Proteins: Structure, Function, and Bioinformatics

WILEY

PROTEINS

RESEARCH ARTICLE

Alternative Conformation Prediction Using Deep Learning With Multi-MSA Strategy and Structural Clustering in CASP16

Qiqige Wuyun[1] | Quancheng Liu[2] | Wentao Ni[3] | Chunxiang Peng[4] | Ziying Zhang[5] | Xiaogen Zhou[5] | Gang Hu[3] | Lydia Freddolino[2,4] | Wei Zheng[3]

[1]Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan, USA | [2]Gilbert S Omenn Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA | [3]NITFID, School of Statistics and Data Science, AAIS, LPMC and KLMDASR, Nankai University, Tianjin, China | [4]Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan, USA | [5]College of Information Engineering, Zhejiang University of Technology, HangZhou, Zhejiang, China

Correspondence: Lydia Freddolino (lydsf@umich.edu) | Wei Zheng (jlspzw@nankai.edu.cn)

Received: 26 May 2025 | **Revised:** 5 September 2025 | **Accepted:** 15 September 2025

Funding: This work was supported in part by the National Natural Science Foundation of China (12426303 to W.Z., 92370128, and 12326611 to G.H. and 62203389 to X.Z.), the Fundamental Research Funds for the Central Universities (054-63253109 to W.Z.), the Tianjin Science and Technology Program (24ZXZSSS00320 to G.H. and W.Z.), Fundamental Research Funds for the Provincial Universities of Zhejiang (RF-C2024006 to X.Z.), the National Institute of Allergy and Infectious Diseases (R01AI134678 and S10OD026825 to L.F.), and the National Science Foundation (MTM2025426 to L.F.).

Keywords: alternative conformation | biomolecule's structure prediction | CASP16 | deep learning | EnsembleFold | multiple sequence alignment | protein complex | protein–nucleic acid complex | structural cluster

ABSTRACT

We report the results from the "MIEnsembles-Server" and "Zheng" groups for structure ensemble predictions in CASP16, both of which employed the EnsembleFold pipeline. Initially, multiple sequence alignments (MSAs) were generated using DeepMSA2 for proteins and rMSA for RNA targets. These MSAs were processed by newly developed deep learning methods—D-I-TASSER2 for protein monomer structure prediction, DMFold2 for protein complex structure prediction, ExFold for RNA structure prediction, and DeepProtNA for protein-nucleic acid complex structure prediction—to yield diverse structural decoys. The generated decoys were clustered into representative models corresponding to distinct conformational states using the structural clustering tool MolClust. Protein monomer targets underwent additional refinement via replica-exchange Monte Carlo (REMC) simulations with D-I-TASSER2, and these refined decoys were re-clustered with MolClust to finalize the ensemble predictions. For the 19 ensemble targets in CASP16, the final EnsembleFold models achieved an average TM-score of 0.657, representing improvements of 10.2% compared to the baseline AlphaFold3 program. Notably, EnsembleFold achieved particularly good performance for hybrid protein/nucleic-acid targets, leading to its efficacy in ensemble prediction tasks. Analysis of the resulting structural ensembles highlighted three significant insights: (i) Models derived from distinct DeepMSA2-generated MSAs typically represent different conformational states for ensemble targets; (ii) REMC simulations significantly enhance model diversity, facilitating the identification of alternative conformations; (iii) The structural clustering approach effectively identifies and selects accurate representative models for each conformational state. We further discuss potential improvements in Quality Assessment (QA) scoring methods that could further enhance the reliability and accuracy of ensemble predictions in the future.

Qiqige Wuyun, Quancheng Liu, and Wentao Ni contributed equally to this study.

1 | Introduction

Protein structure prediction remains a fundamental and extensively studied challenge in structural biology. The 14th Critical Assessment of Protein Structure Prediction experiment (CASP14) represented a transformative milestone [1], primarily due to the unprecedented success of AlphaFold2 [2], an end-to-end deep learning framework capable of generating highly accurate protein models for most targets. Following CASP14, the AlphaFold2 was extended to AlphaFold2-Multimer [3] for predicting the structures of multi-chain protein–protein complexes. More recently, AlphaFold3 [4] has further extended these capabilities, enabling the prediction of a wide range of biomolecular structures—including proteins, RNA, DNA, and ligand complexes—as well as their interactions.

Despite the remarkable achievements of AlphaFold series programs in modeling diverse biomolecules, the training datasets used in developing those methods consist predominantly of static structural data from the Protein Data Bank (PDB) [5]. Consequently, the default AlphaFold series program settings alone are insufficient for reliably predicting proteins undergoing conformational transitions, such as apo/holo structural variations [6], even though many proteins naturally exist in multiple stable conformations [7]. Protein function frequently relies on the coexistence and dynamic exchange between alternative conformational states; for example, functional mechanisms such as induced fit and allosteric regulation inherently involve conformational dynamics, and pathogenic mutations can further influence occupancy among these states [8]. These complexities underscore the importance and challenge of accurately predicting multiple conformations of biomolecules—a frontier that remains unresolved in structural biology.

In CASP15, we introduced two advanced computational methodologies: D-I-TASSER for monomeric protein structure prediction and DMFold for protein complex prediction [9]. D-I-TASSER [10] integrated a deep learning-based multiple sequence alignment (MSA), a Replica Exchange Monte Carlo (REMC) simulation guided by deep learning spatial restraints and knowledge-based potentials, and a structural clustering approach to predict the tertiary structures of protein monomers. DMFold [11] utilized DeepMSA2, a novel multi-source MSA generation algorithm that leverages extensive genomic and metagenomic sequence databases to identify homologous sequences, integrated with AlphaFold2-Multimer's end-to-end modeling modules to predict the quaternary structures of protein complexes. Comparative evaluations in CASP15 demonstrated significant improvements over both the standard AlphaFold2/AlphaFold2-Multimer pipelines and other competing methodologies [9]. Building on these successes, we have since turned our attention to refining and extending these pipelines to enable the prediction of multiple conformational states of biomolecules.

Since CASP15 first introduced the structural ensemble prediction category, numerous methods have emerged and evolved rapidly, fueling active debates and developments within the field [12–14]. One such effort is AF-Cluster [12], which enhances AlphaFold2's capability to sample alternative protein conformations by clustering MSAs according to sequence similarity. This approach proved capable of predicting alternative states of known metamorphic proteins and detected novel conformational states influenced by mutations. In our CASP15 DeepMSA2 pipeline, a diverse set of MSAs can be constructed from multiple data sources and MSA generation sub-pipelines. When models generated from different MSAs show high confidence and consistency, we interpret this as evidence for a single stable conformation. In contrast, when models from distinct DeepMSA2-generated MSAs are diverge yet maintain high confidence, we hypothesized that this might represent the possibility of multiple coexisting conformational states. By generating models from separate MSAs, we have found that we can represent alternative structural conformation states within ensemble targets—an approach conceptually aligned with findings from AF-Cluster.

Beyond the multi-MSA strategy, the REMC simulations employed in D-I-TASSER inherently sample multiple low-energy conformational states by thoroughly exploring the protein energy landscape. Integrating REMC into our monomer ensemble modeling workflow substantially enhances structural diversity. This comprehensive sampling enables the identification of alternative conformations that are more representative of biologically relevant functional states. Moreover, the SPICKER clustering algorithm [15] that is incorporated into the D-I-TASSER pipeline groups decoy structures based on structural similarity and further enables the effective identification of multiple low-energy states. Notably, in our recent study on modeling the SARS-CoV-2 spike protein [10], both open and closed conformational states were captured within a single run of the D-I-TASSER REMC simulation and subsequently distinguished through SPICKER clustering. These findings highlight the potential of combining REMC simulations and structural clustering for modeling alternative conformational states of biomolecules—a direction we continue to pursue.

In CASP16, we developed the EnsembleFold pipeline (Figure 1), represented by the "MIEnsembles-Server" and "Zheng" groups, to participate in ensemble prediction challenges. Building on the robust foundations of D-I-TASSER and DMFold in CASP15, EnsembleFold incorporates several key enhancements: (i) utilization of DeepMSA2-generated MSAs, with each MSA potentially reflecting a distinct conformational state; (ii) integration of deep learning and knowledge-based potentials within an REMC simulation framework for reliable identification of representative models for each state; and (iii) application of structural clustering algorithms to separate and identify models corresponding to different states.

However, there are still potential future improvements for EnsembleFold. In particular, the development of more accurate Quality Assessment (QA) scoring methodologies could further enhance the robustness and accuracy of structural ensemble predictions and identification of stable alternative conformations.

