Computational and Structural Biotechnology Journal 20 (2022) 4481–4489

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Obtaining protein foldability information from computational models of AlphaFold2 and RoseTTAFold

Sen Liu [a,b,c,*], Kan Wu [a,b,c], Cheng Chen [a,b,c]

[a] Key Laboratory of Fermentation Engineering (Ministry of Education) & Cooperative Innovation Center of Industrial Fermentation (Ministry of Education & Hubei Province), Hubei University of Technology, Wuhan 430068, China
[b] National "111" Center for Cellular Regulation and Molecular Pharmaceutics, Hubei University of Technology, Wuhan 430068, China
[c] Hubei Key Laboratory of Industrial Microbiology, Hubei University of Technology, Wuhan 430068, China

ARTICLE INFO

Article history:
Received 4 July 2022
Received in revised form 15 August 2022
Accepted 15 August 2022
Available online 17 August 2022

Keywords:
Protein foldability
AlphaFold
RoseTTAFold
Structure prediction
Solubility

ABSTRACT

The recent breakthrough from AlphaFold2 and RoseTTAFold set a profound milestone for solving the protein folding problem, but they were not explicitly trained to predict protein foldability, i.e., if a protein can really fold into the predicted 3D structure. We wondered if the computational models from AlphaFold2 and RoseTTAFold might carry protein foldability information. Therefore, we predicted the structural models of 159 circular permutants and 158 alanine insertion mutants of the 159-residue dihydrofolate reductase. Our data showed that although AlphaFold2 and RoseTTAFold cannot directly identify unfoldable proteins, the RMSD values of computational models are correlated with protein foldability, with higher RMSD values indicating lower protein foldability. Furthermore, this correlation is independent of secondary structures, and the RMSD values of computational models are quantitatively correlated with protein foldability but not protein functions. Additionally, using a dataset of 129 de novo designed proteins, we showed that inter-model RMSD values between AlphaFold2 models and RoseTTAFold models are a good indicator of protein foldability. At last, we showed that inter-model RMSD values are also useful for evaluating protein solubility by modeling 1664 natural proteins. Our work could be of great value to the design of novel proteins and the prediction of protein foldability.

1. Introduction

Proteins are synthesized as linear chains of amino acids, but generally they could not perform biological functions before forming specific three-dimensional (3D) conformations. The formation of the 3D structure of a peptide sequence, i.e., protein folding, is tremendously challenging because the large dimensionality and its stochasticity. However, the seminal work by Christian Anfinsen and colleagues led to the hypothesis that the peptide sequence of a protein intrinsically determines its specific 3D structure [1]. Since then, protein folding has been described as a searching for the lowest-energy conformation in the energy landscape [2]. This hypothesis further led to the extensive investigation of the protein folding problem, which includes three questions: what is the protein folding code, how to predict the 3D structure of a given peptide sequence, and what are the kinetic folding pathways of proteins [2,3]. The answers of these three questions will let us know if a protein can fold (foldability), what is the 3D conformation (fold) of a protein, and how a protein achieves its 3D conformation (folding kinetics).

With the accumulated contributions of many scientists, remarkable achievements have been made in solving the protein folding problem in the last fifty years [2]. The biggest breakthrough was recently achieved by AlphaFold2 [4] and RoseTTAFold [5]. Both methods took advantages of previous knowledges on protein folds and the recent development of computational algorithms and hardware. By incorporating physical constraints, evolutionary information, neural networks, and GPU computing, AlphaFold2 and RoseTTAFold were able to predict most protein structures (folds) comparable to experimental accuracy. Therefore, these methods are extremely helpful when we want to know the possible 3D structure of a natural protein. However, a virtually predicted model is far from a viable folded protein. It is unknown yet whether a peptide chain could be really foldable when it is virtually folded. For example, David Baker and colleagues recently

* Corresponding author at: Key Laboratory of Fermentation Engineering (Ministry of Education) & Cooperative Innovation Center of Industrial Fermentation (Ministry of Education & Hubei Province), Hubei University of Technology, Wuhan 430068, China.

E-mail address: senliu.ctgu@gmail.com (S. Liu).

https://doi.org/10.1016/j.csbj.2022.08.034
2001-0370/© 2022 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

designed 129 proteins computationally with trRosetta but discovered that 102 proteins did not fold well after expression [6].

Circular permutation is a protein engineering strategy to elucidate the structure–function relationship and folding kinetics of proteins [7]. In circular permutation, the peptide chain of a protein is rearranged by joining the *N*-terminus and the C-terminus while new termini are generated by the cleavage of a peptide bond other than the original ones (Fig. 1a). Therefore, circular permutation changes the connectivity but not the composition of residues of a protein. Besides being a research tool, circular permutation also occurs in natural proteins as an evolution strategy [8]. Studies showed that certain circular permutation affects the folding progress of a protein while keeping the overall structure and function without significant changes [7,9–11]. If the cleavage in a contiguous peptide segment results in the complete loss of the ability of the protein to fold, this sequence region is named as a folding element, which might play key roles in the formation of folding nuclei during the protein folding process [11]. Previous studies proposed that the presence but not the order of folding elements is essential for a protein to be foldable [9,11–13]. Although neither AlphaFold2 nor RoseTTAFold was designed to predict folding foldability or folding kinetics, both methods took advantages of sequence evolution information. Therefore, it is of great interest to know if these state-of-the-art computational methods could distinguish foldable circular permutants from unfoldable ones of a natural protein. That is, if protein foldability information could be extracted from predicted protein folds. A note is that protein folding kinetics is not discussed in this work.

