BMC Bioinformatics

RESEARCH Open Access

M-DeepAssembly: enhanced DeepAssembly based on multi-objective multi-domain protein conformation sampling

Xinyue Cui[1†], Yuhao Xia[1†], Minghua Hou[1], Xuanfeng Zhao[1], Suhui Wang[1] and Guijun Zhang[1*]

[†]Xinyue Cui and Yuhao Xia should be regarded as Joint First Authors.

*Correspondence: zgj@zjut.edu.cn

[1] College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

Abstract

Background: Association and cooperation among structural domains play an important role in protein function and drug design. Despite remarkable advancements in highly accurate single-domain protein structure prediction through the collaborative efforts of the community using deep learning, challenges still exist in predicting multi-domain protein structures when the evolutionary signal for a given domain pair is weak or the protein structure is large.

Results: To alleviate the above challenges, we proposed M-DeepAssembly, a protocol based on multi-objective protein conformation sampling algorithm for multi-domain protein structure prediction. Firstly, the inter-domain interactions and full-length sequence distance features are extracted through DeepAssembly and AlphaFold2, respectively. Secondly, subject to these features, we constructed a multi-objective energy model and designed a sampling algorithm for exploring and exploiting conformational space to generate ensembles. Finally, the output protein structure was selected from the ensembles using our in-house developed model quality assessment algorithm. On the test set of 164 multi-domain proteins, the results show that the average TM-score of M-DeepAssembly is 15.4% and 2.0% higher than AlphaFold2 and DeepAssembly, respectively. It is worth noting that there are models with higher accuracy in ensembles, achieving an improvement of 20.3% and 6.4% relative to the two baseline methods, although these models were not selected. Furthermore, when compared to the prediction results of AlphaFold2 for CASP15 multi-domain targets, M-DeepAssembly demonstrates certain performance advantages.

Conclusions: M-DeepAssembly provides a distinctive multi-domain protein assembly algorithm, which can alleviate the current challenges of weak evolutionary signals and large structures to some extent by forming diverse ensembles using multi-objective protein conformation sampling algorithm. The proposed method contributes to exploring the functions of multi-domain proteins, especially providing new insights into targets with multiple conformational states.

Keywords: Protein structure prediction, Multi-domain protein assembly, Multi-objective energy model, Conformation sampling

Introduction

Since the Mulder proposed the concept of proteins in 1838 [1], proteins have remained a central focus of molecular biology research. Proteins play diverse functional roles in various biological processes [2], such as catalysis of biochemical reactions, transporting nutrients, recognizing and transmitting biological signals [3], etc. Traditional experimental methods require significant time and resources, which made predicting protein structures a challenge in biology for decades [4]. Fortunately, with the development of deep learning, protein structure prediction technology has made significant progress in the past decade [5–9], especially with the emergence of a series of excellent methods such as AlphaFold2 [9] and RoseTTAFold [5], which tackled a 50-year challenge in the field of biology. The success of mapping homologous sequences to a single structure (e.g., AlphaFold2 and RoseTTAFold) does not signify the end of structural biology [4]. For instance, as mentioned in the recent 15th critical assessment of structure prediction (CASP15) review article, there are still challenges in multi-domain proteins when the domain pairing information obtained is weak or the protein structure is large [10].

Structural domains are states situated between the secondary and tertiary structures of a protein [11], representing independent units in evolution, structure, and function [12, 13]. In nature, multi-domain proteins often result from the repetitions and combinations of single domains [2]. Interactions among these structural domains directly impact a protein's conformation and activity, thus influencing physiological processes in organisms. Currently, only one-third of multi-domain proteins have been resolved in the Protein Data Bank (PDB) [14]. Compared to single-domain proteins, multi-domain proteins exhibit higher degrees of freedom in the regions connecting structural domains and in interaction zones [15]. In addition, the lack of effective methods for assembling structural domains and linkers optimization strategy predicts that multi-domain protein structures are more urgent and crucial.

Due to the high computational demands for full-chain modeling of multi-domain proteins, a "divide and conquer" strategy can effectively decompose and address this complex problem [12]. The method involves splitting the full-length sequence based on domain boundaries, generating individual models for each structure, and ultimately assembling each structural domain into a complete full-chain model [16]. For instance, Rosetta [17] employs a strategy that keeps the skeleton of each domain fixed and uses a sampling strategy for the full range of rotamers. AIDA [18] realizes a de novo method guided by ab initio folding potentials, but methods based on de novo- or ab initio-based calculations will lead to the final structure being largely randomly oriented [19]. The representative methods based on structural domain rigid-body docking include DEMO [20] and SADA [19], both of which mainly rely on structure templates. With the advancement of deep learning and the improvement of single-domain prediction, full-chain protein modeling based on deep learning has also become one of the mainstream methods. The E2EDA [12] leverages deep learning to achieve end-to-end multi-domain protein structure prediction, transforming the predicted inter-domain rigid motions into direct assembly of full-chain models. However, E2EDA tends to result in atomic conflict. Meanwhile, our recently developed DeepAssembly [21] assembles multi-domain proteins by predicting inter-domain interactions using a convolutional neural network. DeepAssembly mainly focuses on the inter-domain interactions, whereas AlphaFold2

mainly focuses on the full-chain and ignores inter-domain interactions to some extent. Thereby, better results can be obtained for multi-domain proteins by combining the advantages of both methods.