2 | Methods

2.1 | Overview of "MIEnsembles-Server" Server and "Zheng" Human Group

The approach used by the "MIEnsembles-Server" server group was mainly designed for predicting the alternative conformations of the ensemble targets in CASP16. "MIEnsembles-Server"

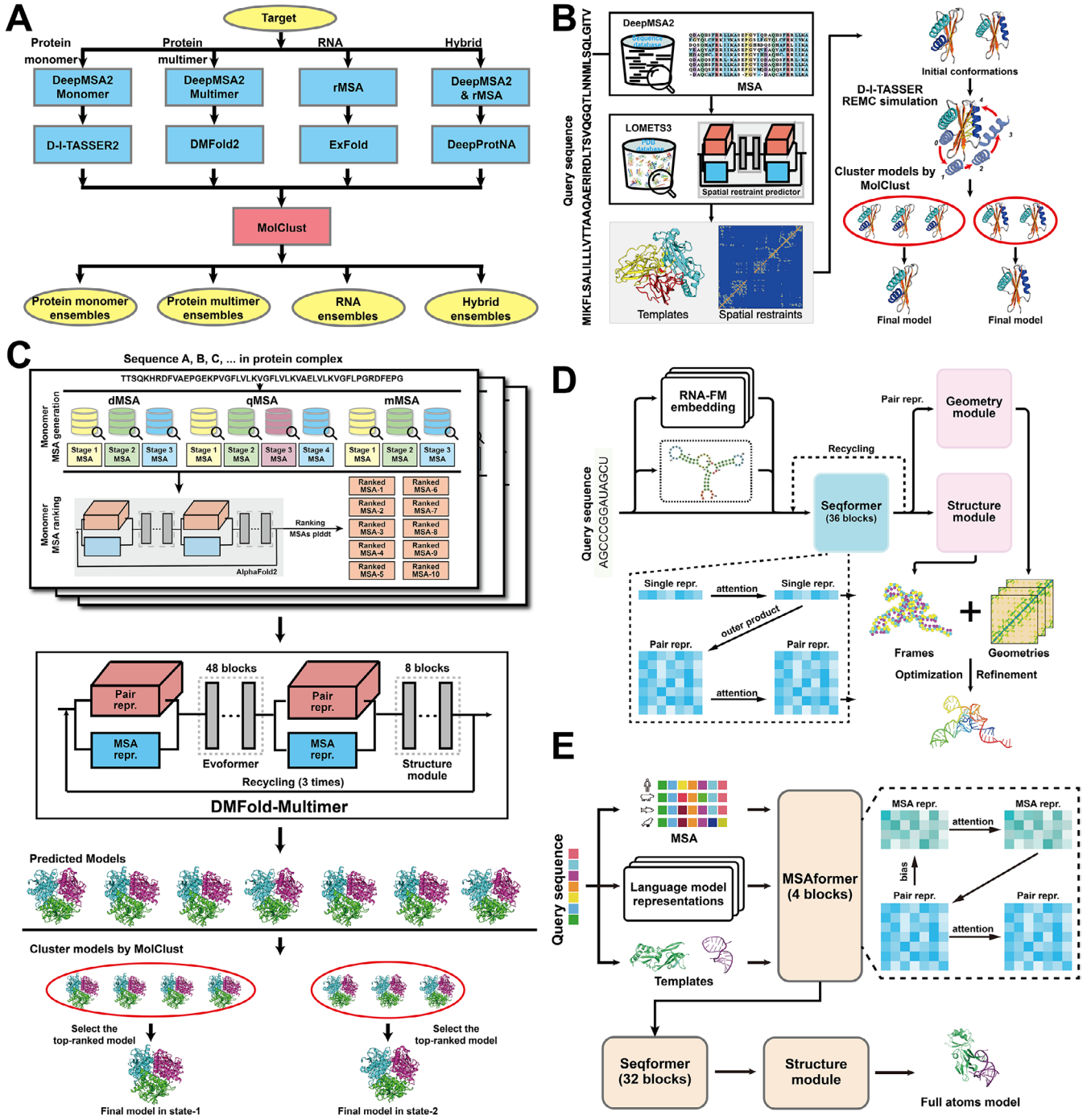

FIGURE 1 | Overview of the pipelines implemented by the "MIEnsembles-Server" server and the "Zheng" group, both of which are based on (A) the EnsembleFold framework. EnsembleFold incorporates four modeling methods: (B) D-I-TASSER2 for protein monomer structure prediction, (C) DMFold2 for protein complex structure prediction, (D) ExFold for RNA structure prediction, and (E) DeepProtNA for protein–nucleic acid complex structure prediction. DeepMSA2 and rMSA are used to generate MSAs for protein- and RNA-related targets, respectively. Multiple MSAs are employed in each method to generate diverse structural decoys, which are subsequently clustered into multi-state models using MolClust.

utilized a hybrid pipeline, called EnsembleFold, for ensemble structure prediction of protein monomers, protein complexes, RNAs, and protein–nucleic acid complexes. In general, EnsembleFold contains four pipelines for generating ensembles of candidate structures, depending on the specific target types involved (Figure 1A): (i) D-I-TASSER2 (Figure 1B) for protein monomer structure prediction that integrates multi-source deep learning spatial restraints and knowledge-based potentials, (ii) DMFold2 (Figure 1C) for protein complex structure prediction that combines the DeepMSA2 [11] MSAs and AlphaFold2 protein folding engine, (iii) ExFold (Figure 1D) for RNA structure prediction that incorporates a pre-trained language model with transformer learning, and (iv) DeepProtNA (Figure 1E) for protein-nucleic acid complex structure prediction that utilizes rMSA RNA MSAs and an AlphaFold2-like framework. The multiple ranked MSAs from DeepMSA2 were fed into those four different methods to generate multi-state decoys (appropriate to the target type in each

case), followed by the use of the structural clustering program, MolClust, for clustering and identifying decoys as multi-state models. The "Zheng" human group also submitted ensemble targets; the algorithm of "Zheng" is nearly identical to that of "MIEnsembles-Server," but with longer simulation time for protein monomers and more MSAs as input for other targets, thus providing a richer array of candidate structures and generally yielding better performance.

2.2 | DeepMSA2 For MSA Generation

The first step of all D-I-TASSER2, DMFold2, ExFold, and DeepProtNA is the construction of MSAs, which builds on the previous DeepMSA2 pipeline. Compared to the version used in CASP15, this updated DeepMSA2 introduces two key improvements: (i) a larger metagenomic sequence database, incorporating data from IMG/M, NCBI, and EBI and clustered at a 50% sequence identity threshold, and (ii) a multi-domain MSA assembly method that combines domain-level MSAs into a full chain-level MSA. The pipeline of the updated DeepMSA2 is shown in Figure S1.

Similar to DeepMSA2, the new pipeline also contains three MSA construction sub-methods: dMSA, qMSA, and mMSA. In dMSA, HHblits [16], Jackhmmer [17] and HMMsearch [17] are used to search the query sequence against the Uniclust30 [18], UniRef90 [19], and Metaclust [20] databases in three stages (labeled stage 1–3 in the order listed above). qMSA is an extended version of dMSA with a new search added between stage 2 and stage 3 of dMSA, where HHblits is used to search the BFD [20] metagenomic database. In addition, a new iteration stage (stage 4) is added in qMSA to search the query through the Mgnify [21] metagenomic database. In mMSA, MSA from qMSA stage 3 is used as the starting point for HMMsearch to search through the huge in-house metagenome database mentioned above. MSAs generated from dMSA, qMSA, and mMSA are input into AlphaFold2 (1-embedding) to predict a set of models. Those MSAs are then ranked by the associated *pLDDT* scores from AlphaFold2.

For multi-domain targets, the same MSA generation method is applied to construct domain-level MSAs based on the predicted domain boundaries. These domain-level MSAs are subsequently assembled into full-length MSAs by linking sequences from the same species. The ranked MSAs are either used directly in protein monomer modeling or paired as multimer MSAs for protein complex modeling.

For heteromeric complexes, an additional selection procedure is employed to generate an optimal set of paired MSAs by combining individual constituent MSAs. The top N ranked MSAs for each constituent protein are chosen to form potential paired MSAs. Each selected MSA for one constituent protein is paired with the MSA of another constituent. For a heteromeric complex containing M different constituent proteins, N^M distinct paired MSAs are generated and evaluated based on a combined score of the depth of the MSAs and pLDDT score of the monomer chains. To guarantee the pipeline could be completed within 3 days, N is set as the maximal value to satisfy $N^M \leq 64$.

2.3 | D-I-TASSER2 for Protein Monomer Ensemble Targets

D-I-TASSER2 is an extended protein monomer structure prediction pipeline based on our CASP15 D-I-TASSER framework. In the D-I-TASSER2 pipeline (Figure 1B), the final monomer MSAs from DeepMSA2 are used as input for AlphaFold2 [2] (8-embedding), OpenFold [22], UniFold [23], ColabFold [24], RosettaFold [25], ESMFold [26], OmegaFold [27], DeepPotential [28], and AttentionPotential (an extension of DeepPotential utilizing an MSA transformer architecture) for the predictions of residue-residue contact maps, distance distributions, inter-residue torsion angles, and hydrogen-bonding networks. Those deep learning–predicted restraints are utilized to guide the REMC folding simulation with the same set of restraints calculated from templates detected by LOMETS3. The full sets of predicted restraints from AttentionPotential and DeepPotential are then fed into DeepFold, an L-BFGS folding system, to produce 10 full-length models. These 10 generated 10 models, as well as five models generated by each of the other structure prediction methods noted above and full-chain level threading templates from the LOMETS3, are used as initial conformations in the REMC folding simulation.

