Fundamental Research 6 (2026) 77–87

Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Article

Multiple conformational states assembly of multidomain proteins using evolutionary algorithm based on structural analogues and sequential homologues

Chunxiang Peng [a], Xiaogen Zhou [a], Jun Liu [a], Minghua Hou [a], Stan Z. Li [b], Guijun Zhang [a,*]

[a] College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China
[b] AI Lab, School of Engineering, Westlake University, Hangzhou 310024, China

ARTICLE INFO

Article history:
Received 7 August 2023
Received in revised form 18 April 2024
Accepted 7 May 2024
Available online 16 May 2024

Keywords:
Domain assembly
Multiple conformational states
Evolutionary algorithm
Deep learning
Protein structure prediction

ABSTRACT

With the breakthrough of AlphaFold2, nearly all single-domain protein structures can be built at experimental resolution. However, accurate modelling of full-chain structures of multidomain proteins, particularly all relevant conformations for those with multiple states remain challenging. In this study, we develop a multidomain protein assembly method, M-SADA, for assembling multiple conformational states. In M-SADA, a multiple population-based evolutionary algorithm is proposed to sample multiple conformational states under the guidance of multiple energy functions constructed by combining homologous and analogous templates with inter-domain distances predicted by deep learning. On a developed benchmark dataset containing 72 multidomain proteins with multiple conformational states, the performance of M-SADA is significantly better than that of AlphaFold2 on multiple conformational states modelling, where 29/72 (40.3%) of proteins can be assembled with a TM-score > 0.90 for two highly distinct conformational states with M-SADA. Furthermore, M-SADA is tested on a developed benchmark dataset containing 296 multidomain proteins with single conformational state, and results show that the average TM-score of M-SADA on the best models is 0.913, which is 5.2% higher than that of AlphaFold2 models (0.868). Results on CASP15 multidomain targets also show that M-SADA can predict new domain arrangements when individual domain structures are correct.

1. Introduction

AlphaFold2 [1], an end-to-end protein structure prediction approach based on attention and equivariant transformer networks, has achieved unprecedented prediction accuracy, as demonstrated in the CASP14 experiment [2,3], representing a major advance in the field as a consequence of long-term efforts [4]. Many protein structures with high resolution can be modelled by AlphaFold2 [1,5]. Accordingly, the DeepMind and the European Molecular Biology Laboratory (EMBL)-European Bioinformatics Institute collaborated to create a new data resource, AlphaFold Protein Structure Database (AlphaFold DB) [6,7], dramatically expanding the structural coverage of the known protein-sequence space using AlphaFold2. However, AlphaFold2 is hardly optimized specifically for multidomain proteins, and AlphaFold2 model's confidence score correlates strongly with the presence of homologues in the Protein Data Bank (PDB) [8,9]. Thus, the structure prediction of multidomain proteins remains relatively unreliable when the number of homologous templates or sequences of related proteins is insufficient for inferencing relative domain orientations [10]. More important, predicted structural models do not capture conformational dynamics [11]. The AlphaFold2 models are usually close, whereas studying the mechanism of action requires knowledge of the broader conformational landscape [11]. This also applies to multidomain proteins, for which function is closely related to changes in tertiary and quaternary structures across multiple conformational states. The development of computational methods to address the gap between single and multiple domains, and that between single and multiple conformational states can yield great dividends [11].

Domain assembly methods can be generally classified into two major categories: *ab initio* methods and template-based approaches. *Ab initio* methods can be adopted to predict multidomain protein full-chain structures when reliable structures are available for individual domains, such as Rosetta [12], AIDA [13] and GalaxyDomDock [10]. However, the domain structures may remain largely randomly orientated by *ab initio* methods [14]. Template-based methods use template information to infer domain orientations, making the orientation between domains more reliable to a certain extent. Representative template-based domain assembly methods include DEMO [9,14] and SADA [15]. In DEMO, docking-based domain assembly simulations are performed to generate full-chain models of multidomain proteins. Further, DEMO-EM is proposed, a domain enhanced modelling method that uses cryo-EM data to generate accurate full-chain structural models for multidomain proteins [16,17]. However, DEMO is heavily dependant on the quality of

* Corresponding author.
E-mail address: zgj@zjut.edu.cn (G. Zhang).

https://doi.org/10.1016/j.fmre.2024.05.003
2667-3258/© 2024 The Authors. Publishing Services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

templates, especially when there are a large number of domains, the assembly accuracy noticeably decreases. In SADA [15], analogous full-chain templates are identified through domain-level structure alignments, and domain orientations are simultaneously searched for generating the full-chain model through a two-stage differential evolution algorithm guided by the energy function, with an inter-residue distance potential predicted by deep learning. However, SADA only used analogous templates to infer inter-domain orientations but homologous templates are also important. Especially with the advent of AlphaFold DB, some high-quality models can be an effective complement to homologous templates. Additionally, deep learning-based distance prediction models employed by these template-based domain assembly methods usually regard multidomain proteins as single domain proteins, which overlook the contribution to inter-domain distance prediction.

Although homologous templates that can be used to model the full-chain structures of multidomain proteins are frequently unavailable due to the limited structures in PDB [13,14], the high-quality models in AlphaFold DB may be an effective complement to PDB [7]. Therefore, the adequate use of sequential homologous and structurally analogous templates may be an important approach for modelling multidomain proteins. In fact, many multidomain proteins are expected to undergo conformational changes involving domain orientations to form different conformational states [10]. For example, the balance of a kinase's active and inactive state must be regulated precisely in a cell. In an oversimplified picture, only these two highly distinct conformations exist for a kinase. While the vast majority of ATP competitive inhibitors bind to the active conformation of kinases, a few small molecules (e.g. the anticancer drug imatinib) bind selectively to the inactive form of ABL1 [18]. For advanced protein structure prediction methods, such as AlphaFold2, although they can accurately predict backbone and side chain conformations, they are for a particular conformational state [18].

In this work, a method called M-SADA is proposed to model multiple conformational states of multidomain proteins, where potential domain orientations are explored through a multiple population-based evolutionary algorithm on the basis of sequential homologues, structural analogues and deep learning predicted inter-domain distance for multidomain proteins. Compared with AlphaFold2 and state of the art domain assembly methods, the proposed method is able to predict more better the full-chain structure and inter-domain orientation and has the ability to explore multiple conformational states of multidomain proteins. On a development multidomain protein dataset with multiple conformational states, the average TM-score of M-SADA is 0.87 for all conformational states, while that of AlphaFold2 is 0.75. Meanwhile, the results on large-scale multidomain protein datasets with single conformational state proteins indicate that M-SADA outperforms advanced domain assembly methods, such as SADA, and full-chain modelling methods, such as AlphaFold2.

2. Materials and methods

2.1. Datasets

To examine the ability of M-SADA to model multidomain proteins with multiple conformational states, its effectiveness to model multidomain proteins with single conformational state, and its domain assembly performance, we tested and discussed M-SADA on three different test sets.

In order to study the ability of M-SADA on modelling multidomain proteins with multiple conformational states, M-SADA is tested on 72 multidomain proteins with multiple conformational states. The benchmark dataset is constructed in accordance with the following criteria: (1) all multidomain proteins in PDB are clustered at 100% sequence identity cut-off, and the TM-scores between different conformations in each cluster (i.e. each multidomain protein) are calculated, (2) the multidomain proteins with at least one pair of conformations with TM-score ≤ 0.75 are selected, (3) the two structures with the largest structural difference in the multidomain proteins are selected as the representative conformational states of the multidomain protein, (4) these multidomain proteins are removed if residues are missing at the domain boundary for representative conformations and (5) the final benchmark set with multiple conformational states (containing 72 multidomain proteins) is generated with a 40% sequence identity cut-off. To the best of our knowledge, this multidomain protein dataset is the first to be constructed systematically to investigate the ability to model multidomain protein structures with multiple conformational states.

In order to test the effectiveness of M-SADA, we reassemble 296 single conformational state multidomain proteins randomly selected from AlphaFold DB. The selection criteria for the 296 proteins are as follows: (1) the missing residues in the crystal structure are less than 10%, (2) the sequence identity with the DeepIDDP training set is < 40% and (3) the sequence identity between each other is < 40%. Text S1 details the selection principles for these thresholds, which are used to select test proteins.

In order to fairly compare the domain assembly performance of M-SADA with other domain assembly methods at the same level, 302 multidomain proteins used in the SADA benchmark dataset with sequence identity to the training set of DeepIDDP < 40% are selected to test the performance of M-SADA [15]. The sequence identity amongst all the 302 proteins is less than 30%. The dataset contains 137 proteins with 2 domains, 61 proteins with 3 domains, 36 proteins with more than 4 domains and 68 proteins with discontinuous domain.