Here, we propose a multi-domain protein assembly method, M-DeepAssembly, which is based on a multi-objective protein conformation sampling algorithm. This method obtains inter-domain interactions using DeepAssembly [21] while extracting distance features of the full-length sequence using AlphaFold2 [9]. Subject to the above features, population-based dihedral angle optimization is performed under the guidance of a multi-objective optimization algorithm to output diverse ensembles [22–24]. Ultimately, the ensembles are screened using our model quality assessment algorithm to output the final structure. M-DeepAssembly was tested on 164 multi-domain proteins and 13 CASP15 multi-domain protein targets, consistently outperforming AlphaFold2. These results suggest that our algorithm can overcome the limitations of DeepAssembly to some extent by exploring large-scale conformation sampling.

Methods

Overview

The pipeline of M-DeepAssembly is illustrated in Fig. 1. The full-length protein sequence is segmented into multiple single-domain sequences using our previously proposed method, the sequence-based domain parser DomBpred [25]. Firstly, the Multiple Sequence Alignment (MSA) feature, inter-domain feature, and template feature are fed into the interaction prediction network to acquire inter-domain interactions (For more details, see DeepAssembly [21]). Simultaneously, the single-domain structures and the full-length sequence distance features were obtained using AlphaFold2 [9]. Based on the single-domain structures, the dihedral angles of the linkers are randomly initialized to generate an initial conformational population. The ensembles are obtained through a

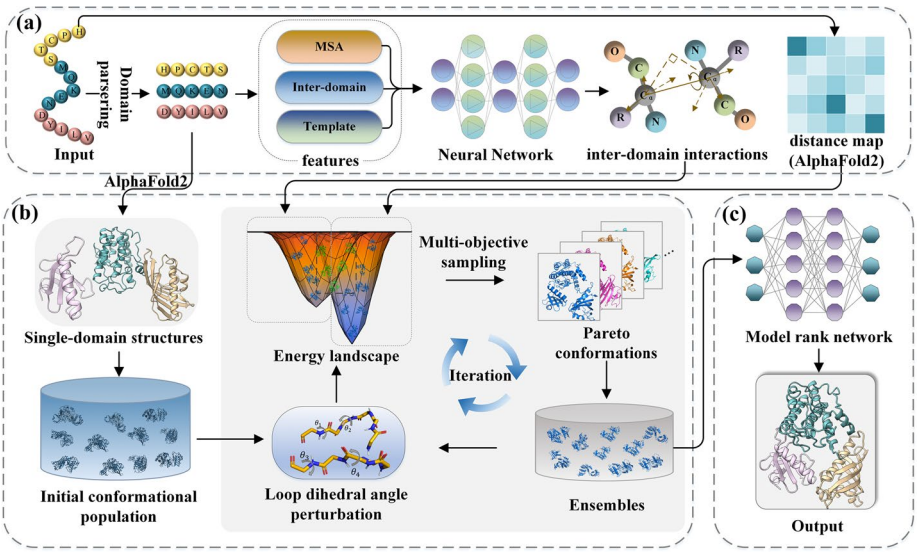

Fig. 1 Pipeline of the M-DeepAssembly. **a** Protein sequence domain parsing and features extraction; **b** Multi-objective protein conformation sampling; **c** The part of MQA, which is used to rank and select output structure from the ensembles

multi-objective conformational sampling. Finally, the model quality assessment (MQA) algorithm is used to select the top-ranking models from given ensembles. The flowchart and specific details of M-DeepAssembly are shown in Additional file 1: Fig. S1.

Material preparation

Based on the input sequence, the MSA feature is obtained by searching Uniclust30 [26] and BFD [27] databases through HHblits [28], the template feature is searched PDB database using HHsearch [28], and the inter-domain feature is extracted through our previously proposed DomBpred [25]. All these features are fed into our DeepAssembly network to predict inter-domain interactions, combined with full-length sequence distance features obtained from AlphaFold2 to construct the multi-objective energy model.

Multi-objective energy model

Generally, it is not possible to find a single solution to meet all tasks because different tasks might conflict with each other [29]. Instead, the multi-objective optimization strategy is able to divide the multi-problem into multiple sub-problems. By solving the sub-problems in parallel, a set of well-distributed pareto optimal solutions is obtained. In our multi-objective conformation sampling algorithm, based on the obtained inter-domain interactions and the full-length sequence distance, two energy functions are constructed to guide the sampling process.