The MSA generated from the DeepMSA2 method is also used to produce sequence profiles or profile Hidden Markov Models (HMMs) to be utilized by the six profile-based threading methods employed by the new version of LOMETS3. Additionally, the contact maps and distance distributions predicted by the deep learning predictors are used by five contact- and distance-based threading methods. In addition to six profile-based threading methods [29–34] and five distance-based threading methods [35–38], we also added three pLM-based threading methods [39–41] to a new development version of LOMETS3. Finally, 140 full-chain level templates (10 templates from each component threading method) are collected by LOMETS3 and then used as initial conformations in the REMC simulation as noted above.

For protein monomer targets, an I-TASSER-based REMC simulation is utilized for generating structural decoy conformations. The REMC simulation is guided by knowledge-based potentials and residue-residue contact maps, distance distributions, inter-residue torsion angles, and hydrogen-bond networks that are predicted by deep learning predictors and calculated from LOMETS3 threading templates. Finally, 10,000 decoy conformations will be produced in this step.

2.4 | DMFold2 For Protein Complex Ensemble Targets

DMFold2 is a protein complex structure prediction pipeline extended from our DMFold method as used in CASP15.

In DMFold2 (Figure 1C), the step after MSA construction from DeepMSA2 is template detection, which is based on a new version of LOMETS (version 4). Compared to LOMETS3 [42], which was used in CASP15, the major update in LOMETS4 is its ability to handle protein complexes. Specifically, for protein heteromers, templates are identified as follows: First, for each constituent chain in the target complex, homologous templates

are identified using LOMETS3 (see Section 2.3). Notably, for each pipeline, templates for individual chains that have already been considered in previous steps are excluded to prevent the similar query constituent chain from hitting identical templates. Next, the templates for each chain are ranked based on their quality (i.e., Z score). Finally, if at least two constituent chains share templates that originate from the same protein complex, those complexes are considered potential complex templates. For protein monomers or homomeric complexes, LOMETS4 monomer templates are directly output and used in the structure model generation step.

In the model generation step, DMFold2 utilizes a modified version of the AlphaFold2 modeling engine. The MSAs from DeepMSA2 and the structure templates from LOMETS4 serve as input features for this modeling engine. Key modifications to the AlphaFold2 modeling engine include: (i) parallel runs with and without templates, (ii) adjusting the dropout rate, (iii) applying different versions of AlphaFold2 pre-trained weights (v1–v3), (iv) generating a higher number of decoys than the default setting (25 models), (v) parallel runs applying or omitting the early stop strategy in AlphaFold2 (v2.3), and (vi) increasing the number of modeling iterations. The final models are ranked by the *pLDDT* score for monomer targets, or by QA scores ($0.2pTM + 0.8ipTM$) for complex targets.

2.5 | ExFold For RNA Ensemble Targets

The RNA structure is predicted by ExFold (Figure 1D), a deep learning-based RNA tertiary structure prediction framework that is developed upon the foundation of DRfold [43]. To construct the initial single-sequence representation, ExFold utilizes pre-trained representations from RNA-FM [44], a large-scale RNA language model, to encode contextual information from single RNA sequences. To complement sequence-level features, pairwise features are derived from the final-layer token embeddings and attention maps of RNA-FM along with predicted secondary structures to construct the pairwise representation. This integrated input enables the model to capture both local base-pairing patterns and long-range dependencies. Inspired by AlphaFold2 [2], ExFold adopts an architecture containing 36 blocks of transformer-based modules, similar to the Evoformer, to facilitate information exchange between single and pair representations. Since RNA structure prediction is performed on a single sequence without using MSAs, column-wise self-attention is omitted. The refined representations are subsequently processed by two parallel modules: a geometry module, which predicts inter-nucleotide geometric restraints, including distances and torsion angles, and a structure module, which outputs coarse-grained 3D coordinates of key atoms (the phosphate P, ribose C4′, and glycosidic N atoms of the nucleobase). Both modules operate in parallel, sharing the same input features and functioning within a single unified model.

To further refine the structure prediction, ExFold incorporates additional supervision signals. Specialized loss functions focusing on base-pairing consistency are employed to better guide the learning of RNA secondary structure features. Given the limited number of experimentally resolved RNA tertiary structures, a self-distillation dataset was conducted based on RNA sequences and secondary structure annotations from the bpRNA [45] database to enhance training. Finally, both the predicted coarse-grained model and the geometric potentials are jointly utilized to guide subsequent RNA structure reconstruction simulations. To produce multi-state ensemble models, all MSAs from rMSA were fed into ExFold.

2.6 | DeepProtNA For Protein–Nucleic Acid Complex Ensemble Targets

The protein–nucleic acid complex structure prediction of EnsembleFold is based on DeepProtNA (Figure 1E), an end-to-end deep learning algorithm that integrates multiple sources of features, including pretrained language model representations, MSAs, and structural templates, to directly generate atomic coordinates through a combination of attention-based networks and a structure module.

To capture rich contextual features from the input sequences, DeepProtNA employs large-scale language models—ESM [26] for proteins and RNA-FM [44] for RNAs. For each biomolecule, embeddings and attention maps are extracted from the final layers. MSAs are generated for protein and nucleic acid sequences using the modified version DeepMSA2 [11] and rMSA [46], respectively. In addition to sequence information, structural templates are retrieved to provide homologous structural context: for proteins, template structures are identified using LOMETS4 (see above), while RNA templates are obtained using BLASTn. The initial MSA representation is constructed by integrating information from both the primary sequence and its MSA. Pairwise features are constructed by combining contextual embeddings and attention-based signals from pretrained language models, along with structural information from templates. All these pairwise features are restricted to encoding intra-chain information. Feature extraction is performed in two stages, drawing inspiration from DeepFoldRNA [47], using four MSAformer blocks followed by 32 Seqformer blocks. In the first stage, the MSA and pair representations are updated through a simplified Evoformer-style module. The single-sequence representation is then extracted from the first row of the MSA representation and further refined alongside the pair representation using a similar attention module that omits column-wise self-attention. The resulting representations are passed to a structure decoder, which directly generates three-dimensional coordinates for each residue or nucleotide. In addition to the predicted structure, DeepProtNA provides QA scores that reflect the reliability of the modeled interactions.

2.7 | MolClust For Clustering Biomolecular Ensembles

For ensemble targets, the MolClust method, which is adapted from SPICKER [15] and US-align [48], is used to cluster the decoys based on structural similarity for protein monomers, RNAs, protein–protein complexes, and protein-nucleic acid complexes. In the MolClust method, the major structural clustering algorithm used in CASP16 is following the SPICKER pipeline, which performs a one-step clustering using a reduced, representative set of decoys and iteratively determines

the pairwise RMSD cutoff through a self-adjusting procedure. The structural similarity matrix used in the main clustering framework was calculated by US-align, which is a universal structural alignment platform that enables accurate and efficient alignment of protein, RNA, and DNA monomers and complexes by optimizing a unified TM-score objective function with a heuristic alignment searching algorithm. In the MolClust framework, the TM-score was used as the clustering threshold, rather than RMSD, to maintain consistency with US-align's alignment metrics. The centers of clusters with the largest number of members and highest QA scores are selected as representative models for potential alternative conformations.

2.8 | Runtime of EnsembleFold

To evaluate computational cost, we benchmarked the runtime for our pipeline against the standalone version of AlphaFold3, as the runtime of the AF3 server cannot be directly measured and the number of models generated by the CASP16 "AF3-server" group is unknown. Unlike AlphaFold3, which produces a limited set of models, our pipeline generates substantially larger ensembles (from a few hundred to several thousand per target; Table S1), thereby increasing both conformational exploration and computational burden. Thus, for each target, we calculated the time required to generate ten models for both methods. As shown in Figure S2, EnsembleFold consistently required 10%–20% longer runtimes than AlphaFold3 (200–1400 s vs. 150–1200 s per 10 models), reflecting its deeper conformational sampling. While this introduces additional computational expense, it also yields broader conformational ensembles that proved necessary for difficult targets. These results highlight a key trade-off between computational cost and conformational coverage.

3 | Results

According to the official definitions provided by CASP16, a total of 24 targets were classified as ensemble targets. Of these, 21 targets with known stoichiometry were released during Phase 1, including 11 protein-only targets (T1214, T1249v1, T1249v2, T1228v1, T1228v2, T1239v1, T1239v2, T1294v1, T1294v2, T1200, and T1300), six RNA-only targets (R1203, R1253v1, R1253v2, R1283v1, R1283v2, and R1283v3), and four protein–nucleic acid complex targets (M1228v1, M1228v2, M1239v1, and M1239v2). As the experimental structures for T1200 and T1300 were not available, performance evaluations were conducted on the remaining 19 targets.

3.1 | Overall Performance of the EnsembleFold in CASP16

Currently, there is no universally accepted standard for evaluating ensemble structure predictions. In this study, we primarily assessed prediction accuracy using the TM-score against each reference conformation (see below). However, due to the inherent complexity of ensemble targets, target-specific considerations and processes were necessary for robust evaluation.