2.2. Methods

The pipeline of M-SADA is shown in Fig. 1. Starting from the input full-chain sequence and computationally predicted (or experimentally solved) domain structures, the full-chain structural analogues are detected from our developed multidomain protein structure database (MPDB) [15,19]. Then, the homologous templates are successively searched from MPDB, PDB, AlphaFold DB90 and AlphaFold DB according to the input full-chain sequence. Subsequently, the analogous and homologous templates are used to design multiple energy functions, which include different template restraints, physical constraints and residue distance information predicted by the in-house inter-domain distance prediction method DeepIDDP [20]. Based on multiple energy functions, domain orientations are optimized by a designed multiple population-based evolutionary algorithm. Finally, full-chain models from different populations are selected and ranked using our developed model quality assessment method DeepUMQA2 [21].

2.2.1. Templates for domain assembly

Templates are important for modelling protein structures. Structural information from templates is considerably more reliable than that from other sources, particularly when the target protein and the template are highly homologous. However, homologous templates that can be used to model the full-chain structure of multidomain proteins are often unavailable [14,20], because only partial structures (such as single-domain structures) have been experimentally solved for most multidomain proteins due to technical difficulties [14]. In practice, the structural space of protein-protein interfaces is small (< 1,000 structurally distinct interfaces) [5]. Similarly, the structural space of domain-domain interfaces is also limited. In some cases, structural analogous templates can be used to model multidomain protein structure, because domains often interact in a similar way in the quaternary orientations if the domains have similar tertiary structures [15]. The comprehensive use of inter-domain interaction information provided by sequential homologues and structural analogues is necessary for domain assembly. Here, homologous templates are detected according to the input sequence, and analogous templates of the full-chain are identified according to the structural similarity between domain models and the proteins in MPDB. Different templates are used to construct multiple different energy functions for

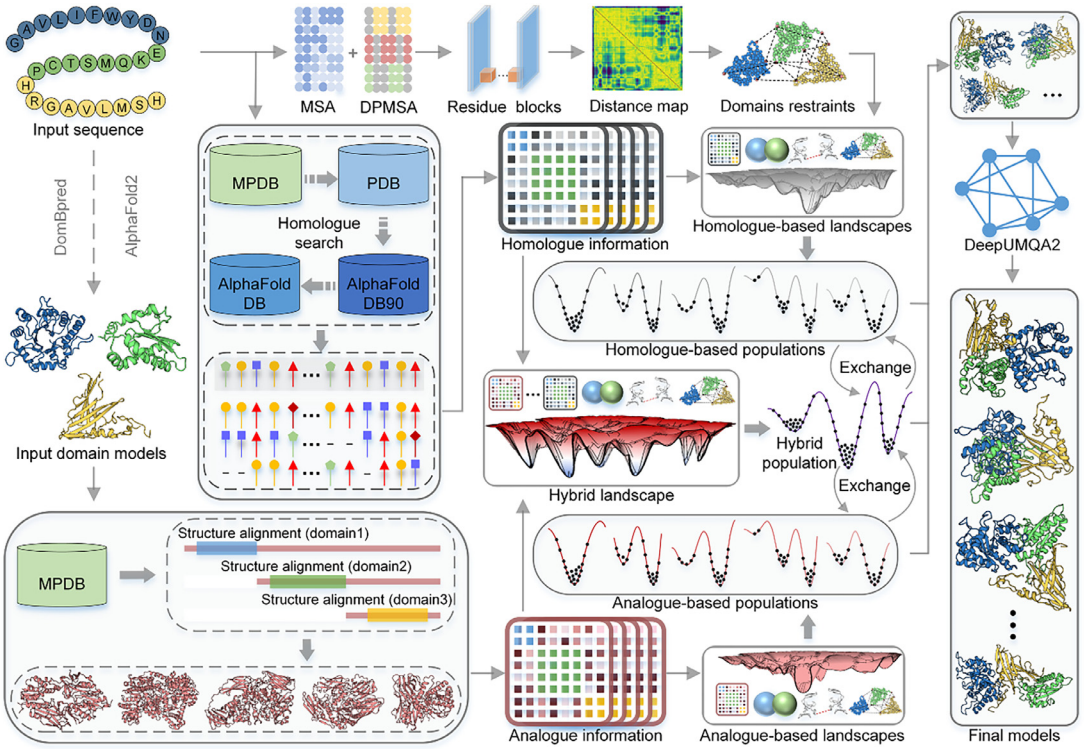

Fig. 1. Pipeline of M-SADA. Starting from the input full-chain sequence and individual domain structures, the structural templates are first identified from MPDB by structural alignment, the homologous templates are successively searched from MPDB, PDB, AlphaFold DB90 and AlphaFold DB. Meanwhile, the residue distances between different domains are predicted by an inter-domain distance predictor DeepIDDP. Subsequently, the analogous templates, homologous templates and predicted inter-domain distances are used to design multiple energy functions, which include different template restraints, physical constraints and spatial restraints predicted by deep learning. Based on multiple energy functions, domain orientations are optimized by a multiple population-based evolutionary algorithm. Finally, the lowest energy full-chain models from different populations are selected and ranked using model quality assessment method DeepUMQA2.

domain assembly, because they are complementary for modelling multidomain proteins to a certain extent [15].

2.2.2. Multidomain protein structure database update

Domains often interact with each other to perform biological functions, and thus, some multidomain proteins in the PDB are identical in sequence but different in structure. These diverse structures with the same sequence may also be important. Thus, we add multidomain protein structures with identical sequences but structural differences to MPDB as follows: (i) clustering all proteins in PDB with 100% sequence identity and then calculating the TM-scores between the proteins in each cluster, (ii) selecting the clusters with at least one pair of proteins with TM-score $\leq$ 0.85 and (iii) selecting the two proteins with the largest structural differences in each selected cluster to MPDB.

2.2.3. Analogous template identification

Structurally analogous proteins comprise an important class of templates for building the full-chain structure of multidomain proteins. Here, analogous templates for full-chain structures are identified from MPDB through a two-stage procedure based on TM-align [22], as shown in Fig. 2.

In the first stage, each domain is aligned to a template by TM-align, irrespective of the overlap. The quality of the template is evaluated by a local similarity LS_{score}, and 200 templates with the highest LS_{score}s are used in the next stage, where LS_{score} is similarly defined as that in our previously developed SADA [15].

In the second stage, the domain models are structurally aligned on each template selected in the first stage, where the overlap is not allowed in the alignments of different domains. The top five analogous templates with the highest LS_{score}s in the second stage are selected for domain assembly. For each one of the top five analogous templates, if there is a domain that cannot be correctly aligned to the analogous template (TM-score of the domain < 0.5), then M-SADA splits input domain structures into two parts and identifies analogous templates separately. The templates that can cover domains1–(n − 1) and the templates that can cover domains(n − 1)–n are independently detected from the previously selected templates. The template with the highest LS_{score} for domains1–(n–1) and the template with the highest LS_{score} for domains(n–1)–n are selected for domain assembly. Details of analogous template identification can be found in Text S2.

2.2.4. Homologous template recognition

Homologous proteins share a common ancestor and usually have similar structures in homologous sequence regions, making them pivotal for predicting protein structures lacking experimental data [23]. However, available protein structures deposited in the PDB are limited (about 200,000 publicly available structures as of January 2023) [24], accounting for less than 0.1% of the protein sequence database UniProtKB/TrEMBL [25], resulting in some homologous proteins not being available for structural modelling. However, AlphaFold DB has the potential to be a useful complement to PDB [7], especially for modelling the full-chain structure of multidomain proteins. Therefore, we introduce AlphaFold DB to compensate for the lack of homologous structures available in PDB. The process of detecting the homologous proteins of the target sequence can be divided into four stages that correspond to the searching for four structure databases (MPDB, PDB, AlphaFold DB90 and AlphaFold DB) through HMMER program [26,27], where AlphaFold DB90 represents the models with an average (per-residue local distance difference test) pLDDT $\geq$ 90 in AlphaFold DB. The details can be found in Text S3.

