DockQ rates and robust geometric validity, the model provides a truly usable structural foundation for downstream workflows such as affinity prediction, binder design, and virtual screening.

Nevertheless, important limitations remain. The deficit in antibody–antigen near-native recovery, for instance, indicates that certain complex types are still open challenges. This suggests that relying solely on structural diffusion may be insufficient for the hardest tasks, pointing toward the need for stronger physical priors or targeted, large-scale domain data in future iterations. Further updates on Melo-1 and related releases will be available at `https://www.atomelody.com`.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021.

[2] Josh Abramson, Jonas Adler, Jack Dunger, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493–500, 2024.

[3] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, et al. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024.

[4] Lihang Liu, Shanzhuo Zhang, Yang Xue, et al. Technical report of helixfold3 for biomolecular structure prediction, 2024.

[5] Chai Discovery. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024.

[6] Xinshi Chen, Yuxuan Zhang, Chan Lu, et al. Protenix: Advancing structure prediction through a comprehensive alphafold3 reproduction. *bioRxiv*, 2025.

[7] ByteDance Seed. Protenix-v2: Broadening the reach of structure prediction and biomolecular design. https://github.com/bytedance/Protenix/blob/main/docs/PX2.pdf, 2026.

[8] Aureka Research. Folding, reasoning, and scaling with open-source drug discovery engine, 2026.

[9] EvolutionaryScale / Biohub. Esmfold2 protein structure prediction. https://biohub.ai/models/esmfold2, 2026.

[10] Yi Zhou, Chan Lu, Yiming Ma, et al. Seedfold: Scaling biomolecular structure prediction, 2025.

[11] Yifan Xu et al. Benchmarking all-atom biomolecular structure prediction with foldbench. *Nature Communications*, 2025.

[12] Sankar Basu and Bjorn Wallner. Dockq: a quality measure for protein-protein docking models. *PLoS ONE*, 11(8), 2016.

[13] Janani Durairaj et al. Have protein-ligand co-folding methods moved beyond memorisation? *bioRxiv*, 2025.

[14] Wenzhi Ma, Zhenyu Liu, Jincai Yang, Chan Lu, Hanyu Zhang, and Wenzhi Xiao. From dataset curation to unified evaluation: Revisiting structure prediction benchmarks with pxmeter. *bioRxiv*, 2025.

[15] Marco Biasini et al. OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D*, 69:701–709, 2013.

[16] Anthropic. Claude protein binder design dataset. https://huggingface.co/datasets/Anthropic/claude-protein-binder-design, 2026.

[17] Max D. Overath et al. Predicting experimental success in de novo binder design: A meta-analysis of 3,766 experimentally characterised binders. *bioRxiv*, 2025.

[18] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[19] The OpenFold3 Team. OpenFold3-preview. https://github.com/aqlaboratory/openfold-3, 2025.

[20] Susan Leung, Michael J. Bodkin, Frank von Delft, Paul Brennan, and Garrett M. Morris. Sucos is better than RMSD for evaluating fragment elaboration and docking poses. *ChemRxiv*, 2019.

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

[22] Lihang Liu, Yang Liu, Xianbin Ye, Shanzhuo Zhang, Yuxin Li, Kunrui Zhu, Yang Xue, Jingbo Zhou, Xiaonan Zhang, and Xiaomin Fang. Reshaping biomolecular structure prediction through strategic conformational exploration with helixfold-s1. *Nature Machine Intelligence*, 8:1115–1126, 2026.