In detail, the 19 ensemble targets were classified into five categories according to the nature of their conformational states and release formats. *Category* I includes M1228v1/v2, T1228v1/v2, M1239v1/v2, T1239v1/v2, and T1294v1/v2, where both conformational states of each ensemble were released as individual targets sharing identical stoichiometry. *Category* II consists of T1249v1/v2 and R1253v1/v2, where both conformational states of each ensemble were also released as individual targets sharing identical stoichiometry; however, in these cases, participants were explicitly informed of the structural features distinguishing each conformational state. *Category* III includes R1283v1/v2/v3, where each alternative state was released as a separate target with different stoichiometries. *Category* IV contains only T1214, in which the alternative conformation arises from ligand binding; for this target, participants were required to predict only the ligand-bound conformation. *Category* V consists only of R1203, which presents two alternative conformational states under a single target entry; for this target, participants were permitted to predict and submit either conformational state.

Given the characteristics of these categories, we adopted the following TM-score evaluation strategies. For *Category* II, *Category* III, and *Category* IV, each conformational state was treated as an independent target, with the experimental structure of each state directly compared to the corresponding predictions, following the standard protocol used for single-state targets. For *Category* I, participants were not informed of the structural features distinguishing the v1 and v2 structures. Consequently, submissions from different groups could have the v1 and v2 labels swapped relative to the experimental structures. To address this ambiguity, all submitted models for v1 and v2 were jointly evaluated to account for potential state swapping. Specifically, TM-scores were computed for all possible pairings between predicted and experimental structures (i.e., prediction v1 vs. experimental v1, prediction v2 vs. experimental v2, prediction v1 vs. experimental v2, prediction v2 vs. experimental v1). The optimal mapping was then identified to maximize the total TM-score under the constraint that each prediction set (v1 or v2) was mapped exclusively to one experimental structure state. For *Category* V (R1203), in which two conformational states were provided under a single target entry, predictions were mapped to experimental structure states in a one-to-one manner to maximize the total TM-score. The final evaluation score was defined as the average TM-score of the optimal mappings.

The average TM-scores of the best-scoring models among the submitted models for 123 servers and human groups participating in CASP16 are listed in Figure 2A. Notably, human groups accounted for seven of the top 10 positions, including all of the top four. The "Zheng" group achieved the highest overall performance among 123 groups (36 server groups and 87 human groups), with an average TM-score of 0.657 (Table S2). Among automated servers, "MIEnsembles-Server" delivered the best performance, with a TM-score of 0.610 (Table S2).

The "AF3-server" group, representing a control run of AlphaFold3, ranked eighth with a TM-score of 0.596. Figure 2B,C shows the resulting head-to-head comparisons between our groups and "AF3-server." Overall, the "Zheng" group generated models for ensemble targets with average TM-scores of 0.657, which were 10.2% better than the models

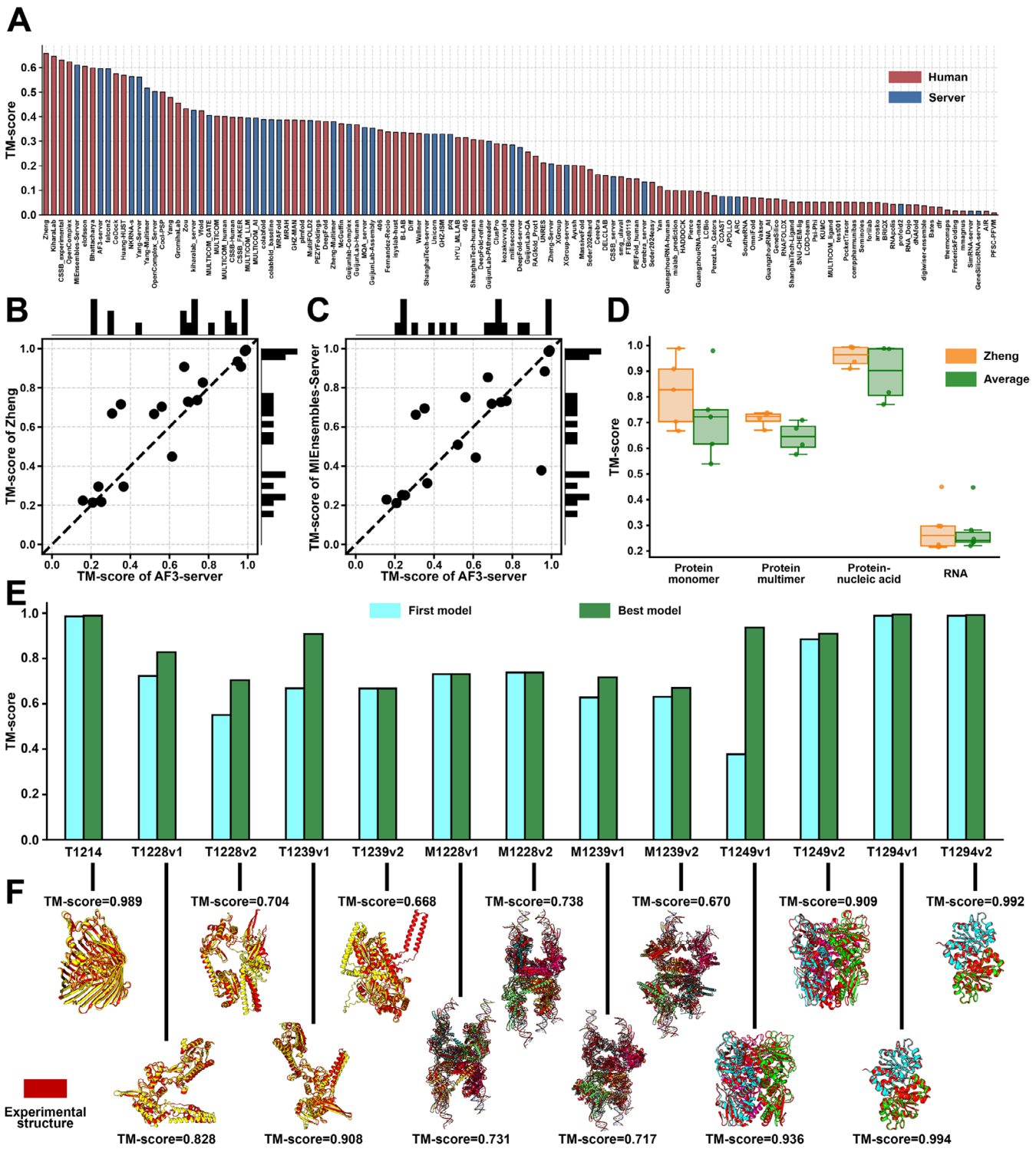

FIGURE 2 | Overall performance of the "MIEnsembles-Server" server and "Zheng" human group in CASP16. (A) Average TM-scores of the best-scoring models among all submissions from 123 participating servers and human groups in CASP16. (B) Head-to-head comparison of TM-scores between the "Zheng" group and the "AF3-server." (C) Head-to-head comparison of TM-scores between the "MIEnsembles-Server" and the "AF3-server." (D) Comparison of average TM-scores for the "Zheng" group versus all other groups across four categories of ensemble targets. (E) TM-scores of the first model and the best model from the "Zheng" group for 13 protein-related ensemble targets in CASP16. (F) "Zheng" group's best models superposed to corresponding experimental structures on 13 CASP16 protein-related ensemble targets.

generated by "AF3-server," with a p value of 3.41E-02 (Table S3). "MIEnsembles-Server" also outperformed this baseline by approximately 2.3%. Furthermore, the "Zheng" group and "MIEnsembles-Server" achieved relatively higher TM-scores on 68.4% (=13/19) and 57.9% (=11/19) of the 19 ensemble targets, respectively, compared to "AF3-server."

To further investigate the reasons underlying the high ranking of the "Zheng" group, we analyzed prediction performance across four ensemble target categories: monomeric proteins, multimeric proteins, protein–nucleic acid complexes, and RNA-only targets. Figure 2D summarizes the "Zheng" group's performance across different ensemble target types. For protein-only targets, the "Zheng" group achieved good results, with average TM-scores of 0.819 and 0.958 (Table S4) for monomeric and multimeric proteins, which are 13.5% and 7.5% higher than those of the average over all other groups, respectively. A moderate decrease was observed for protein–nucleic acid complexes with an average TM-score of 0.714 for the "Zheng" group, which is still 10.8% higher than that for the average over all other groups. In contrast, RNA-only targets proved particularly challenging, with the "Zheng" group achieving an average TM-score of 0.284. This trend was consistent across all groups, as reflected by the overall average TM-score of 0.278 for RNA-only targets, highlighting the limitations of current modeling methods for these structures. These results demonstrate that the "Zheng" group's performance on protein-related targets was markedly higher than the overall average, whereas their predictions for RNA-only targets were only marginally better than the field average.

Further insights into the "Zheng" group's performance on 13 protein-related ensemble targets, evaluated on both the first models and the best models, is illustrated in Figure 2E. The "Zheng" group achieved average TM-scores of 0.736 for the first models and 0.830 for the best models (Table S5). Regarding the best-scoring models, the "Zheng" group constructed correct folds (TM-score >0.5) for all targets. Moreover, 53.8% (7/13) of these predictions matched the experimental structures with TM-scores exceeding 0.8. A comprehensive overview of all best-scoring models is provided in Figure 2F.