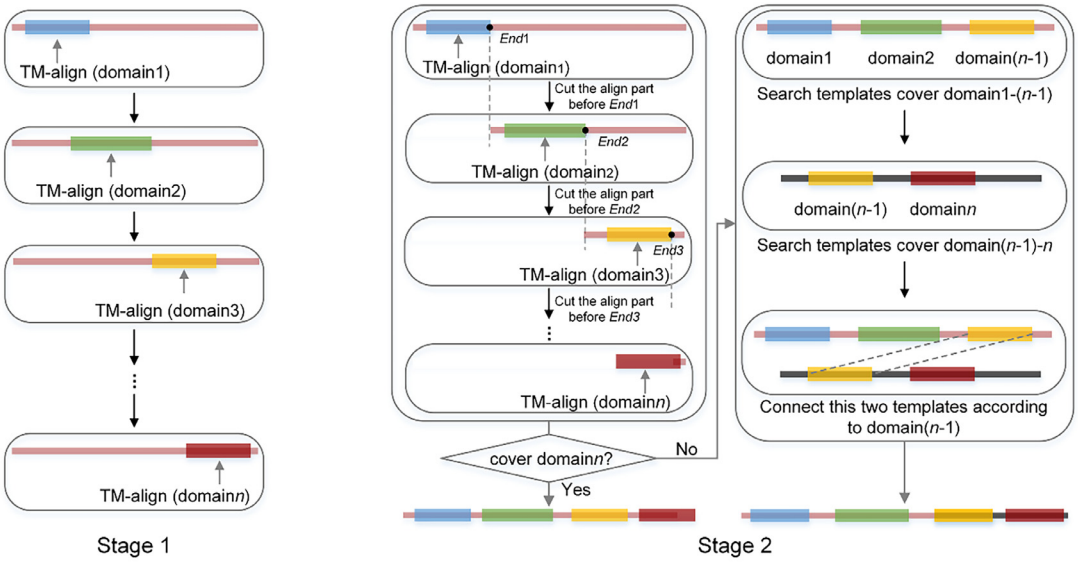

Fig. 2. Process of analogous template identification.

2.2.5. Inter-domain distance prediction neural network

To our knowledge, few distance prediction methods are optimized specifically for the distance of the inter-domain residue pairs of multidomain proteins. Currently, most multidomain protein structure assembly methods use inter-domain distances extracted directly from the distances predicted by methods developed for common distance prediction, possibly reducing ability to capture orientations between domains.

In this work, our recently developed inter-domain distance prediction network (DeepIDDP) is introduced into M-SADA [20]. In DeepIDDP, a data enhancement strategy, called domain-pair multiple sequence alignment, is firstly proposed to enhance the inter-domain coevolutionary information of multidomain proteins. Then, two new features, namely, the inter-domain residue coupling score and the inter-domain average contact potential, are designed to focus on inter-domain residue pairs. Finally, a deep network that combines the attention mechanism and a deep residual block is constructed to predict the inter-domain residue distance distribution of multidomain proteins. DeepIDDP was trained on 9,248 multidomain proteins (8,820 for training and 464 for validation) selected from MPDB. The details of parameters and configurations for DeepIDDP can be referenced in [20].

2.2.6. Energy functions for domain assembly

For multidomain proteins, capturing the interactions between domains is difficult for a single template in some cases because domains often interact with each other to perform more complex functions, resulting in a diversity of orientations between domains. Here, we design multiple energy functions based on different template information to guide domain assembly.

For homologue- or analogue-based populations, we design different energy functions (E^{p}_{total}) for each population based on different templates to guide domain assembly, where p represents the energy function used in the p-th population. E^{p}_{total} is defined as follows:

$$E^{p}_{\text{total}} = w_1 E_{\text{clash}} + w_2 E_{\text{bd}} + w_3 E_{\text{dist}} + w_4 E^{p}_{\text{tpl}} \quad (1)$$

where E_{clash} is to prevent clash between atoms, E_{bd} is to prevent neighbouring domains from being too far apart during the assembly simulation and thus not satisfy the nature of multidomain proteins, E_{dist} is the energy of the inter-domain distance derived from DeepIDDP, which provides the inter-domain information predicted by the deep learning, and $E^{p}_{\text{tpl}}$is the template-based energy function which provides the inter-domain information from templates. The details of each energy term are described in Text S4.

For the hybrid population, we use all the detected template information to design a hybrid energy function ($E^{\text{hybrid}}_{\text{total}}$) to guide domain assembly. $E^{\text{hybrid}}_{\text{total}}$ is defined as follows:

$$E^{\text{hybrid}}_{\text{total}} = w_5 E_{\text{clash}} + w_6 E_{\text{bd}} + w_7 E^{\text{hybrid}}_{\text{dist}} + w_8 E^{\text{hybrid}}_{\text{tpl}} \quad (2)$$

where each energy term is described in Text S4.

2.2.7. Multiple population-based evolutionary algorithm for domain assembly

Differential evolution, introduced by Storn and Price [28], stands out as a highly effective and widely recognized evolutionary algorithm (EA) dedicated to solving global optimization [29]. In the past years, many variants with automated tuning or ensemble of mutation strategies and parameters have been proposed to enhance the search capability of algorithm [29,30]. Amongst them, multiple population based evolutionary algorithms has been considered amongst the most efficient for global optimization [30]. Moreover, multiple population based evolutionary algorithm is suitable for M-SADA, which has several different energy functions and the energy functions are related to each other. For different energy functions, a multiple population-based evolutionary algorithm is proposed to explore optimal solutions, in which population diversity can be maintained because different populations can be located in different search spaces. Solutions amongst different populations can be exchanged rather than independently optimized, enabling the finding of more potential solutions efficiently. The flowchart of the multiple population-based evolutionary algorithm is shown in Fig. S1 and described in Text S5. The proposed algorithm can further improve domain assembly accuracy while ensuring the diversity of domain assembly results.

The same number of populations are set based on the number of energy functions. In each population, a two-stage differential evolution algorithm is proposed to explore and exploit the optimal solution under the guidance of the corresponding energy function. During the simulation, the solutions amongst different populations are exchanged for each learning period generation, where the best k solutions from each population are used to replace the worst $P \times k$ solutions in the hybrid population. Here, P is the number of populations in addition to the hybrid population. Then, the worst k solutions of each population are replaced with the best k solutions of the hybrid population. When the populations converge, the solution with the lowest E^{p}_{total} in each population is selected and ranked by DeepUMQA2 for output. The details of parameters and configurations for DeepUMQA2 can be referenced in [21].

For the hybrid population, inspired by ultrafast shape recognition [31,32], three different solutions are selected based on the cosine similarity of the solutions because multiple template information is utilized in the hybrid population and different stable solutions may be able to increase the diversity of solutions. The rules for the selection of solutions in the hybrid population are as follows: (1) the solution with the lowest $E_{\text{total}}^{\text{hybrid}}$ is selected, named $S_{\text{hybrid, best}}$, (2) the solution with the minimum cosine similarity to $S_{\text{hybrid, best}}$ is selected, name $S_{\text{hybrid, Mbest}}$ and (3) the solution with the minimum cosine similarity to $S_{\text{hybrid, Mbest}}$ is selected, named $S_{\text{hybrid, MMbest}}$. When the cosine similarity amongst the solutions is smaller, the structural difference amongst the models generated by these solutions is greater. Refer to Text S5 for algorithm description and parameter setting of the multiple population-based evolutionary algorithm.

3. Results and discussion

3.1. Modelling multidomain proteins with multiple conformational states

Domains between multidomain proteins frequently interact to perform higher-level functions. Therefore, the conformational states of some multidomain proteins are often not unique. These different conformational states are critical for further studies on protein function and drug design. However, currently available protein structure modelling methods do not seem to specifically address this issue. In this section, we investigate whether M-SADA exhibits the ability to model the full-chain structures of 72 multidomain proteins with different conformational states, and then compare it with state-of-the-art method AlphaFold2 and the latest multiple conformational states folding method MultiSFold [33]. MultiSFold is a recently developed approach for exploring the structures of proteins across multiple conformational states. It's noteworthy, however, that MultiSFold was not explicitly designed for multidomain proteins. As there are few effective methods for modelling the structures of multidomain protein with multiple conformational states, we choose AlphaFold2 predicted structure, as a reference to further analyse the performance of M-SADA. Here, the structure models of AlphaFold2 in each multidomain protein are predicted using its standalone package. In M-SADA, the domain structures for assembly come from AlphaFold2 predicted top 1 model, and the templates are removed using 100% sequence identity cut-off.