The energy function of the inter-domain interactions for the target conformation x is described as follows:

$$f_{\text{inter}}(x) = \sum_{i=1}^{l}\sum_{j=1}^{l}\left| d_{i,j} - d^{*}_{i,j} \right|, \tag{1}$$

where l is the number of residues in the linkers; and $d_{i,j}$ represent the predicted C_α distance between the i-th and j-th residues, and $d^{*}_{i,j}$ represents C_α distance between the i-th and j-th residues of the target conformation.

In fact, capturing the full-length protein structural interactions can compensate to some extent for the lack of information on inter-domain interactions. To address this, we designed an energy function to measure protein full-length sequence interactions. Referring to our previous work [30], the energy function is designed as follows:

$$f_{\text{full}}(x) = \sum_{w=1}^{L}\sum_{q=1}^{L}\frac{d_{w,q} - d^{*}_{w,q}}{d'}, \tag{2}$$

$$d' = \log\left(|w-q| + \varepsilon\right),\ w \neq q, \tag{3}$$

where L represents the length of protein sequence; $d_{w,q}$, $d^{*}_{w,q}$ represent the predicted C_α distance between the w-th and q-th residues and the C_α distance of the target conformation, respectively; ε is a very small number to avoid d' being zero.

Based on Eqs. (1) and (2), we constructed a multi-objective energy model $f(x)$, which is described as follows:

$$min_x f(x) = [f_1(x), f_2(x), \ldots, f_k(x)], \tag{4}$$

$$s.t.\ x \in \Omega. \tag{5}$$

where $f_i(x), i \in \{1, 2, \ldots, k\}$ is the i-th energy function; k represents the total number of energy functions, x is the decision vector in the decision space Ω, in this study, it can be regarded as a conformation in the ensembles. The algorithm optimizes all energy functions simultaneously by exploiting the shared information among them. For the two conformations x^a and x^b in the ensembles, for all energy functions $f_i(x)$, if $f_j(x^a) \leq f_j(x^b)$, $\forall j \in \{1, 2, \ldots, m\}$ and $f_q(x^a) < f_q(x^b), \exists q \in \{1, 2, \ldots, m\}$[31], then x^a is said to dominate $x^b (x^a \prec x^b)$. If there does not exist a decision vector dominates x^* in the whole decision space, then x^* is called a pareto optimal solution or non-dominated solution (see Fig. 2).

Initialization

The single-domain protein structures predicted by AlphaFold2 [9] are connected from N-terminal to C-terminal in sequence order, and the dihedral angles of the linkers are randomly perturbed to generate N full-length conformations that form the initial population. The random initial dihedral angle vector $\Theta_i^{\text{initial}}$ of the i-th conformation is set according to the following equation:

$$\Theta_i^{\text{initial}} = (2\boldsymbol{X} - \boldsymbol{I}) d_{\max}, \quad i \in \{1, 2, \ldots, N\}, \tag{6}$$

where $\boldsymbol{X}$ is an M-dimensional vector composed of random numbers between 0 and 1, $\boldsymbol{I}$ is an M -dimensional vector of ones, and $d_{\max}$ represents the maximum perturbation value for initial angles. The above M represents the dihedral angle dimension of the target conformation, i.e., the dimension of the decision optimization variable.

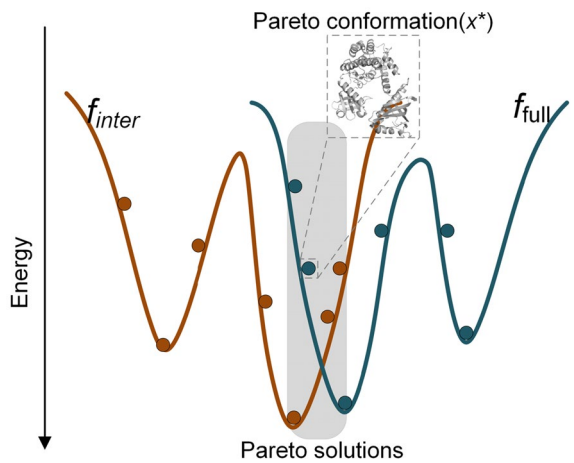

Fig. 2 Energy landscape for multi-objective sampling, the protein conformation corresponding to the lower energy (pareto solutions) is obtained under the guidance of two energies

Crossover and mutation operators

To update the population, we interact the linker fragments between the conformations in the population to construct new conformations to help jump out of the local energy basin. Therefore, it is necessary to design crossover and mutation operators for new conformation generation, which can expand the conformational space by perturbing the dihedral angles of the multi-domain protein linkers. For each individual $\Theta_i = \{\theta_{i,j}\}_{i=1,\ldots,N,j=1,\ldots,M}$, it is a vector composed of M dihedral angles of the i-th conformation, we use the following equation for mutation and crossover.