3.2 | Contributions of Multi-MSA, REMC Simulation, and Structural Clustering to the Performance of EnsembleFold

To elucidate the reasons underlying EnsembleFold's strong performance in CASP16, we performed detailed case-by-case analyses of representative ensemble targets. As ensemble structure prediction is a relatively new and rapidly developing research area, the number of available ensemble targets remains limited. In CASP16, a total of 19 ensemble targets were provided, encompassing a diverse range of target types, including proteins, RNAs, and protein–nucleic acid complexes. Given this diversity, our analyses focused on examining representative cases across these target categories to better understand the strengths of the EnsembleFold approach.

T1214 is a protein consisting of 677 amino acids, adopting a beta fold architecture capable of binding a pyrroloquinoline quinone ligand. The primary structural distinction between its apo and holo forms is localized to a specific beta strand, highlighted in blue (apo) and red (holo) in Figure 3A. For T1214, threading retrieved only apo homologs; the top-ranked templates were apo structures (PDB ID: 6v81), which still have a TM-score of 0.953 to the experimental structure of T1214. Figure 3B displays the distribution of TM-scores for decoy structures aligned to the apo and holo states, generated during the initial AI-based stage using DMFold (i.e., without any REMC simulations). Remarkably, the decoys spontaneously segregate into three clusters (Figure 3C). Upon tracing the origins of these clusters, we found that each was derived from a distinct MSA input: qMSA stage 3 (Neff = 390.3), dMSA stage 1 (Neff = 860.5), and qMSA stage 2 (Neff = 106.9) (Figures S3 and S4). The corresponding models achieved TM-scores of 0.947, 0.976, and 0.973 for Clusters 1, 2, and 3, respectively. This finding suggests that MSA diversity directly contributes to the sampling of alternative conformational states.

To further investigate the role of MSA diversity in sampling alternative conformational states, we applied two complementary modeling strategies for T1214. In the first strategy, MSA subsampling was performed by adjusting the parameters *max_extra_msa* and *max_msa_clusters* in DMFold. Specifically, *max_extra_msa* was varied across eight values [16, 32, 64, 128, 256, 512, 1024, 5120], while *max_msa_clusters* was set to half of the chosen depth and limited to 512. A total of 800 models were generated (100 per setting), and the five top-ranked models were selected for analysis. This approach also yielded highly accurate predictions, with TM-scores of 0.986 for the first model and 0.991 for the best model. Focusing on individual subsampling depths, we observed that reducing the MSA size from the default 5120 increased the likelihood of generating models closer to the holo state. This effect peaked at a subsampling size of 256. Larger depths limited sampling variability, while smaller depths introduced greater structural variability and reduced model accuracy (Figure S5). In the second strategy, MSA masking was applied by replacing columns with "X" in the β-strand of interest, producing an additional ensemble of 800 models. This approach also yielded accurate final models, with TM-scores of 0.986 and 0.988 for the top-ranked and best models respectively (Figure S6). This demonstrates that local sequence perturbation can also influence conformational sampling.

However, after we included REMC simulation step used in D-I-TASSER2, the decoy ensemble expanded to include a new, prominent cluster, as shown in Figure 3D. Notably, the models in this newly formed cluster were significantly more similar to the holo state compared to the original AI-generated models, even though the ligand itself was not included during modeling (Figure 3E). EnsembleFold generated more accurate final models with a TM-score of 0.986 for the first model and 0.989 for the best model. Furthermore, when the corresponding ligand was added to the best EnsembleFold model, the ligand fit well within the binding site without steric clashes (Figure 3F). This result underscores the ability of REMC sampling, guided by both deep learning and knowledge-based potentials, to explore alternative conformations beyond those directly encoded by the input MSAs. This case study highlights two critical insights: first, MSAs constructed by DeepMSA2 can capture alternative structural states in ensemble targets; and second, REMC simulation is essential for expanding conformational diversity and identifying biologically relevant models. Together, these components are pivotal for the accurate modeling of protein ensembles even in the absence of ligand information.

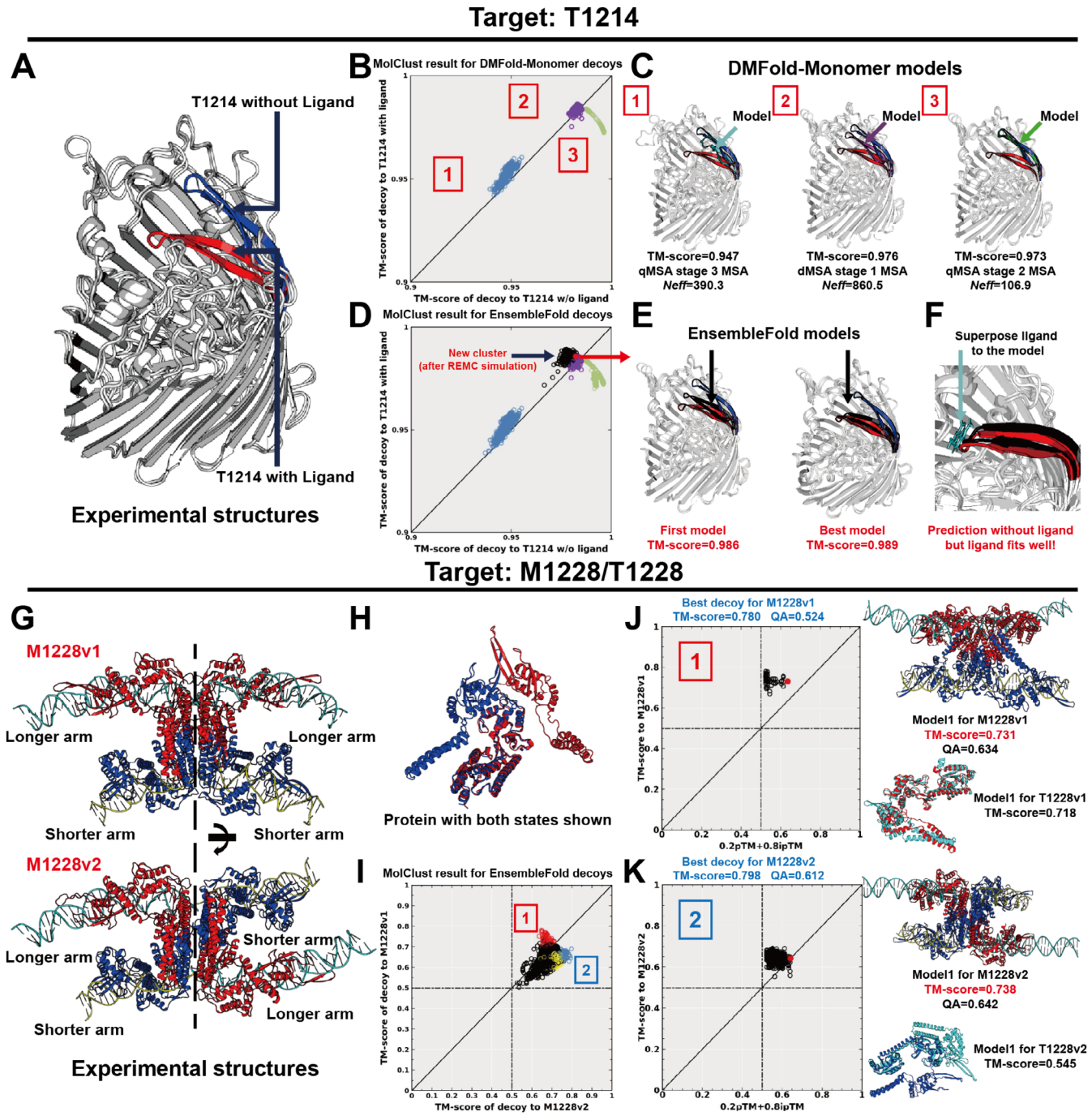

FIGURE 3 | Case studies of T1214 and M1228/T1228 to illustrate the factors contributing to EnsembleFold's strong performance in CASP16. Panels (A–F) present the case study for T1214, while panels (G–K) detail the case study for M1228/T1228. (A) Experimental structure of T1214. (B) MolClust clustering result of DMFold decoys for T1214. (C) DMFold models corresponding to three representative clusters for T1214. (D) MolClust clustering result of EnsembleFold decoys for T1214. (E) The first and best-scoring EnsembleFold models for T1214. (F) Local β-strand segment with superposed ligand in the EnsembleFold model for T1214. (G) Experimental structures for the two distinct conformational states of M1228. (H) Experimental structure for monomer T1239. (I) MolClust clustering result of EnsembleFold decoys for M1228. (J) The best-scoring decoy for M1228v1 and its corresponding EnsembleFold model. (K) The best-scoring decoy for M1228v2 and its corresponding EnsembleFold model.

To assess the influence of nearest structural neighbors, we evaluated the structural accuracy of both threading templates and the best available homology structures for each target. The two sets of templates yielded highly similar average TM-scores (0.50 for threading vs. 0.51 for homology) for 9 protein-related targets (Figure S7), suggesting no significant advantage of one over the other. In contrast, our pipeline consistently produced models with higher TM-scores than either template-derived baseline for every Ensemble target, with particularly notable improvements for M1228 and M1239, and an overall average TM-score of 0.88. These results indicate that the accuracy of our method is not primarily determined by template quality, but instead by the integration of DeepMSA2 alignments and REMC refinement.