To investigate whether the output model contains multiple conformational states, for each multidomain protein, the TM-scores between each native conformational state structure and all assembled (or predicted) models are calculated, and the maximum TM-score is taken. The TM-score is utilized to measure the similarity between two protein structures, with its values ranging from (0,1] [34]. A higher TM-score indicates greater similarity between the two structures. Here, a higher TM-score indicates that the accuracy of the predicted (or assembled) model is superior, bringing it closer to the native structure, which in turn reflects better performance of the algorithm. The results are summarized in Table S1 and the details for each multidomain protein cluster are shown in Table S2. For all conformations of the 72 multidomain proteins, M-SADA achieves an average TM-score of 0.87 compared to 0.77 for AlphaFold2 and 0.78 for MultiSFold. However, some multidomain proteins may contain similar conformational states, which affects the analysis of the multiple conformational states modelling ability to a certain extent. Therefore, we focus on the two conformational states with the maximum structural difference for each multidomain protein cluster, and the two conformational states are named state1 and state2, respectively. On the 72 multidomain proteins, the average TM-score of M-SADA for conformational state1 achieves 0.88, 17.3% higher than that of AlphaFold2 (0.75) for conformational state1 and 14.3% higher than the average TM score of MultiSFold (0.77) for conformational state 1. For conformational state2, M-SADA obtains an average TM-score of 0.86 while AlphaFold2 is 0.76 and MultiSFold is 0.77. For the 72 multidomain proteins containing two conformational states with the maximum structural difference, M-SADA obtains a higher average TM-score (0.87), and the comparison between M-SADA and AlphaFold2 for each protein is shown in Fig. 3a. For the 65 out of 72 multidomain proteins, the average TM-score of M-SADA is better than that of AlphaFold2. Table 1 summarizes the number of multidomain proteins for which the M-SADA, AlphaFold2 and MultiSFold models satisfy different TM-score cut-offs on two different conformational states. On 29 proteins, the TM-scores of M-SADA models exceed 0.90 in both highly distinct conformational states. However, the number of multidomain proteins with TM-scores above 0.90 in both conformational states for the AlphaFold2 models is 2 and for the MultiSFold is 3.

For a more intuitive analysis of the modelling performance on two representative conformational states, the TM-score boxplot and distribution for the two different states of each protein are shown in Fig. 3b. From this figure, we can draw the conclusion that M-SADA can accurately model two different conformational states on more proteins than state-of-the-art modelling method AlphaFold2 and the latest multiple conformational states folding method MultiSFold. The result also validates that the multiple population-based evolutionary algorithm proposed in M-SADA is able to generate models with diversity, which enables M-SADA to explore multiple conformational states of multidomain proteins.

A representative example is shown in Fig. 3c, which is the MspJI restriction endonuclease in complex and belongs to a family of restriction enzymes that cleave DNA containing 5-methylcytosine (5mC) or 5-hydroxymethylcytosine (5hmC) [35]. In subunits A and B of the tetramer, the binding and cleavage domains are close together ('closed' conformation, protein 4r28B), and in subunits C and D they are farther apart with respect to each other ('open' conformation, protein 4r28C) [35]. For AlphaFold2, the 'closed' conformation is accurately modelled, whilst the 'open' conformation is not modelled correctly. By contrast, M-SADA successfully modelled the 'closed' and 'open' conformations after the individual domain models were reassembled by M-SADA, achieving TM-scores of 0.91 and 0.95, respectively. These results indicate M-SADA can probably be used to model different conformational states of multidomain proteins.

3.2. Reassembly of multidomain protein from AlphaFold DB

Although AlphaFold2 has achieved remarkable success in modelling protein structures, the multidomain protein full-chain modelling accuracy appears to be worse on average than that for the constituent domains [4]. This shows the necessity of further effort on inter-domain orientation modelling for protein structure prediction. To test the effectiveness of M-SADA in modelling multidomain protein full-chain structures, M-SADA is used to reassemble 296 multidomain proteins with single conformational state, where the structures of individual domain models are from the AlphaFold2 full-chain model and 100% sequence identity cut-off is used to exclude templates for M-SADA. The accuracy of the reassembled models is compared with the accuracy of the models in AlphaFold DB by taking the TM-score of the M-SADA structures and the crystal structures and comparing it to the TM-score of the AlphaFold2 structure and the crystal structures. Here, we use the AlphaFold2 model to refer to the model deposited in AlphaFold DB. For the AlphaFold2 models at different pLDDT cut-offs, Table 2 presents the average accuracy of the AlphFold2 models and that of models reassembled by M-SADA and SADA. The median TM-scores of the full-chain models generated by different methods are summarized in Table S3. M-SADA-top1 represents the first model ranked by our developed DeepUMQA2 [21], which is currently one of state-of-the-art methods for protein monomer and complex model quality estimation. On the three-month CAMEO dataset (March 11 to June 04 2022,), the Pearson correlation coefficient of DeepUMQA2 is 0.899 at the global evaluation level. The coefficient shows the correlation between the predicted and real global scores of all protein models,

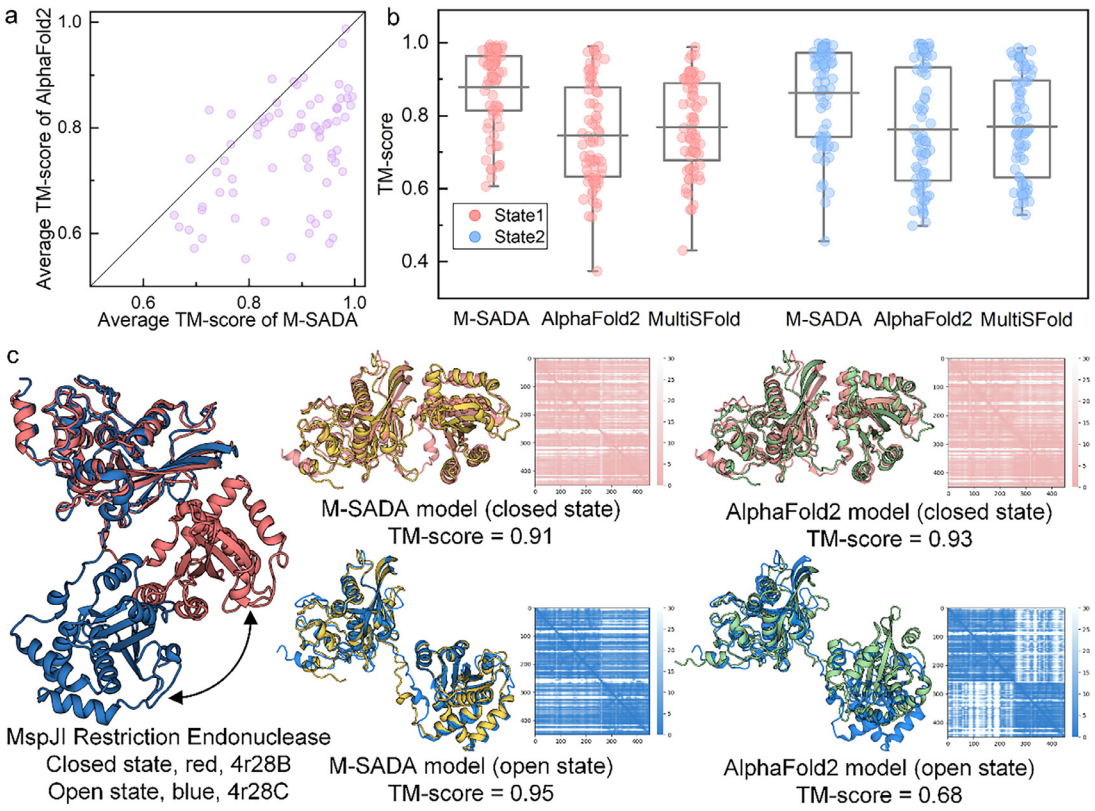

Fig. 3. Comparison of M-SADA with AlphaFold2 and MultiSFold on 72 multidomain proteins containing two conformational states with the maximum structural difference. (a) Head-to-head comparison between the average TM-scores of each multidomain protein generated by AlphaFold2 and M-SADA. (b) TM-score boxplot and distribution for different conformational states on M-SADA, AlphaFold2 and MultiSFold. The grey horizontal line in the box represents the average TM-scores. (c) A representative example is showing M-SADA performance in modelling different conformational states. The red and blue cartoons are native structures of different conformational states. The yellow and green cartoons represent the M-SADA model and AlphaFold2 model, respectively.

Table 1
Number of multidomain proteins for which M-SADA, AlphaFold2 and MultiSFold models satisfy the different TM-score cut-offs on two different conformational states.

Method	The number of multidomain proteins			
	$TM\text{-score}_{s1,\,s2} > 0.75$	$TM\text{-score}_{s1,\,s2} > 0.80$	$TM\text{-score}_{s1,\,s2} > 0.85$	$TM\text{-score}_{s1,\,s2} > 0.90$
M-SADA	46	44	38	29
AlphaFold2	14	7	4	2
MultiSFold	21	12	7	3

Note: $TM\text{-score}_{s1,\,s2}$ > cut-off represents the TM-score of the best model on state1 and the TM-score of the best model on state2 are both greater than cut-off.

Table 2
Summary of modelling results for M-SADA, SADA and AlphaFold2 on 296 multidomain proteins.