The dihydrofolate reductase of Escherichia coli (EC 1.5.1.3; referred as DHFRcoli in this work) is an intensively studied model in protein circular permutation. By generating circular permutants at all 158 neighbored residue pairs of this 159-residue protein, Masahiro Iwakura and colleagues defined ten folding elements of DHFRcoli based on experimental data [11]. Breaking any one of these folding elements abolished the correct folding of DHFRcoli. Interestingly, these folding elements did not strictly align with the secondary structural units, or the substrate and coenzyme binding sites [11]. Nonetheless, the residues in these folding elements were correlated with the formation of the folding nuclei in the early folding events [11].

In this work, we systematically modelled the 3D structures of the 159 circular permutants of DHFRcoli with both AlphaFold2 and RoseTTAFold. Although not all permutants were experimentally foldable and active, they were all virtually folded and highly resembled the X-ray structure of the wild-type protein, indicating that these computational methods cannot evaluate the foldability of proteins. However, when aligned to the wild-type structure, the theoretical models of unfoldable permutants had larger RMSD values than the foldable permutants. Furthermore, the correlation between RMSD values and foldability was independent of secondary structures. These findings were confirmed by modeling another 158 alanine insertion mutants of DHFRcoli [14] using both AlphaFold2 and RoseTTAFold. Furthermore, using a dataset of 129 de novo designed proteins, we found that inter-model RMSD values between AlphaFold2 models and RoseTTAFold models were a good indicator for evaluating protein foldability when no wild-type structure information is available. At last, we showed that inter-model RMSD values could also be used to evaluate protein solubility based on a dataset of 1664 natural proteins. Our work indicated that the computational models from AlphaFold2 and RoseTTAFold could be used to evaluate protein foldability.

2. Methods

2.1. Construction of the sequences of the circular permutants

The circular permutated sequences were constructed according to the construction methods described in [10,11,21]. A five-glycine (GGGGG) sequence flanked the C-terminal residue (residue 159) of the wild-type sequence, connecting the downstream sequence in the permutated protein. When the *N*-terminal residue was not methionine (M), the extra M was added as required in protein expression. However, as describe in [21], an extra M was also added for the DHFRcoli-20 sequence since this was done in protein expression to protect the removal of M20 by the methionyl-aminopeptidase.