Notably, although hybrid targets—comprising both protein and nucleic acid components—posed significantly greater challenges compared to protein-only targets, EnsembleFold demonstrated exceptional performance in this category (Figure S8). For example, "MIEnsembles-Server" achieved the highest TM-score of 0.748 among all 55 participating groups (18 server groups and 37 human groups) on the 18 hybrid targets in CASP16 (Table S6). The ability to accurately model these hybrid targets substantially contributed to EnsembleFold's strong overall performance on ensemble targets.

An illustrative case is the DNA–protein ensemble target M1228, which presents two states (v1 and v2). Here, the protein component adopts two distinct forms: one binding to a longer DNA arm and the other to a shorter DNA arm. In state v1, both long DNA arms are positioned on the same side, whereas in v2, they twist to opposite orientations (Figure 3G,H). Application of EnsembleFold to this hybrid target resulted in a TM-score distribution of predicted models aligned to both experimental states (Figure 3I). Clustering of the predicted models revealed two dominant clusters, with the highest QA scores of 0.634 and 0.642, each accurately corresponding to the two experimental conformations and achieving TM-scores of 0.731 and 0.738 for the v1 and v2 states, respectively (Figure 3J,K). Interestingly, for the protein subcomponent T1228, the models were automatically parsed from the hybrid target predictions, and these derived models for T1228 also demonstrated strong performance. These results demonstrate the capacity

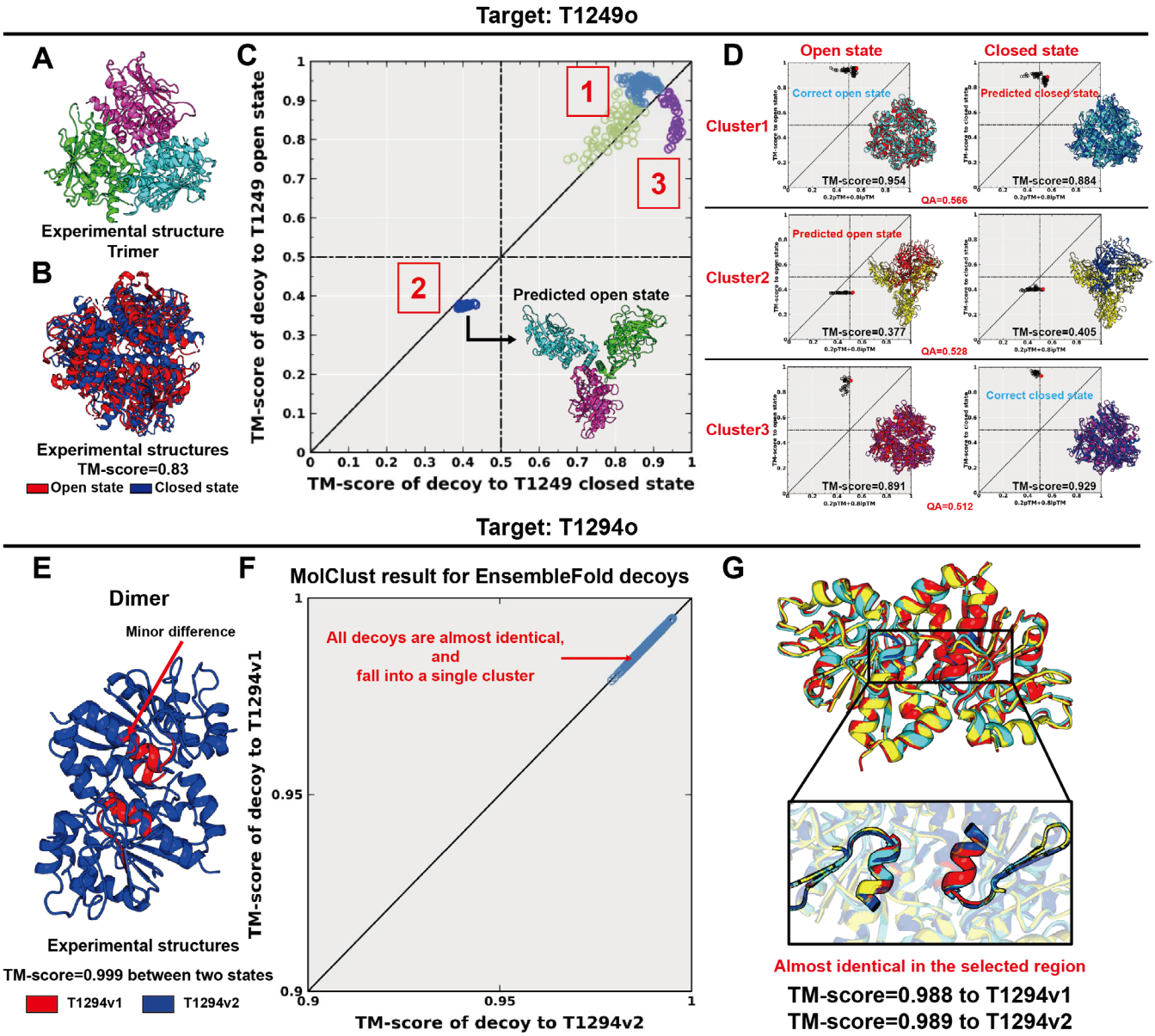

FIGURE 4 | Case studies of T1249o and T1294o to illustrate the challenges and problems in protein ensemble prediction by EnsembleFold. Panels (A–D) present the case study for T1249o, and panels (E–G) detail the case study for T1294o. (A) Experimental structure of the T1249o trimer. (B) Open and closed conformational states of T1249o. (C) Clustering of decoys for T1249o. (D) EnsembleFold models mapping to the open and closed states for the top three clusters of T1249o. (E) Experimental structure of the T1294o dimer. (F) MolClust clustering result of EnsembleFold decoys for T1294o. (G) The best EnsembleFold model superposed with the two experimental structures of T1294o, highlighting the subtle conformational shift between those targets.

of EnsembleFold to accurately generate and distinguish models for hybrid ensemble targets comprising both protein and nucleic acid components. The structural clustering algorithm played a crucial role in correctly identifying the representative model for each state.

A similar case is presented by M1239/T1239 (Figure S9). Compared to M1228 and T1228, these targets share similar structural characteristics but involve larger and more structurally complex protein monomers. Despite this increased complexity, the modeling workflow applied to M1239 and T1239 closely parallels the approach described above for M1228/T1228, and the prediction results for M1239 and T1239 remained similarly robust.

Overall, our findings indicate that three methodological advances are particularly critical for the accurate and robust modeling of alternative states in challenging ensemble targets: (i) the utilization of MSAs generated by DeepMSA2, where each MSA may reflect a distinct conformational state; (ii) the integration of deep learning and knowledge-based potentials within a REMC simulation framework, enabling correct identification of representative models for each state; and (iii) the application of structural clustering algorithms to identify and separate models corresponding to different states.

3.3 | What Went Wrong in the Current Version of EnsembleFold?

Although EnsembleFold demonstrated strong performance in modeling protein ensemble targets, there are still some areas that need to be improved (Figure 4). In particular, the current QA scoring function may fail to identify the optimal model when conformational differences between states are subtle and highly localized, reflecting a general and ongoing challenge in the field of developing highly accurate QA methods.

A clear example can be observed in the case of T1249, a trimeric protein complex (Figure 4A) exhibiting two distinct conformational states: an open state and a closed state (Figure 4B). After applying structural clustering to the EnsembleFold decoys, four major clusters were identified (Figure 4C). The top three clusters—colored light blue, blue, and purple—were ranked according to the highest QA scores: 0.566 for Cluster 1, 0.528 for Cluster 2, and 0.512 for Cluster 3. Thus, model selection was performed by choosing the top-ranked models from Cluster 1 and Cluster 2. Notably, Cluster 2 was predicted as representing the open state due to its more open structural conformation; however, it did not correspond well to the experimentally determined open conformation, with a TM-score of only 0.377. Further analysis revealed that the most accurate models matching the experimental open and closed states were actually located in Cluster 1 and Cluster 3, achieving TM-scores of 0.954 and 0.929 for the open and closed states, respectively (Figure 4D). This misclassification was primarily due to the limitations of the current QA scoring function, which led to the selection of suboptimal models despite the presence of accurate ones in other clusters. This example underscores a critical area for improvement: while EnsembleFold is capable of generating accurate models and the clustering approach effectively separates distinct conformational states, the current QA scoring scheme may fail to identify the optimal model for selection. Enhancing the accuracy of QA scores is therefore essential for improving final model selection in future versions of EnsembleFold.

Another class of challenge is exemplified by the case of T1294, a dimeric protein complex with two conformational states that differ only by subtle structural changes localized to a small region (Figure 4E). Structural comparison of the experimental models shows a TM-score of 0.999 between the two states, reflecting nearly identical overall conformations. EnsembleFold generated high-quality decoys for T1294; however, due to the minimal differences between the two experimental structure states, all decoys fell into a single structural cluster during MolClust analysis (Figure 4F). As a result, none of the predicted models captured the subtle local variation present between the two states, although predicted models have high TM-scores of 0.988 and 0.989 for the v1 and v2 states, respectively (Figure 4G). This limitation highlights a fundamental challenge in ensemble modeling: current clustering and model selection approaches are insufficiently sensitive to minor conformational changes, especially when these differences are highly localized.