Method	Average TM-score			
	pLDDT < 80	80 ≤ pLDDT < 90	90 ≤ pLDDT	All
M-SADA-best	0.674	0.863	0.964	0.913
M-SADA-top1	0.628	0.850	0.956	0.901
SADA	0.564	0.815	0.945	0.879
AlphaFold2	0.524	0.801	0.940	0.868

Note: pLDDT < 80 represents the AlphaFold2 models with an average pLDDT < 80.

and reliability (performance) increases with correlation coefficient [21]. M-SADA-best represents the best model amongst the output models.

Overall, the best models and top1 models of M-SADA obtain a higher TM-score than the models of AlphaFold2 and SADA. On average, the average TM-score for the M-SADA top1 models is 0.901, higher than AlphaFold2 models (0.868, 3.8%) and SADA models (0.879, 2.5%). The average TM-score for the best models of M-SADA is 5.1% higher than that of the AlphaFold2 models. For the AlphaFold2 models with an average pLDDT ≥ 90, the average TM-score of the full-chain models assembled by M-SADA is slightly higher than that of AlphaFold2 models. However, for the AlphaFold2 models with an average pLDDT

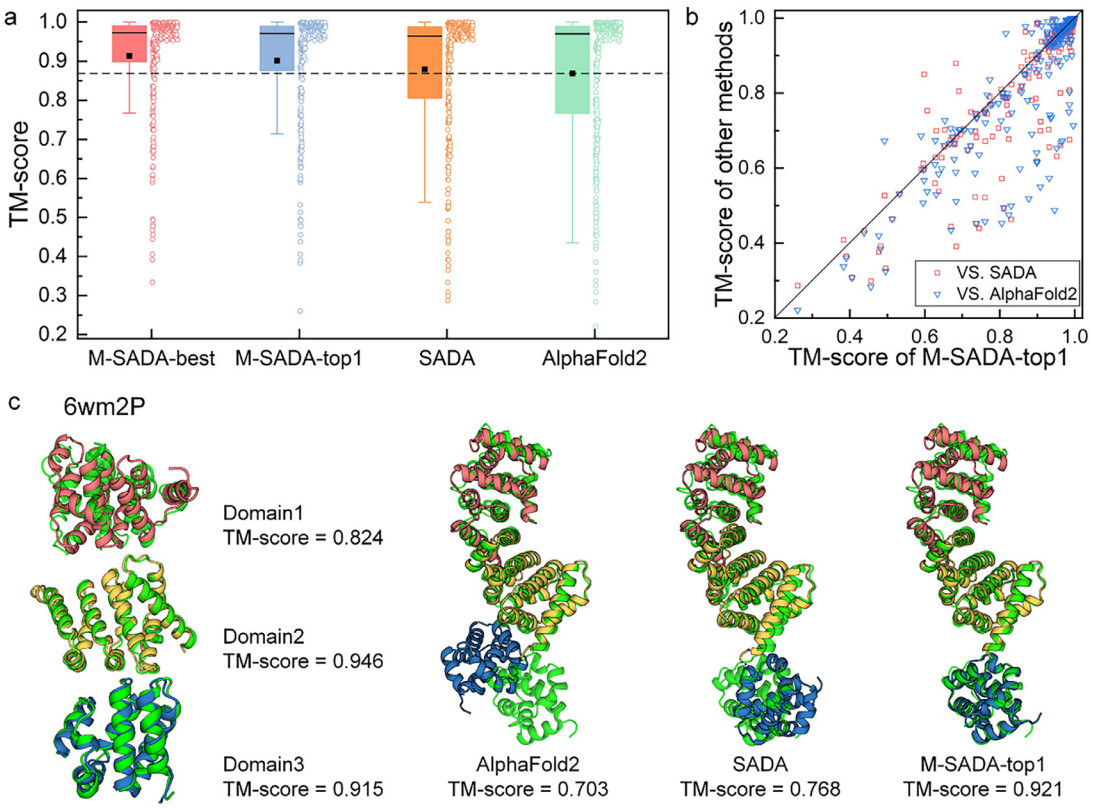

Fig. 4. Comparison between M-SADA with SADA and AlphaFold2. (a) TM-score boxplot and distribution for different methods. The black horizontal line and black square are median and average TM-scores, respectively. (b) TM-score comparison between M-SADA models with AlphaFold2 and SADA models. (c) A representative example shows M-SADA can build better full-chain models, where the models of AlphaFold2, SADA and M-SADA are structurally aligned on native structure. The green cartoon represents the native structure, and different domains are marked with different colours.

≤ 80, the average TM-score of the top1 models of M-SADA achieves 0.628, which is 19.8% higher than that of AlphaFold2 (0.524) and 11.3% higher than that of SADA (0.564). The TM-score boxplot and distribution are illustrated in Fig. 4a. For the TM-score of each protein, a comparison of M-SADA with AlphaFold2 and SADA is shown in Fig. 4b, which indicates the quality of the AlphaFold2 models can be improved by M-SADA. In particular, for the 84 proteins cannot be accurately predicted by AlphaFold2 (TM-score < 0.8), the top1 full-chain models of M-SADA achieve an average TM-score of 0.729 while the average TM-score of AlphaFold2 is 0.618. For the best M-SADA models, they achieve a higher TM-score.

A representative example is shown in Fig. 4c. For 6wm2P with 3 domains, although AlphaFold2 accurately predicts all the domain models (TM-score = 0.824, 0.946 and 0.915), the TM-score of the full-chain model for AlphaFold2 is lower than the average TM-score of the constituent domains because the orientation of the third domain is not modelled correctly. In SADA, the orientation of the third domain is also not constructed correctly, resulting in a TM-score of 0.768 for the full-chain model. However, all domain orientations are modelled correctly by M-SADA, resulting in a TM-score of 0.921 for the top1 full-chain model.

These results also indicate that the accuracy of full-chain modelling is lower than the average accuracy of the constituent domains for some multidomain proteins, and M-SADA can probably improve the quality of AlphaFold2 models to a certain extent. Compared with SADA, M-SADA uses different types of templates and a more advanced inter-domain distance prediction network to generate multiple energy functions. Then, a multiple population-based evolutionary algorithm is used to generate more potential solutions. Therefore, M-SADA can generate more accurate models. The association between the AlphaFold2 model quality and the percentage of the cases improved after M-SADA reassembly is displayed in Fig. S2 to assist in deciding in which cases M-SADA can be utilized to enhance the quality of AlphaFold2 models. Amongst the 296 multidomain proteins, for the AlphaFold2 models with an average pLDDT ≤ 90, more than 75% of the top1 models reassembled by M-SADA have higher TM-scores than before.

In addition, we investigated the relationship between the average TM-scores of input domain models and TM-score of the final assembled model on 296 proteins, as shown in Fig. S3a. However, in practical scenarios, the native structure is often unavailable. We further investigated the correlation between the average pLDDT scores of input domain models and the TM-score of the final assembled model, as shown in Fig. S3b. Given that the M-SADA algorithm does not alter the structure of user-provided domain models, the quality of these input models influences the accuracy of the full-chain assembled models. From Fig. S3b, the correlation between the accuracy of input domain models and the accuracy of full-chain multidomain protein structures assembled by M-SADA is evident, and the Pearson correlation coefficient is 0.68. In addition, the identification of analogous templates is based on the domain structures, meaning that the accuracy of domain models has a consequential impact on the quality of template identification. Furthermore, our inter-domain distance prediction network, DeepIDDP, also leverages structural features of domains. Therefore, the quality of domain models will affect the prediction accuracy of DeepIDDP [20]. Inputting high-quality domain models (or domain models with high pLDDT scores) facilitates M-SADA in generating more accurate full-chain structures.

3.3. Overall results of domain assembly on experimental domains

The accuracy of the single-domain model affects the performance of the domain assembly methods. In order to objectively compare the

Table 3
Results of domain assembly of M-SADA, SADA, DEMO and AIDA on 302 multidomain proteins.

Methods	Average TM-score	Median TM-score	#TM-score ≥ 0.8
M-SADA-best	0.86	0.92	216
M-SADA-top1	0.83	0.90	200
SADA	0.79	0.83	161
DEMO	0.73	0.73	119
AIDA	0.62	0.62	33

Note: #TM-score ≥ 0.8 represents the number of models with TM-score ≥ 0.8.

domain assembly performance of M-SADA with other methods, M-SADA is tested on 302 multidomain proteins and compared with AIDA [13], DEMO [14] and SADA [15]. This benchmark includes 137 proteins with 2 domains, 61 proteins with 3domain, 36 proteins with ≥ 4 domains and 68 proteins with discontinuous domain.

We reassemble the 302 multidomain proteins based on experimentally solved domains, where 30% sequence identity used to filter available templates to minimize the homologous contaminations [14]. Table 3 summarizes the modelling results of M-SADA and other methods, and TM-score is calculated by comparing assembled model and native structure. The details for each type of multidomain proteins are shown in Table S4.