$$\theta_{i,j}^{\text{trial}} = \theta_{i,j}^{\text{target}} + F\left(\theta_{f_1,j} - \theta_{f_2,j}\right),\ \ i,f_1,f_2 \in \{1,\ 2,\ldots,N\},\ i \neq f_1 \neq f_2, j \in \{1,\ 2,\ldots,M\} \tag{7}$$

$$\theta_{i,j} = \begin{cases} \theta_{i,j}^{\text{trail}}, & if\ rand(0,1) \leq p_c \\ \theta_{i,j}^{\text{target}}, & otherwise \end{cases},\ \ i \in \{1,\ 2,\ldots,N\}, j \in \{1,\ 2,\ldots,M\} \tag{8}$$

where $\theta_{i,j}^{\text{target}}$ represents the j-th dihedral angle of linkers in the i-th conformation, $\theta_{f_1,j}$ and $\theta_{f_2,j}$ are the dihedral angles corresponding to two conformations randomly selected from the ensembles generated after multi-objective sampling, $\theta_{i,j}^{\text{trail}}$ is the trial dihedral angle corresponding to conformation after mutation, F is a scaling factor; rand(0, 1) is a random number between 0 and 1, p_c represents the crossover factor, when rand(0, 1) is less than p_c, the mutation dihedral angle $\theta_{i,j}^{\text{trail}}$ is selected as the final dihedral angle ($\theta_{i,j}$), otherwise, the $\theta_{i,j}^{\text{target}}$ is selected.

Population update

The conformations after crossover and mutation are used to generate non-dominated solutions through multi-objective conformation sampling. For conformation x^*, which is called a non-dominated solution if there does not exist a conformation x in decision space that satisfies $x \prec x^*$. In addition, for the dominated x' conformation, we receive with probability p_a. All obtained x^* and x' are used to update the population and form ensembles. The final ensembles are output when the maximum number of iterations $G_{\max}$ is reached (see Additional file 1: Fig. S1 and Table S1).

Model selection

Finally, we use an in-house developed model quality assessment [32–35] algorithm (DeepUMQA-rank) to select the top-ranking (top1) model in the final ensembles generated by M-DeepAssembly. This algorithm employs a neural network based on an axial attention mechanism, which integrates knowledge derived from protein sequence features and protein family structural features to assess the quality of the protein structures. The algorithm will be available in a preprint soon.

Results

Dataset

For a fair comparison, we used the same test set as in DeepAssembly, which consists of 164 TM-score < 0.8 multi-domain proteins from the *H. sapiens* proteome in the AlphaFold database [36]. Among these, there are 104 proteins with 2 domains, 30 proteins with 3 domains, 30 proteins with more than 3 domains, and a maximum of 7 domains.

In addition, we collected 13 multi-domain protein targets of CASP15 as a test set for objective comparison with AlphaFold2 [9] and AlphaFold3 [37].

Results of the benchmark set

To test the performance of M-DeepAssembly, we compared it with DeepAssembly and AlphaFold2 on 164 non-redundant multi-domain proteins where single-domain structures are predicted by AlphaFold2 [9]. The results are summarized in Fig. 3a. Here, we use TM-score (template modeling score) [38] and RMSD (root mean square deviation) as the evaluation metrics to measure the topological similarity of protein structures. The results show that more than 60% of the proteins obtained better performance than AlphaFold2 and AlphaFold3 [37]. Specifically, the average TM-score of the models generated by M-DeepAssembly (0.704) is 15.4% higher and the RMSD (10.419 Å) is 23.3% lower than the models generated by AlphaFold2 (TM-score: 0.610, RMSD: 13.590 Å). Similarly, compared to models generated by AlphaFold3 (TM-score: 0.645, RMSD: 13.444 Å), M-DeepAssembly achieved a 9.15% higher TM-score and a 22.5% lower RMSD. Interestingly, there are better models in our ensembles, although they have not been selected by DeepUMQA-rank. The average TM-score of the best model is 13.8% and 6.4% (Table 1) higher than AlphaFold3 and DeepAssembly, respectively (see Additional file 1: Table S2 and Table S3), which suggests that MQA is crucial for model selection and is an important research field in the future. Overall, we think that the reason for the improvement of M-DeepAssembly lies in the multi-objective conformation sampling strategy. Moreover, with the breakthroughs in single-domain structure prediction,