4 | Conclusions

We present the results and analyses of the EnsembleFold pipeline, which was used in the CASP16 experiment to generate the "MIEnsembles-Server" and "Zheng" group results for structural ensemble prediction. Our results demonstrate that EnsembleFold, through the integration of multi-source deep learning predictions, advanced MSA generation, REMC-based conformational sampling, and structural clustering, achieved strong performance across diverse ensemble targets, including proteins, RNAs, and protein–nucleic acid complexes. Notably, EnsembleFold consistently outperformed the AlphaFold3 baseline and established itself among the top-performing methodologies for ensemble prediction, particularly excelling on complex hybrid targets relative to other participating methods.

Based on analysis of the CASP16 targets, we highlighted three key methodological advances that contributed to EnsembleFold's performance. First, leveraging multiple DeepMSA2-generated MSAs proved critical in capturing alternative conformational states, with each MSA often reflecting a distinct structural ensemble. Second, the incorporation of REMC simulations, guided by both deep learning–predicted restraints and knowledge-based potentials, substantially broadened the sampling of conformational space, enabling more accurate modeling of biologically relevant alternative states. Third, the application of MolClust for structural clustering facilitated effective identification and selection of representative models for each conformational state, ensuring that the resulting ensembles accurately reflected the underlying diversity present in the experimental data. The iterative MSA generation step and the REMC simulation are more critical components for EnsembleFold: the former provides deep coevolutionary information that underlies accurate inter-residue

restraints, while the latter ensures conformational sampling and refinement to escape template bias and further explore possible conformations.

Importantly, we also found that when MSAs are sufficiently deep, variations among alternative MSAs can approximate a form of sampling, since each provides slightly different co-evolutionary constraints. Furthermore, the use of different sequence databases inherently increases MSA diversity, which can further shape the conformational ensembles obtained. However, when sequence depth is limited, this approximation breaks down—MSA variations cannot mimic true sampling, and the variability largely reflects loss of signal. Thus, the generation of distinct conformers is not random "sampling of structure space," but reflects how MSA diversity, database choice, and perturbation reshape the restraints provided to the model.

Case studies further highlighted our pipeline's ability to model and distinguish states in challenging hybrid complexes. However, our investigations also revealed critical limitations. Chief among these limitations is the dependence on the current QA scoring schemes, which in some cases led to suboptimal model selection despite the presence of highly accurate models within the predicted ensembles. Additionally, the detection and modeling of minor, localized conformational differences—such as those observed in T1294—remains a significant challenge for both clustering algorithms and model selection criteria.

Addressing the limitations identified here represents an important direction for future research. In particular, the development of more sensitive and context-aware QA scoring functions, as well as clustering algorithms capable of adaptively resolving fine-grained conformational heterogeneity, will be essential to further enhance the robustness and accuracy of ensemble structure prediction. Moreover, improved strategies for modeling RNA and hybrid targets, which continue to pose substantial difficulties across all methods, are needed to advance the field.

Author Contributions

Qiqige Wuyun: writing – original draft, methodology, investigation, data curation, visualization. **Quancheng Liu:** methodology, validation, data curation, writing – review and editing. **Wentao Ni:** methodology, visualization, writing – review and editing, writing – original draft, data curation. **Chunxiang Peng:** data curation, methodology. **Ziying Zhang:** data curation. **Xiaogen Zhou:** data curation, writing – review and editing. **Gang Hu:** writing – review and editing, methodology, funding acquisition, resources, data curation. **Lydia Freddolino:** funding acquisition, data curation, investigation, resources, writing – review and editing, project administration. **Wei Zheng:** funding acquisition, investigation, writing – original draft, writing – review and editing, methodology, data curation, resources, visualization, project administration, supervision.

Acknowledgments

The authors thank Dr. Pengshuo Yang for discussion and assistance. The authors are grateful to Jonathan Poisson, Brock Palen, and the staff of the University of Michigan Advanced Research Computing team for IT support.

Conflicts of Interest

L.F. is a scientific advisory board member and paid consultant for CircNova Inc.; however, CircNova played no role in the funding, design, performance, or interpretation of the present work, and the present manuscript is not directly used by or related to CircNova's ongoing work.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Peer Review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/prot.70059.

References

1. A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult, "Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XIV," *Proteins: Structure, Function, and Bioinformatics* 89, no. 12 (2021): 1607–1617.

2. J. Jumper, R. Evans, A. Pritzel, et al., "Highly Accurate Protein Structure Prediction With AlphaFold," *Nature* 596, no. 7873 (2021): 583–589.

3. R. Evans, M. O'Neill, A. Pritzel, et al., "Protein Complex Prediction With AlphaFold-Multimer," *BioRxiv* (2022).

4. J. Abramson, J. Adler, J. Dunger, et al., "Accurate Structure Prediction of Biomolecular Interactions With AlphaFold 3," *Nature* 630, no. 8016 (2024): 493–500.

5. S. K. Burley and H. M. Berman, "Open-Access Data: A Cornerstone for Artificial Intelligence Approaches to Protein Structure Prediction," *Structure* 29, no. 6 (2021): 515–520.

6. T. Saldano, N. Escobedo, J. Marchetti, et al., "Impact of Protein Conformational Diversity on AlphaFold Predictions," *Bioinformatics* 38, no. 10 (2022): 2742–2748.

7. K. Henzler-Wildman and D. Kern, "Dynamic Personalities of Proteins," *Nature* 450, no. 7172 (2007): 964–972.

8. A. Stein, D. M. Fowler, R. Hartmann-Petersen, and K. Lindorff-Larsen, "Biophysical and Mechanistic Models for Disease-Causing Protein Variants," *Trends in Biochemical Sciences* 44, no. 7 (2019): 575–588.

9. W. Zheng, Q. Wuyun, L. Freddolino, and Y. Zhang, "Integrating Deep Learning, Threading Alignments, and a Multi-MSA Strategy for High-Quality Protein Monomer and Complex Structure Prediction in CASP15," *Proteins: Structure, Function, and Bioinformatics* 91, no. 12 (2023): 1684–1703.

10. W. Zheng, Q. Wuyun, Y. Li, et al., "Deep-Learning-Based Single-Domain and Multidomain Protein Structure Prediction With D-I-TASSER," *Nature Biotechnology* (2025): 1–13.

11. W. Zheng, Q. Wuyun, Y. Li, C. Zhang, L. Freddolino, and Y. Zhang, "Improving Deep Learning Protein Monomer and Complex Structure Prediction Using DeepMSA2 With Huge Metagenomics Data," *Nature Methods* 21, no. 2 (2024): 279–289.

12. H. K. Wayment-Steele, A. Ojoawo, R. Otten, et al., "Predicting Multiple Conformations via Sequence Clustering and AlphaFold2," *Nature* 625, no. 7996 (2024): 832–839.

13. S. Mansoor, M. Baek, H. Park, G. R. Lee, and D. Baker, "Protein Ensemble Generation Through Variational Autoencoder Latent Space Sampling," *Journal of Chemical Theory and Computation* 20, no. 7 (2024): 2689–2695.

14. D. Wu and L. Feng, "Robust Prediction of Multiple Protein Conformations With Entropy Guidance," *BioRxiv* (2025).

15. Y. Zhang and J. Skolnick, "SPICKER: A Clustering Approach to Identify Near-Native Protein Folds," *Journal of Computational Chemistry* 25, no. 6 (2004): 865–871.

16. M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment," *Nature Methods* 9, no. 2 (2011): 173–175.

17. L. S. Johnson, S. R. Eddy, and E. Portugaly, "Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure," *BMC Bioinformatics* 11 (2010): 431.

18. M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Soding, and M. Steinegger, "Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments," *Nucleic Acids Research* 45, no. D1 (2017): D170–D176.

19. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and UniProt C, "UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches," *Bioinformatics* 31, no. 6 (2015): 926–932.

20. M. Steinegger and J. Soding, "Clustering Huge Protein Sequence Sets in Linear Time," *Nature Communications* 9, no. 1 (2018): 2542.

21. L. Richardson, B. Allen, G. Baldi, et al., "MGnify: The Microbiome Sequence Data Analysis Resource in 2023," *Nucleic Acids Research* 51, no. D1 (2023): D753–D759.

22. G. Ahdritz, N. Bouatta, C. Floristean, et al., "OpenFold: Retraining AlphaFold2 Yields New Insights Into Its Learning Mechanisms and Capacity for Generalization," *Nature Methods* 21, no. 8 (2024): 1514–1524.

23. Z. Li, X. Liu, W. Chen, et al., "Uni-Fold: An Open-Source Platform for Developing Protein Folding Models Beyond AlphaFold," *BioRxiv* (2022).

24. M. Mirdita, K. Schutze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "ColabFold: Making Protein Folding Accessible to all," *Nature Methods* 19, no. 6 (2022): 679–682.

25. M. Baek, F. DiMaio, I. Anishchenko, et al., "Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network," *Science* 373, no. 6557 (2021): 871–876.

26. Z. Lin, H. Akin, R. Rao, et al., "Evolutionary-Scale Prediction of Atomic-Level Protein Structure With a Language Model," *Science* 379, no. 6637 (2023): 1123–1130.