Compared to the average TM-score of previously developed SADA (0.79), the TM-score of the M-SADA-top1 model on 302 proteins achieves an improvement of 5.1% to 0.83, where the *P*-value between M-SADA and SADA is 1.64e-9. Compared to DEMO and AIDA, the accuracy of M-SADA models is improved by 13.7% and 33.9%, respectively. In particular, for m4dom proteins, the improvement in accuracy of the full-chain models is more significant, with an average accuracy improvement of 12.1% for the M-SADA-top1 model compared to the second-best method. In addition, the results also indicate that the performance of M-SADA decreases with increasing domains, probably because the quality of the available templates decreases and the accuracy of the inter-residue distances predicted by DeepIDDP decreases with increasing domains, which affects the precision of the energy functions.

Overall, the full-chain models assembled by M-SADA are better, and the *P*-values of SADA, DEMO and AIDA indicate statistically significant differences between the methods. The reason for such discrepancy is that M-SADA uses different types of templates and a more advanced inter-domain distance map to generate multiple energy functions that contain more potential orientations between domains.

3.4. Evaluation of inter-domain distance prediction

In this section, we analyse the difference between the accuracy of the inter-domain distances predicted by the DeepIDDP [20] and that of GeomNet [36] using in SADA on 302 multidomain proteins, and only the pairs of atoms with predicted distance ≤15 Å and belonged to different domains are analysed. The results are summarized in Table 4. Here, 2dom, 3dom, m4dom and 2dis are used to refer to the multidomain proteins with 2 domains, 3 domains, ≥ 4 domains and discontinuous domains, respectively.

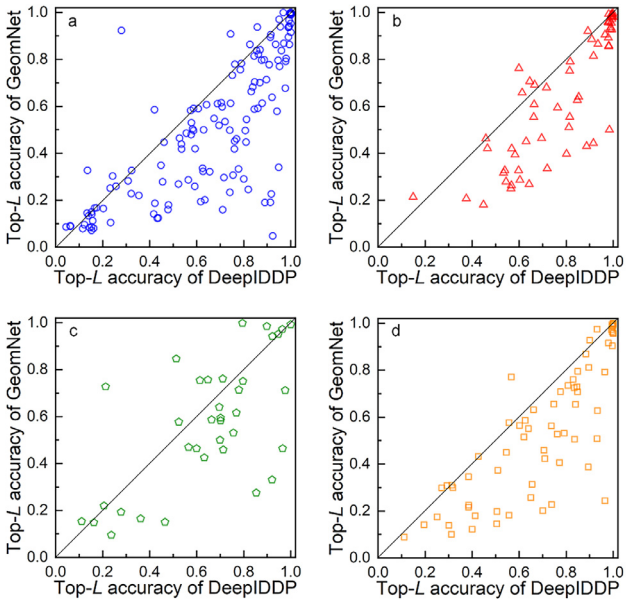

Fig. 5. Comparison between DeepIDDP in M-SADA and GeomNet in SADA on each protein. (a), (b), (c) and (d) represent the comparison on 137 proteins with 2 domains, 61 proteins with 3domains, 36 proteins with ≥ 4 domains and 68 proteins with discontinuous domains, respectively.

Overall, DeepIDDP outperforms GeomNet on the multidomain protein test set. According to the results in Table 4, we found that for those proteins with ≥ 4 domains, the improvement of mean absolute error (MAE) metrics was not as significant as that of 2dom, 3dom and 2dis proteins, but was still 13.5% higher than GeomNet. This is due to the limited number of multidomain protein structures available in the PDB. In particular, when the number of domains is greater than three, the number of multidomain proteins available with training is significantly reduced [15], which may be the main reason for the reduced accuracy of DeepIDDP on m4dom proteins compared to other types of multidomain proteins. Fig. 5 shows top-*L* accuracy comparisons between DeepIDDP and that GeomNet on each test protein, which intuitively shows the improved accuracy of DeepIDDP on different types of proteins compared to GeomNet. The DeepIDDP, trends for top-*L* accuracy and MAE are the same for different types of multidomain proteins.

Table S5 lists detailed results of M-SADA when predicted inter-domain distance is not used, which demonstrates that DeepIDDP predicted distance contributes to M-SADA accuracy. Compared to not using predicted distance, the accuracy of M-SADA is improved by 7.5% when the predicted inter-domain distance information is used to guide the domain assembly. In particular, the improvement is more significant as the number of domains increasing, with the average TM-score improving from 0.73 to 0.79 (an 8.2% increase) for the 61 proteins with

Table 4
Summary of the results of inter-domain distance prediction for 302 multidomain proteins.

Method	MAE					Top-*L*	Precision	Recall
	2dom	3dom	m4dom	2dis	All			
DeepIDDP	2.66	2.80	3.26	2.82	2.80	0.69	0.54	0.67
GeomNet	3.46	3.68	3.70	3.54	3.55	0.56	0.42	0.62

Note: MAE represents the mean absolute error of all residue pairs between domains, where the distance of the residue pairs is less than 15 Å. Top-*L* represents the prediction accuracy of the top *L* for the inter-domain residues.

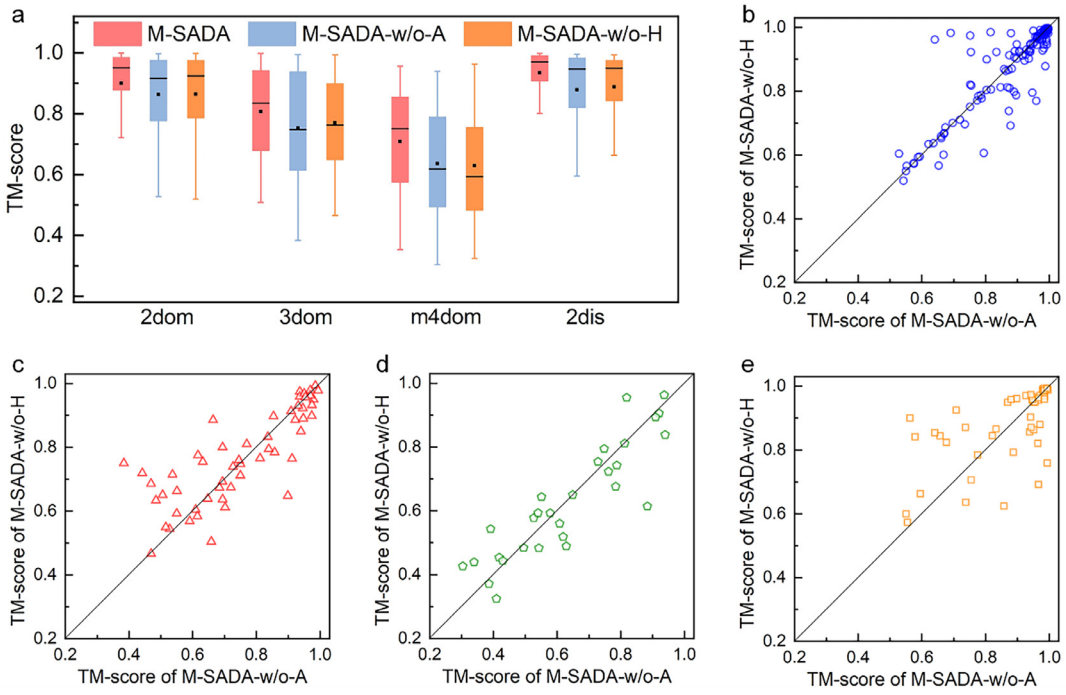

Fig. 6. Domain assembly results of M-SADA using different templates. (a) Boxplot for the TM-score of the best models generated using M-SADA, M-SADA-w/o-A and M-SADA-w/o-H on 249 proteins. M-SADA-w/o-A and M-SADA-w/o-H refer to M-SADA without analogous and homologous templates, respectively. The black horizontal line and square are the median and mean, respectively. (b), (c), (d) and (e) are accuracy comparisons between M-SADA without utilizing homologous and analogous templates for proteins with 2 domain, 3domains, ≥ 4 domains and discontinuous domains, respectively.

3 domains and from 0.55 to 0.68 (an 8.2% increase) for the 36 proteins with ≥ 4 domains.

3.5. Effect of homologous and analogous templates on M-SADA

In order to investigate the effect of different types of templates on the accuracy of M-SADA, here we test the accuracy of M-SADA only using templates identified by different methods. To eliminate the effect of the accuracy of the domain structure on M-SADA, we test M-SADA on the multidomain proteins with experimental domains, and only the best models are discussed including the models in hybrid population.