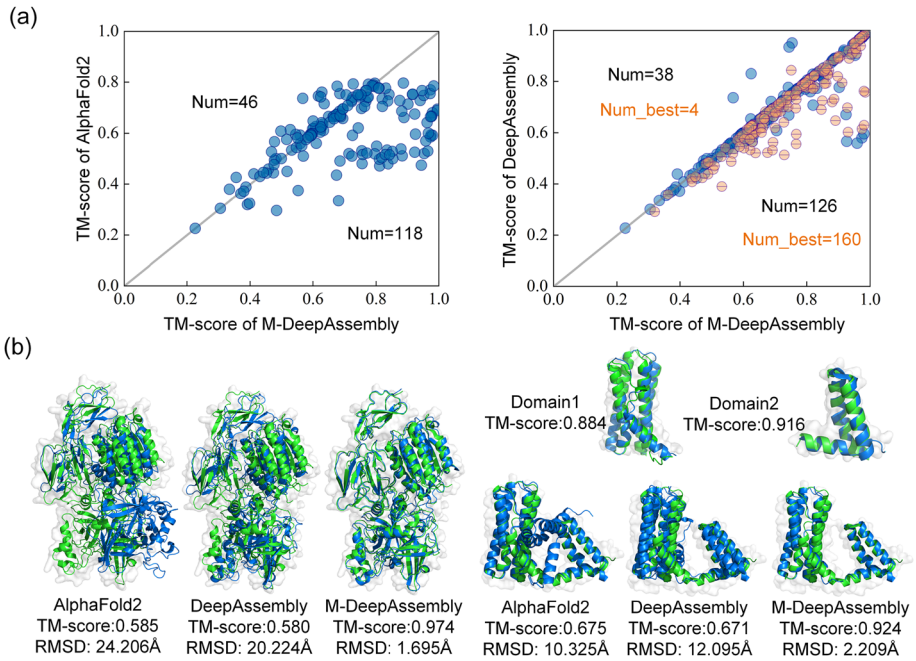

Fig. 3 **a** Performance comparison of M-DeepAssembly and AlphaFold2, DeepAssembly on the test set, Num represents the number of proteins on both sides of the split line, and Num_best is the number of results of the best model comparison in the test set; **b** M-DeepAssembly compares the prediction performance on the two cases with AlphaFold2 and DeepAssembly, respectively

Table 1 Comparison of M-DeepAssembly with other methods on 164 multi-domain proteins.

Method	TM-score		RMSD (Å)	
	top1	best	top1	best
AlphaFold2	0.610	–	13.590 Å	–
AlphaFold3	0.645	–	13.444 Å	–
DeepAssembly	0.690	–	11.019 Å	–
M-DeepAssembly	0.704	0.734	10.419 Å	8.774 Å

top1 in the table represents the top-ranking models, and best represents the best model in the ensembles. The results of AlphaFold3 were predicted using the official server.

Table 2 Comparison of M-DeepAssembly with other methods on 13 multi-domain proteins in CASP15.

Method	TM-score		RMSD (Å)	
	top1	best	top1	best
AlphaFold2	0.567	–	20.578 Å	–
AlphaFold3	0.544	–	25.439 Å	–
M-DeepAssembly†	0.573	0.591	16.795 Å	13.904 Å
M-DeepAssembly‡	0.741	0.762	8.336 Å	7.414 Å

The results of AlphaFold2 come from the official evaluation CASP15 website. M-DeepAssembly† and M-DeepAssembly‡ represent the assembly using AlphaFold2 and experimental single-domain structures, respectively. The results of AlphaFold3 were predicted using the official server.

M-DeepAssembly is expected to accelerate the solution of multi-domain protein structures. All experiments were performed on a machine equipped with an NVIDIA TITAN RTX GPU, a 16-core CPU, and 32 GB of RAM.

In particular, it is worth noting that M-DeepAssembly provides a lightweight method for predicting longer sequence proteins through a "divide-and-conquer" strategy. For example, the crystal structure chain *f* of complement C3b (Fig. 3b, left side) is composed of 5 domains (PDB: 2xwbf [39]). AlphaFold2 and DeepAssembly only obtain the TM-score accuracies of 0.585 and 0.580 on this multi-domain target, whereas the TM-score for each single-domain structure predicted by AlphaFold2 are 0.973, 0.974, 0.975, 0.925 and 0.977, respectively. Our M-DeepAssembly achieves high-precision (0.974) assembly by combining the advantages of the above two baseline methods. The right side of Fig. 3b is a multi-domain structure composed of all alpha helices (PDB: 3ajmA), which protein with important roles in cell signaling and blood vessel development. The TM-score of the two single-domain structures predicted by AlphaFold2 are 0.884 and 0.916, respectively, while the full-chain prediction accuracy is only 0.675. However, the model predicted by M-DeepAssembly achieved a TM-score of 0.924. These results further demonstrate that higher prediction accuracy can be achieved through multi-objective conformation sampling that integrates inter-domain interactions and full-length sequence distance features.

Results of CASP15 multi-domain targets

For objectively evaluate the performance of M-DeepAssembly, we collected a total of 13 multi-domain targets based on the "domain definition" from the CASP15 official website. Results for AlphaFold2 and our M-DeepAssembly are shown in Table 2, where

the results of AlphaFold2 are from the official evaluation. The average TM-scores of the models generated by AlphaFold2 and M-DeepAssembly are 0.567, 0.573, respectively (see Additional file 1: Table S4 and Table S5 for details). The average TM-score of the best model in the ensembles is 0.591, which is 4.2%, 8.6% higher than AlphaFold2 and AlphaFold3, respectively (see Additional file 1: Table S6 for details). Especially on target T1121, M-DeepAssembly showed a 24.5% increase in TM-score (0.925) compared to AlphaFold2 (0.743). Additionally, we performed assembly using experimental single-domain structures, and the average TM-score generated by M-DeepAssembly was increased to 0.741. The results suggest that deviations in the prediction accuracy of single-domain structures directly lead to poor prediction results of full-length multi-domain proteins, which also indicates that single-domain proteins are still not completely solved.