27. R. Wu, F. Ding, R. Wang, et al., "High-Resolution de Novo Structure Prediction From Primary Sequence," *BioRxiv* (2022).

28. Y. Li, C. Zhang, D. J. Yu, and Y. Zhang, "Deep Learning Geometrical Potential for High-Accuracy Ab Initio Protein Structure Prediction," *Iscience* 25, no. 6 (2022): 104425.

29. D. Xu, L. Jaroszewski, Z. Li, and A. Godzik, "FFAS-3D: Improving Fold Recognition by Including Optimized Structural Features and Template Re-Ranking," *Bioinformatics* 30, no. 5 (2014): 660–667.

30. N. Ben-Tal, A. Meier, and J. Söding, "Automatic Prediction of Protein 3D Structures by Probabilistic Multi-Template Homology Modeling," *PLoS Computational Biology* 11, no. 10 (2015): e1004343.

31. J. Söding, "Protein Homology Detection by HMM–HMM Comparison," *Bioinformatics* 21, no. 7 (2005): 951–960.

32. T. Lengauer, J. Ma, S. Wang, Z. Wang, and J. Xu, "MRFalign: Protein Homology Detection Through Alignment of Markov Random Fields," *PLoS Computational Biology* 10, no. 3 (2014): e1003500.

33. S. Wu and Y. Zhang, "MUSTER: Improving Protein Sequence Profile–Profile Alignments by Using Multiple Sources of Structure Information," *Proteins: Structure, Function, and Bioinformatics* 72, no. 2 (2008): 547–556.

34. H. Zhou and Y. Zhou, "Fold Recognition by Combining Sequence Profiles Derived From Evolution and From Depth-Dependent Structural Alignment of Fragments," *Proteins: Structure, Function, and Bioinformatics* 58, no. 2 (2004): 321–328.

35. C. M. Deane, W. Zheng, Q. Wuyun, et al., "Detecting Distant-Homology Protein Structures by Aligning Deep Neural-Network Based Contact Maps," *PLoS Computational Biology* 15, no. 10 (2019): e1007411.

36. S. Bhattacharya, R. Roche, B. Moussad, and D. Bhattacharya, "DisCovER: Distance- and Orientation-Based Covariational Threading for Weakly Homologous Proteins," *Proteins: Structure, Function, and Bioinformatics* 90, no. 2 (2021): 579–588.

37. D. W. A. Buchan, D. T. Jones, and A. Valencia, "EigenTHREADER: Analogous Protein Fold Recognition by Efficient Contact Map Threading," *Bioinformatics* 33, no. 17 (2017): 2684–2690.

38. S. Ovchinnikov, H. Park, N. Varghese, et al., "Protein Structure Determination Using Metagenome Sequence Data," *Science* 355, no. 6322 (2017): 294–298.

39. K. Kaminski, J. Ludwiczak, K. Pawlicki, V. Alva, and S. Dunin-Horkawicz, "pLM-BLAST: Distant Homology Detection Based on Direct Comparison of Sequence Representations From Protein Language Models," *Bioinformatics* 39, no. 10 (2023): btad579.

40. W. Liu, Z. Wang, R. You, et al., "PLMSearch: Protein Language Model Powers Accurate and Fast Sequence Search for Remote Homology," *Nature Communications* 15, no. 1 (2024): 2775.

41. L. Pantolini, G. Studer, J. Pereira, J. Durairaj, G. Tauriello, and T. Schwede, "Embedding-Based Alignment: Combining Protein Language Models With Dynamic Programming Alignment to Detect Structural Similarities in the Twilight-Zone," *Bioinformatics* 40, no. 1 (2024): btad786.

42. W. Zheng, Q. Wuyun, X. Zhou, Y. Li, P. L. Freddolino, and Y. Zhang, "LOMETS3: Integrating Deep Learning and Profile Alignment for Advanced Protein Template Recognition and Function Annotation," *Nucleic Acids Research* 50, no. W1 (2022): W454–W464.

43. A. Gustaf, B. Nazim, F. Christina, et al., "OpenFold: Retraining AlphaFold2 Yields New Insights Into Its Learning Mechanisms and Capacity for Generalization," *Nature Methods* 21 (2024): 1514–1524.

44. J. Chen, Z. Hu, S. Sun, et al., "Interpretable RNA Foundation Model From Unannotated Data for Highly Accurate RNA Structure and Function Predictions," *arXiv* (2022).

45. P. Danaee, M. Rouches, M. Wiley, D. Deng, L. Huang, and D. Hendrix, "bpRNA: Large-Scale Automated Annotation and Analysis of RNA Secondary Structure," *Nucleic Acids Research* 46, no. 11 (2018): 5381–5394.

46. C. Zhang, Y. Zhang, and A. M. Pyle, "rMSA: A Sequence Search and Alignment Algorithm to Improve RNA Structure Modeling," *Journal of Molecular Biology* 435, no. 14 (2023): 167904.

47. R. Pearce, G. S. Omenn, and Y. Zhang, "De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials From Deep Learning," *BioRxiv* (2022).

48. C. Zhang, M. Shine, A. M. Pyle, and Y. Zhang, "US-Align: Universal Structure Alignments of Proteins, Nucleic Acids, and Macromolecular Complexes," *Nature Methods* 19, no. 9 (2022): 1109–1115.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Number of models generated per target by the "Zheng" group in CASP16. **Table S2:** The average TM-scores of all 123 groups (36 server groups and 87 human groups) on 19 macromolecular ensemble targets in CASP16, in descending order of performance. EnsembleFold was registered as "Zheng" and "MIEnsembles-Server" (bold font) in CASP16. **Table S3:** The comparisons of "Zheng" and "MIEnsembles-Server" with "AF3-Server" on 19 macromolecular ensemble targets in CASP16. *p* values were calculated between TM-scores by "Zheng"/"MIEnsembles-Server"

and "AF3-Server" using paired one-sided Student's *t*-tests. **Table S4:** Comparison of average TM-scores between the "Zheng" group and all other groups across four types of macromolecular ensemble targets. "Average" TM-scores were calculated by averaging non-zero values per target from all other groups, and then taking the average across targets within each target type. **Table S5:** TM-scores of the first model and the best model for "Zheng" group on 13 CASP16 protein-related ensemble targets. The best model refers to the highest TM-score among the submitted models by each group. **Table S6:** The average TM-scores of all 55 groups (18 server groups and 37 human groups) on 18 hybrid targets in CASP16. TM-scores for conventional hybrid targets were obtained from the CASP16 results webpage (https://predictioncenter.org/casp16/results.cgi?tr_type=hybrid) For M1228v1/v2 and M1239v1/v2, TM-scores were recomputed using the ensemble-specific evaluation protocol. EnsembleFold was registered as "Zheng" and "MIEnsembles-Server" (bold font) in CASP16. **Figure S1:**. Schematic of the updated DeepMSA2 pipeline used in our methods with two key improvements: (i) a larger metagenomic sequence database, and (ii) a multi-domain MSA assembly method that combines domain-level MSAs into a full chain-level MSA. **Figure S2:** Comparison of the running time required to generate a single model for each target using AlphaFold3 and EnsembleFold. For large targets (R1283v3 and R1253v1/v2), smaller stoichiometries were used to enable the sampling of more decoys under limited computational resources. **Figure S3:** The MSA origins of the 3 DMFold clustering models for T1214. **Figure S4:** Number of non-gap residues at each position in the three MSAs for T1214. (A) The stage 1 of dMSA. (B) The stage 2 of qMSA. (C) The stage 3 of qMSA. In all panels, the beta-strand region that undergoes conformational changes is highlighted in red. **Figure S5:** The comparison of TM-scores of decoys generated by the MSA subsampling strategy for T1214 with and without the corresponding ligand. (A) Results across all MSA depth settings (max_extra_msa), with different colors representing distinct subsampling configurations. (B–I) Results for each individual subsampling configuration. In all panels, the top-ranked model is highlighted by a red star. **Figure S6:** The comparison of TM-scores of decoys generated by the MSA masking strategy for T1214 with and without the corresponding ligand. (A) Results across all masked regions, with different colors representing distinct regions. (B–I) Results for each individual region. In all panels, the top-ranked model is highlighted by a red star. **Figure S7:** The impact of available templates on modeling Ensemble targets. RNA monomer targets are excluded because EnsembleFold does not perform template searches for RNA monomer targets. For nucleic acid–containing complexes, only monomeric templates were used in EnsembleFold, and thus no threading template results are available for these targets. **Figure S8:** The average TM-scores of all 55 groups (18 server groups and 37 human groups) on 18 hybrid targets in CASP16. TM-scores for conventional hybrid targets were obtained from the CASP16 results webpage (https://predictioncenter.org/casp16/results.cgi?tr_type=hybrid). For M1228v1/v2 and M1239v1/v2, TM-scores were recomputed using the ensemble-specific evaluation protocol. **Figure S9:** Case study of M1239/T1239. (A) The experimental structures for the two distinct conformations of M1239. (B) The experimental structure for monomer T1239. (C) MolClust result of EnsembleFold decoys for M1239. (D) The best decoy for M1239v2 and corresponding EnsembleFold model. (E) The best decoy for M1239v1 and corresponding EnsembleFold model.