53 of 302 proteins have no available homologous templates when 30% sequence identity is used to exclude templates. Therefore, on the remaining 249 proteins, we investigate the effect of different templates on M-SADA. Table S6 contains M-SADA accuracy of only using templates identified by different methods that demonstrate that different templates contribute to M-SADA performance. Overall, the average TM-score is 0.82 when M-SADA only used structural analogous templates or homologous templates. However, the average TM-score achieve 0.86 when M-SADA utilizes both types of templates, a 4.9% improvement over using one type of templates. Fig. 6a shows TM-score boxplot of different templates contribute to M-SADA, and TM-score comparisons between using analogous and homologous templates are shown in Fig. 6b-e.

On 54 3dom and 29 m4dom proteins, compared to M-SADA-w/o-A, the accuracy of M-SASD is improved by 8.0% (0.75 to 0.81) and 10.9% (0.64 to 0.71), respectively, and compared to M-SADA-w/o-H, the accuracy of M-SASD is improved by 5.2% (0.77 to 0.81) and 12.7% (0.63 to 0.71), respectively. Fig. 6b-e shows that different types of templates may be complementary, and provide M-SADA with different types of template information, improving the M-SADA accuracy when they are combined, especially for the proteins with ≥ 3 domains. For some targets, particularly novel targets, it may not be possible to detect available homologous templates through homology search tools. However, templates identified at the domain level through structural alignment may provide valuable inter-domain interaction information. For some other targets, the inter-domain information provided by analogous templates was not as comprehensive as that from homologous templates. For instance, in the case of the protein 3l76A with 2 domains, the analogous templates identified by M-SADA in the MPDB did not offer high-quality inter-domain information, resulting in a TM-score of only 0.79 for the assembled models. However, the homologous template identified by M-SADA in the AlphaFold DB90 (AlphaFold2 models) provided reliable inter-domain information, resulting in a TM-score of 0.90 for the assembled model based on homologous templates. The TM-score of the final model reached 0.96 when M-SADA utilized information from both types of templates.

We assembled the 302 test proteins by setting weights of template-based energy of M-SADA to 0 to examine the effect of templates on M-SADA accuracy, and Table S7 contains the detailed results. The average accuracy of M-SADA decreased from 0.86 to 0.79. Especially for the 36 m4dom proteins, the accuracy decreased by 13.2% (from 0.68 to 0.59). These results also demonstrate the combination of templates and predicted inter-domain distances is crucial for the performance of M-SADA, especially for the proteins with more domains. In addition, we also discuss the effect of the models in AlphaFold DB on the performance of M-SADA in Text S6, and results indicate using protein models in AlphaFold DB is helpful in improving the domain assembly accuracy of M-SADA.

3.6. CASP15 multidomain targets

M-SADA is compared with SADA, DEMO, AlphaFold2, trRosetta [37,38] and ESMFold [39] on all CASP15 multidomain targets, where the full-chain structures of M-SADA, SADA and DEMO are generated based on structural domains individually predicted by AlphaFold2. Here, M-SADA, SADA, DEMO and AlphaFold2 are run by using time cut-off (2022–05–01) in PDB. trRosetta is one of the representative methods in energy minimization protein structure prediction methods

with deep learning predicted constraints [40]. ESMFold is an end-to-end approach based on language models that rapidly infers the structure from the sequence without the need for MSA and template information, which is one of the representative methods in protein language models [40]. Therefore, we additionally compared these two methods with M-SADA on CASP15 multidomain protein full-chain structure modelling. Table S8 summarizes the prediction accuracy of each method for each multidomain target, and comparisons between them are shown in Fig. S4.

In terms of average TM-score, M-SADA-top1 obtains 0.633, which is 5.3%, 20.6%, 7.5%, 64.8% and 31.1% higher than SADA (0.601), DEMO (0.525), AlphaFold2 (0.589), trRosetta (0.384) and ESMFold (0.483), respectively. The accuracy of the full-chain models assembled by M-SADA is higher than the accuracy of the full-chain models predicted (or assembled) by DEMO, trRosetta and ESMFold on all targets. As shown in Fig. S4, M-SADA obtained better full-chain models on 7 targets compared to AlphaFold2. For targets T1137s1, T1137s2, T1137s3, T1137s4 and T1137s6, TM-scores for the full-chain structures built by all methods are low, all below 0.50. These results also indicate that CASP15 targets are challenging. On the 5 targets, M-SADA performed unsatisfactory, but is still comparable or better than other methods. This may be due to the inaccurate domain models predicted by AlphaFold2, which hinder M-SADA from obtaining correct domain orientations. The crystal structures of each domain extracted from native structures provided by CASP15 are further used as input and a higher accuracy (TM-score = 0.795) for all CASP15 multidomain targets is achieved with M-SADA. Results show that TM-scores of full-chain structures for T1137s1, T1137s2, T1137s3, T1137s4 and T1137s6 are significantly improved by M-SADA using experimental domains. For the target T1137s5, the top1 model generated by M-SADA, which had a TM-score of 0.462, did not attain the accuracy of SADA model (0.527) and AlphaFold2 model (0.517). This difference is partly due to the accuracy of the input domain models, which affects the performance of M-SADA. The TM-scores for the domain models predicted for T1137s5 were 0.74 and 0.54, respectively. An improvement in accuracy was observed upon the incorporation of experimental domains. The structure of this protein is not compact but rather extended, making the inference of inter-domain distances challenging and leading to lower accuracy of assembled models. Moreover, within models generated by M-SADA, the most accurate full-chain model achieved a TM-score of 0.533, but this model was not selected by the DeepUMQA2. For the target T1121, the optimal model produced by M-SADA, with a TM-score of 0.935, was comparable in accuracy to models predicted by AlphaFold2, but the best model of M-SADA was not selected by the DeepUMQA2. Results of CASP15 multidomain targets indicate M-SADA predicts the domain orientations of full-chain structures better than the advanced end-to-end method. We also show that M-SADA can predict new domain arrangements in CASP15 when individual domain structures are correct.

4. Conclusion

Deep learning techniques have enabled remarkable progress in protein structure prediction, allowing high-accuracy models for most single-domain proteins and some multidomain proteins. However, compared with single-domain protein prediction, the full-chain modelling of multidomain proteins remains relatively unreliable. More importantly, the domains of multidomain proteins frequently interact with each other, and therefore, multidomain proteins usually have multiple conformational states. The modelling of multidomain proteins with multiple conformational states is also essential for further studying protein functions and interaction mechanisms. However, fewer computational approaches have been developed to address this problem.

This work develops a multiple conformational states domain assembly method, M-SADA, which enables the modelling of different conformational states of multidomain proteins. In M-SADA, the high-quality structural analogous templates are first identified from the MPDB based on the input domain structures, and the available homologous templates are identified from MPDB, PDB and AlphaFold DB. Based on the different templates, specific energy functions combined with deep learning predicted restraints are designed to guide domain assembly. Finally, a multiple population-based evolutionary algorithm is proposed to explore domain orientation based on multiple energy functions.

M-SADA is tested on 72 multidomain proteins with multiple conformational states. Consequently, M-SADA correctly generates full-chain models with different states on the majority of multidomain proteins and is 16.0% higher than AlphaFold2 (0.75) in terms of the average TM-score, and 29/72 (40.3%) of proteins can be assembled with a TM-score > 0.90 for highly distinct conformational states with M-SADA while AlphaFold2 does so in only 2/72 (2.8%) of proteins. M-SADA exhibits superior performance in modelling multiple conformational states of multidomain proteins mainly because of the following reasons: (i) it generates multiple conformational landscapes, and (ii) it achieves cooperation and optimization between different landscapes through a multiple population-based evolutionary algorithm. M-SADA is also applied to reassemble 296 human multidomain proteins from AlphaFold DB, and most cases achieve better models. In particular, for the 84 proteins cannot be accurately predicted by AlphaFold2 (TM-score < 0.8), the top1 full-chain models of M-SADA achieve an average TM-score of 0.729, an improvement of 18.0% over the accuracy of AlphaFold2 models (0.618). Domain assembly of M-SADA is also tested on 302 multidomain proteins with experimental domains, and results show that M-SADA significantly better than the advanced methods. Meanwhile, the results of ablation experiment indicate that sequential homologues and structural analogues are complementary.

In spite of some achievements reported in M-SADA, it has the potential for improving in future. For these multidomain proteins with weak domain-domain interaction, the performance of M-SADA was not satisfactory. For this type of protein, the inference of inter-domain distances is challenging. Even with the use of experimental domains, there remains significant room for improvement in the assembly accuracy. In addition, multiple MSAs are not considered by DeepIDDP. Therefore, the inter-domain distance map predicted by DeepIDDP lacks diversity. If diverse inter-domain distance maps can be predicted, the ability to model multidomain proteins with multiple conformational states will be likely improved. In the M-SADA assembly process, the domain structures are kept invariant while the domain structures may change due to the interactions between domains in the real protein folding process. Therefore, considering the interactions between domains and performing local optimization of structural domains in the assembly process may improve the accuracy of both single-domain structures and full-chain structures. Moreover, a gap still exists between the best model of M-SADA and the top1 model assessed by DeepUMQA2. Therefore, further improving the performance of model quality assessment for multidomain proteins is also the next direction of our research. Since there is some similarity between the interactions between domains of multidomain proteins and the interactions between the monomers of protein complexes [41]. The M-SADA has some potential to be extended to assemble protein monomers, which is expected to assemble larger protein complex structures.