Ablation study

To evaluate the performance impact of M-DeepAssembly on each component, we designed two ablation experiments on 164 benchmark datasets to analyze the impact of full-length sequence distance features and inter-domain interactions on model performance. The first experiment (M-DeepAssembly-w/o-inter) only considers full-length sequence distance features, while the second experiment (DeepAssembly-w/o-full) only uses inter-domain interactions. The test results of different versions of M-DeepAssembly on the benchmark test set are shown in Fig. 4 (see Additional file 1: Table S7 for detailed results).

The average TM-score and RMSD of the best models in the ensembles (M-DeepAssembly best) generated by M-DeepAssembly are 0.734 and 8.774 Å, respectively. While the top-ranking (top1) model (M-DeepAssembly) exhibits an average TM-score of 0.704 and RMSD of 10.419 Å, the average TM-score of M-DeepAssembly (0.704) is 2.0% and 37.8% higher than M-DeepAssembly-w/o-full (0.690) and M-DeepAssembly-w/o-inter (0.511), respectively. The comparative results of the ablation experiments highlight the excellent performance of the inter-domain interactions provided by DeepAssembly.

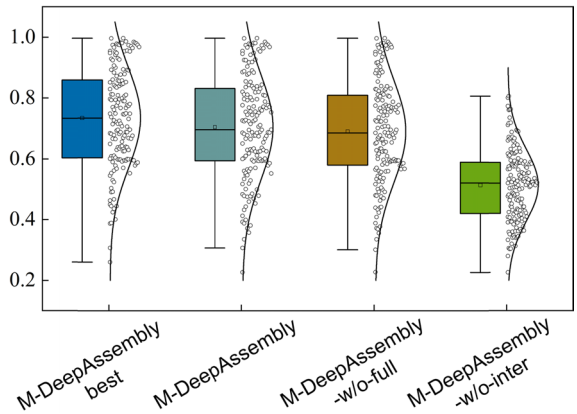

Fig. 4 M-DeepAssembly ablation experimental results, where M-DeepAssembly best represents the best model in the obtained ensembles. M-DeepAssembly-w/o-full represents a version that dose not include the full-length sequence distance features predicted by AlphaFold2, and M-DeepAssembly-w/o-inter represents a version that does not contain the inter-domain interactions

Simultaneously, they provide further evidence that the complementation of full-length sequence distance features and inter-domain interactions can improve the accuracy of the prediction model.

Case study: insights from multiple conformations in ensembles

Interestingly, our testing revealed that our method has the potential to detect dynamic changes in protein conformation. For example, two conformations (Fig. 5(a)) comprising the Streptococcus pneumoniae response regulator spr1814 [40, 41] were present simultaneously in our obtained ensembles. The spr1814 plays an important role in facilitating DNA binding and transcriptional activation [42, 43], which consists of two conformations, each conformation consisting of an N-terminal receiver domain (residues 1–119) and the C-terminal effector domain (residues 141–199). The inter-domain interactions of spr1814 are predominantly mediated by salt bridges present in two different residues. In conformation A, the salt bridge is formed between Glu188 and Lys84, Arg91 and Glu195, with interactions between the receiver structural domain and effector structural domain. However, when the conformation shifts to conformation B (Fig. 5b), the salt bridges disappear, resulting in a 74° rotational change (Fig. 5a) in the effector domain of conformation B relative to A [44]. This rotation leads to loosening of the inter-domain interactions in conformation B, subsequently causing phosphorylation to mediate dimerization of the receiver domain. M-DeepAssembly successfully captured two distinct conformational states and achieved TM-scores of 0.985 and 0.861. (Fig. 5b).

Discussion

The stability of multi-domain proteins usually depends on interactions with other protein cofactors, and the high flexibility of the hinge region between domains makes the prediction accuracy of multi-domain proteins much lower than that of single-domain structures. Therefore, how to make full use of the breakthroughs in single-domain structure prediction techniques to accurately reveal and predict the relative orientation relationships among the structural domains of proteins has become the key to research.

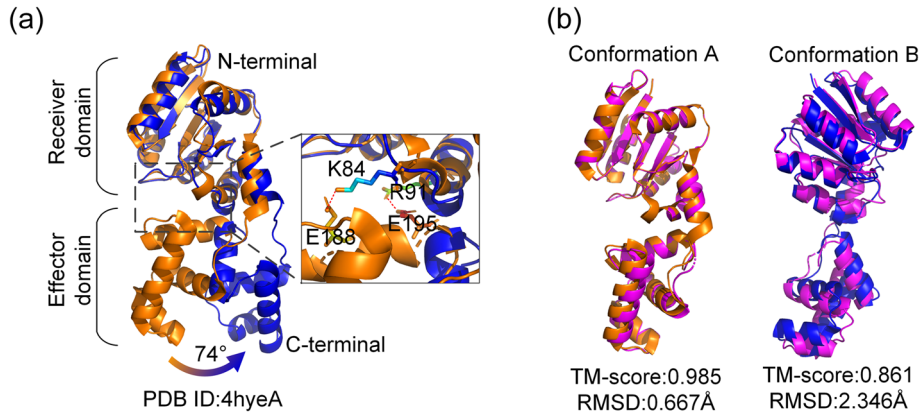

Fig. 5 **a** Two conformations of 4hyeA, which are components of spr1814. **b** Comparison of the two structures predicted by M-DeepAssembly (magenta color) with native conformation A (orange color) and native conformation B (blue color), respectively

For the multi-domain protein structure prediction problem, we developed M-DeepAssembly based on existing structure predictors. In M-DeepAssembly, we focus more on methodological innovation and practical application. The method extracts the inter-domain and intra-domain distance information predicted by DeepAssembly [21] and AlphaFold2 [9], and combines them with a multi-objective optimization algorithm to explore conformational space to generate diverse conformations, thereby improving the structure prediction accuracy of multi-domain proteins to some extent. Our method was compared to AlphaFold2, AlphaFold3, and DeepAssembly on 164 multi-domain proteins from the *H. sapiens* protein and 13 multi-domain targets from the 15th critical assessment of structure prediction (CASP15). With single-domain data predicted by AlphaFold2, M-DeepAssembly successfully improves the prediction accuracy of more than 60% of the proteins in 164 test sets compared to all other methods. Although our method performs well on the aforementioned test sets, M-DeepAssembly currently struggles with prediction tasks for proteins longer than 1500 amino acids. Further works are needed for the problem of structure prediction of proteins in multi-domain complexes.

Conclusion

Proteins should not be viewed as just a single static structure, but rather as a conformational ensemble with multiple accessible states [45]. The flexible regions of proteins are often the key to protein function, and the study of the motion of multi-domain protein hinge regions and the multiple conformational states they induce is important for revealing biological processes and their regulatory mechanisms.

Our future research will focus on exploring the multiple conformations and conformational ensembles of proteins to further reveal the dynamic properties of proteins and their roles in biology.

Abbreviations

CASP	Critical assessment of techniques for protein structure prediction
PDB	Protein data bank
MSA	Multiple sequence alignment
MQA	Model quality assessment

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06131-2.

Additional file 1.

Additional file 2.

Acknowledgements

We would like to thank current and previous members of the IOBIO-Lab for helpful discussions and suggestions. We thank P.Z. for helping to develop the server and organize the experimental data.

Author contributions

G.Z. conceptualized and supervised this study and wrote the paper. X.C. performed the test, analyzed the data, and wrote the paper. Y.X. developed the DeepAssembly, analyzed data, and wrote the paper. M.H. designed the multi-objective algorithm part and wrote the paper. X.Z. developed the DeepUMQA-rank and wrote the paper. S.W. analyzed data and wrote the paper. All authors have read and approved the final manuscript.

Funding

This work is supported by the National Key R&D Program of China (2022ZD0115103), the National Nature Science Foundation of China (62173304), Zhejiang Provincial Special Support Program for High-Level Talents (2023R5248).

Availability of data and materials
All the data used in the methods of this paper are listed in Additional file 2 and are available at https://www.rcsb.org/. The M-DeepAssembly online server is freely accessible at http://zhanglab-bioinf.com/M-DeepAssembly/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 22 January 2024 Accepted: 3 April 2025
Published online: 05 May 2025