Research data

M-SADA server is freely available at http://zhanglab-bioinf.com/M-SADA/.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

This work has been supported by the National Science and Technology Major Project (2022ZD0115103), the National Nature Science Foundation of China (62173304 and 62203389), the Key Project of Zhejiang Provincial Natural Science Foundation of China (LZ20F030002).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.fmre.2024.05.003.

References

[1] J. Jumper, R. Evans, A. Pritzel, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (2021) 583–589.
[2] A. Kryshtafovych, T. Schwede, M. Topf, et al., Critical assessment of methods of protein structure prediction (CASP)—round XIV, Proteins: Struct., Funct., Bioinf. 89 (2021) 1607–1617.
[3] X. Zhou, W. Zheng, Y. Li, et al., I-TASSER-MTD: A deep-learning-based platform for multi-domain protein structure and function prediction, Nat. Protoc. 17 (2022) 2326–2353.
[4] R. Pearce, Y. Zhang, Toward the solution of the protein structure prediction problem, J. Biol. Chem. 297 (2021) 100870.
[5] J. Skolnick, M. Gao, H. Zhou, et al., AlphaFold 2: Why It works and its implications for understanding the relationships of protein sequence, structure, and function, J. Chem. Inf. Model. 61 (2021) 4827–4831.
[6] M. Varadi, S. Anyango, M. Deshpande, et al., AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models, Nucleic Acids Res. 50 (2022) D439–D444.
[7] K. Zhao, Y. Xia, F. Zhang, et al., Protein structure and folding pathway prediction based on remote homologs recognition using PAthreader, Commun. Biol. 6 (2023) 243.
[8] D.T. Jones, J.M. Thornton, The impact of AlphaFold2 one year on, Nat. Methods. 19 (2022) 15–20.
[9] X. Zhou, C. Peng, W. Zheng, et al., DEMO2: Assemble multi-domain protein structures by coupling analogous template alignments with deep-learning inter-domain restraint prediction, Nucleic Acids Res. 50 (2022) W235–W245.
[10] J. Choi, T. Park, S. Yul Lee, et al., GalaxyDomDock: An Ab initio domain–domain docking web server for multi-domain protein structure prediction, J. Mol. Biol. 434 (2022) 167508.
[11] S. Subramaniam, G.J. Kleywegt, A paradigm shift in structural biology, Nat. Methods. 19 (2022) 20–23.
[12] A.M. Wollacott, A. Zanghellini, P. Murphy, et al., Prediction of structures of multidomain proteins from structures of the individual domains, Protein Sci. 16 (2007) 165–175.
[13] D. Xu, L. Jaroszewski, Z. Li, et al., AIDA: ab initio domain assembly for automated multi-domain protein structure prediction and domain–domain interaction prediction, Bioinformatics 31 (2015) 2098–2105.
[14] X. Zhou, J. Hu, C. Zhang, et al., Assembling multidomain protein structures through analogous global structural alignments, Proc. Natl. Acad. Sci. USA 116 (2019) 15930.
[15] C. Peng, X. Zhou, Y. Xia, et al., Structural analogue-based protein structure domain assembly assisted by deep learning, Bioinformatics 38 (2022) 4513–4521.
[16] X. Zhou, Y. Li, C. Zhang, et al., Progressive assembly of multi-domain protein structures from cryo-EM density maps, Nat. Comput. Sci. 2 (2022) 265–275.
[17] Z. Zhang, Y. Cai, B. Zhang, et al., DEMO-EM2: Assembling protein complex structures from cryo-EM maps through intertwined chain and domain fitting, Briefings Bioinf. 25 (2024) bbae113.
[18] M. Schauperl, R.A. Denny, AI-based protein structure prediction in drug discovery: Impacts and challenges, J. Chem. Inf. Model. 62 (2022) 3142–3156.
[19] C. Peng, X. Zhou, Y. Xia, Y. Zhang, G. Zhang, MPDB: A unified multi-domain protein structure database integrating structural analogue detection, bioRxiv. (2021). https://doi.org/10.1101/2021.10.27.466092.
[20] F. Ge, C. Peng, X. Cui, et al., Inter-domain distance prediction based on deep learning for domain assembly, Briefings Bioinf. (2023) bbad100.
[21] J. Liu, K. Zhao, G. Zhang, Improved model quality assessment using sequence and structural information by enhanced deep neural networks, Briefings Bioinf. (2022) bbac507.
[22] Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on the TM-score, Nucleic Acids Res. 33 (2005) 2302–2309.
[23] J. Boekhorst, B. Snel, Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties, BMC Bioinf. 8 (2007) 356 -356.
[24] H.M. Berman, J. Westbrook, Z. Feng, et al., The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[25] A. Bateman, M.J. Martin, C. O'Donovan, et al., UniProt: A hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212.
[26] J. Mistry, R.D. Finn, S.R. Eddy, et al., Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions, Nucleic Acids Res. 41 (2013).
[27] S.R. Eddy, A new generation of homology search tools based on probabilistic inference, Genome informatics, Int. Conf. Genome Inf. 23 (2009) 205–211.
[28] R. Storn, K. Price, Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces, J. Glob. Optim. 11 (1997) 341–359.
[29] X. Zhou, C. Peng, J. Liu, et al., Underestimation-assisted global-local cooperative differential evolution and the application to protein structure prediction, IEEE Trans. Evol. Comput. 24 (2020) 536–550.
[30] X. Li, L. Wang, Q. Jiang, et al., Differential evolution algorithm with multi-population cooperation and multi-strategy integration, Neurocomputing 421 (2021) 285–302.
[31] P.J. Ballester, W.G. Richards, Ultrafast shape recognition to search compound databases for similar molecular shapes, J. Comput. Chem. 28 (2007) 1711–1723.
[32] X. Hao, G. Zhang, X. Zhou, et al., A novel method using abstract convex underestimation in Ab-initio protein structure prediction for guiding search in conformational feature space, IEEE-ACM Trans. Comput. Biol. Bioinform. 13 (2016) 887–900.
[33] M. Hou, S. Jin, X. Cui, et al., Protein multiple conformation prediction using multi-objective evolution algorithm, Interdisciplinary Sciences: Computational Life Sciences, 2024.
[34] J. Xu, Y. Zhang, How significant is a protein structure similarity with TM-score=0.5? Bioinformatics 26 (2010) 889–895.
[35] J.R. Horton, H. Wang, M.Y. Mabuchi, et al., Modification-dependent restriction endonuclease, MspJI, flips 5-methylcytosine out of the DNA helix, Nucleic Acids Res. 42 (2014) 12092–12101.
[36] J. Liu, G. He, K. Zhao, G. Zhang, De novo protein structure prediction by incremental inter-residue geometries prediction and model quality assessment using deep learning, bioRxiv (2022). https://doi.org/10.1101/2022.01.11.475831.
[37] J. Yang, I. Anishchenko, H. Park, et al., Improved protein structure prediction using predicted interresidue orientations, Proc. Natl. Acad. Sci. USA. 117 (2020) 1496.
[38] H. Su, W. Wang, Z. Du, et al., Improved protein structure prediction using a new multi-scale network and homologous templates, Adv. Sci. 8 (2021) e2102592.
[39] Z.M. Lin, H. Akin, R.S. Rao, et al., Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (2023) 1123–1130.
[40] C. Peng, F. Liang, Y. Xia, et al., Recent advances and challenges in protein structure prediction, J. Chem. Inf. Model. 64 (2023) 76–95.
[41] Y. Xia, K. Zhao, D. Liu, et al., Multi-domain and complex protein structure prediction using inter-domain interactions from deep learning, Commun. Biol. 6 (2023). https://doi.org/10.1038/s42003-023-05610-7.

Author profile

Chun-Xiang Peng received the Ph.D. degree in control science and engineering from College of Information Engineering from Zhejiang University of Technology in 2023. His research interests include intelligent optimization, computational biology and bioinformatics.

Gui-Jun Zhang received the Ph.D. degree in control theory and control engineering from Shanghai Jiao Tong University in 2004. He is currently a professor in the College of Information Engineering, Zhejiang University of Technology. His research interests include intelligent information processing, optimization theory and algorithm design, machine learning, and bioinformatics. In recent years, he has published over 100 journal papers and three monographs (books).