References

1. Mulder GJ. Sur la composition de quelques substances animales. Bull Sci Phys Nat Neerl. 1838;1838(104):9.
2. Apic G, Huber W, Teichmann SA. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. J Struct Funct Genomics. 2003;4:67–78.
3. Morris R, Black KA, Stollar EJ. Uncovering protein function: from classification to complexes. Essays Biochem. 2022;66(3):255–85.
4. Zhou Y, Litfin T, Zhan J. 3= 1+ 2: How the divide conquered de novo protein structure prediction and what is next? Natl Sci Rev. 2023;10(12):nwad259.
5. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373(6557):871–6.
6. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577(7792):706–10.
7. Xu J, Mcpartlon M, Li J. Improved protein structure prediction by deep learning irrespective of co-evolution information. Nat Mach Intell. 2021;3(7):601–9.
8. Yang J, Anishchenko I, Park H, et al. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci. 2020;117(3):1496–503.
9. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
10. Ozden B, Kryshtafovych A, Karaca E. The impact of AI-based modeling on the accuracy of protein assembly prediction: insights from CASP15. Proteins. 2023;91(12):1636–57.
11. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci. 2003;100(21):12105–10.
12. Zhu H, Xia Y, Zhang G. E2EDA: Protein domain assembly based on end-to-end deep learning. J Chem Inf Model. 2023;63(20):6451–61.
13. Kinch LN, Grishin NV. Evolution of protein structures and functions. Curr Opin Struct Biol. 2002;12(3):400–8.
14. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. Nucleic Acids Res. 2000;28(1):235–42.
15. Zhou X, Zheng W, Li Y, et al. I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. Nat Protoc. 2022;17(10):2326–53.
16. Zheng W, Li Y, Zhang C, et al. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins. 2019;87(12):1149–64.
17. Wollacott AM, Zanghellini A, Murphy P, et al. Prediction of structures of multidomain proteins from structures of the individual domains. Protein Sci. 2007;16(2):165–75.
18. Xu D, Jaroszewski L, Li Z, et al. AIDA: ab initio domain assembly for automated multi-domain protein structure prediction and domain–domain interaction prediction. Bioinformatics. 2015;31(13):2098–105.
19. Peng C, Zhou X, Xia Y, et al. Structural analogue-based protein structure domain assembly assisted by deep learning. Bioinformatics. 2022;38(19):4513–21.
20. Zhou X, Hu J, Zhang C, et al. Assembling multidomain protein structures through analogous global structural alignments. Proc Natl Acad Sci. 2019;116(32):15930–8.
21. Xia Y, Zhao K, Liu D, et al. Multi-domain and complex protein structure prediction using inter-domain interactions from deep learning. Commun Biol. 2023;6(1):1221.
22. Gunantara N. A review of multi-objective optimization: methods and its applications. Cogent Eng. 2018;5(1):1502242.
23. Zhao K, Liu J, Zhou X, et al. MMpred: a distance-assisted multimodal conformation sampling for de novo protein structure prediction. Bioinformatics. 2021;37(23):4350–6.
24. Xia Y, Peng C, Zhou X, et al. A sequential niche multimodal conformational sampling algorithm for protein structure prediction. Bioinformatics. 2021;37(23):4357–65.
25. Yu Z, Peng C, Liu J, et al. DomBpred: protein domain boundary prediction based on domain-residue clustering using inter-residue distance. IEEE/ACM Trans Comput Biol Bioinform. 2022;20(2):912–22.
26. Mirdita M, Von Den Driesch L, Galiez C, et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017;45(D1):D170–6.

27. BFD. https://bfd.mmseqs.com/.
28. Steinegger M, Meier M, Mirdita M, et al. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinform. 2019;20:1–15.
29. Lin X, Zhen H, Li Z, et al. Pareto multi-task learning. Adv Neural Inf Process Syst. 2019; 32.
30. Hou M, Jin S, Cui X, et al. Protein multiple conformation prediction using multi-objective evolution algorithm. Interdiscip Sci. 2024;16(3):519–31.
31. Zitzler E, Thiele L. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. IEEE Trans Evol Comput. 1999;3(4):257–71.
32. Liu D, Zhang B, Liu J, et al. Assessing protein model quality based on deep graph coupled networks using protein language model. Brief Bioinform. 2024;25(1):bbad420.
33. Liu J, Liu D, Zhang G. DeepUMQA3: a web server for accurate assessment of interface residue accuracy in protein complexes. Bioinformatics. 2023;39(10):btad591.
34. He G, Liu J, Liu D, et al. GraphGPSM: a global scoring model for protein structure using graph neural networks. Brief Bioinform. 2023;24(4):219.
35. Hiranuma N, Park H, Baek M, et al. Improved protein structure refinement guided by deep learning based accuracy estimation. Nat Commun. 2021;12(1):1340.
36. Varadi M, Anyango S, Deshpande M, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50(D1):D439–44.
37. Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493–500.
38. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004;57(4):702–10.
39. Forneris F, Ricklin D, Wu J, et al. Structures of C3b in complex with factors B and D give insight into complement convertase formation. Science. 2010;330(6012):1816–20.
40. Park AK, Moon JH, Oh JS, et al. Crystal structure of the response regulator spr1814 from Streptococcus pneumoniae reveals unique interdomain contacts among NarL family proteins. Biochem Biophys Res Commun. 2013;434(1):65–9.
41. Park AK, Bong SM, Moon JH, Chi YM. Crystallization and preliminary X-ray crystallographic studies of DesR, a thermosensing response regulator in a two-component signalling system from Streptococcus pneumoniae. Struct Biol Cryst Commun. 2009;65(7):727–9.
42. Da Re S, Schumacher J, Rousseau P, et al. Phosphorylation-induced dimerization of the FixJ receiver domain. Mol Microbiol. 1999;34(3):504–11.
43. Lewis RJ, Scott DJ, Brannigan JA, et al. Dimer formation and transcription activation in the sporulation response regulator Spo0A. J Mol Biol. 2002;316(2):235–45.
44. Poornam GP, Matsumoto A, Ishida H, et al. A method for the analysis of domain movements in large biomolecular complexes. Proteins. 2009;76(1):201–12.
45. Degiacomi MT. Coupling molecular dynamics and deep learning to mine protein conformational space. Structure. 2019;27(6):1034–40.

Publisher's Note