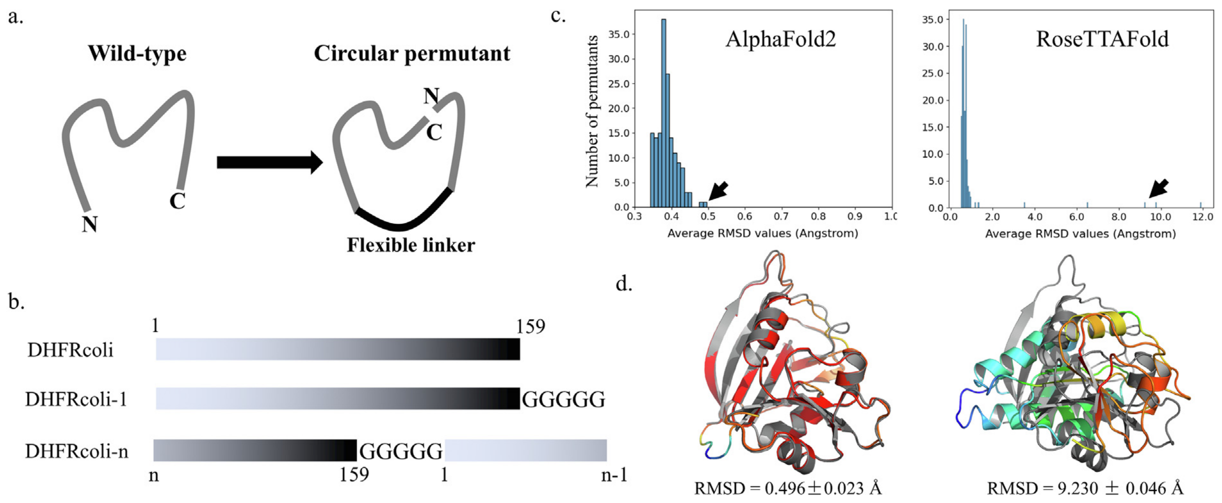

Fig. 1. Construction of the circular permutants and model quality. (a) Scheme of circulate permutation. The original termini are connected by a flexible peptide linker, and new termini are introduced by the cleavage of a peptide bond elsewhere. (b) Scheme of the sequence construction of the DHFRcoli circular permutants. The marked numbers are residue numbers. A five-glycine linker was added to connect the original termini. (c) The distributions of the backbone RMSD values of the 159 AlphaFold circular permutants. The pointed models by the arrows are shown in D. (d) A predicted model from AlphaFold2 for the experimentally unfoldable circular permutant DHFRcoli-92, and A predicted model from RoseTTAFold for the experimentally foldable and active circular permutant DHFRcoli-79. The RMSD values were averaged from five models and calculated as the backbone RMSD aligned to the X-ray structure (PDB ID: 1RX4). The X-ray structure of the wild-type DHFRcoli is colored in gray and the models are colored as spectrum by pLDDT values of Cα atoms (red: higher pLDDT values; blue: lower pLDDT values). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.2. Construction of the sequences of the alanine insertion mutants

As describe in [14], an alanine insertion mutant had an alanine residue inserted between two neighbored residues. For the 195-residue DHFRcoli, there would be 158 insertion sites. However, When the nth residue is an alanine, (n-1)An and nA(n + 1) have the same sequence. So the actual number of the alanine insertion mutants were 145 since DHFRcoli contains 13 alanine residues.

2.3. Acquiring of the experimental data

The experimental data of the circularly permutants were obtained from the reference [11]. From the original paper, we could not determine the relative expression level (solubility) of the permutants, so we assigned foldability (foldable vs unfoldable) according to the enzymatic activity data and the CD data. The experimental data of the permutant 54 (DHFRcoli-54) was missing, but it was assigned as foldable since this position was not included in any folding element [11]. The experimental data of the alanine insertion mutants were obtained from the reference [14]. The alanine insertion mutant 67 (DHFRcoli-67A68) did not have the precipitant ratio data, but this mutant was assigned as foldable in this work, since it was not included in any folding elements. The data were extracted from the figures using WebPlotDigitizer (https://automeris.io/WebPlotDigitizer).

2.4. Acquiring of the sequences of the de novo design and E. coli

The sequences of the E. coli proteins reported in Niwa et al. [16] were downloaded from the NCBI protein sequence database by matching the JW ID, the locus tag K-12, the locus tag MG1655, and the gene name. The sequences of the de novo designs of the Baker group [6] were obtained from the associated information of the paper. The experimental data were obtained from the papers as well.

2.5. Protein structure prediction with AlphaFold2

The 3D structures of the DHFRcoli sequences were predicted with AlphaFold2 using MMseqs2 v1.2 [22]. Templates were used, and the Amber force field was used to relax the model structures. The multiple sequence alignment (MSA) mode was MMseqs2 (UniRef + Environment), and the pair mode was set as “unpaired + paired”. The recycle number was three. For each sequence, five unrelaxed and five relaxed models were generated, and the relaxed models were used in followed analyses.

2.6. Protein structure prediction with RoseTTAFold

The 3D structures of the DHFRcoli sequences were predicted with the RoseTTAFold using the end-to-end version. The checked out main version was 20,210,803 with UniRef30 HHsuite (2020.06), BFD (id30_c90), and pdb100 (2021Mar03). For each sequence, five models were generated and used for analyses.

2.7. RMSD calculation

The RMSD values between the predicted models and the crystal structure (PDB ID 1RX4) were calculated with the “align” function in Pymol [23]. Only backbone non-hydrogen atoms (CA + CB + C + O) were used in the alignment. The aligned structures were prepared in Pymol and the models were colored by pLDDT scores in spectrum.

3. Results

3.1. Foldable and unfoldable circular permutants have folded 3D models from AlphaFold2 and RoseTTAFold

DHFRcoli is a 159-residue protein with well characterized crystal structures. As a control, we modelled the structure of the wild-type protein using AlphaFold2 and RoseTTAFold. The model quality was evaluated by the backbone RMSD value between the model and the X-ray structure (PDB ID 1RX4). The RMSD values of the AlphaFold2 model and the RoseTTAFold model were 0.369 ± 0.022 Å and 0.529 ± 0.003 Å respectively (Figure S1), indicating that both methods recaptured the native structure with high accuracy.

In 158 circular permutants, a five-glycine linker was inserted between the original *N*-terminal and C-terminal residues [11]. Therefore, this five-glycine linker was also appended to the C-terminus of the wild-type sequence to serve as the reference sequence. This linker was optimized and did not perturb the core structure and function of DHFRcoli according to the previous investigation [10] and an X-ray structure containing four glycine residues at the C-terminus (PDB ID: 5UII). Since this sequence also could be considered as an additional circular permutant with the first residue being the original *N*-terminus, it was referred as DHFRcoli-1, leading to 159 circular permutants in total. Accordingly, the circular permutant with the nth residue as the *N*-terminus was named as DHFRcoli-n in this work (Fig. 1b). Among the 159 permutants (including DHFRcoli-1), 86 were assigned as foldable and 73 were unfoldable based on the experimental data from trimethoprim (TMP) sensitivity assay, activity assay, and circular dichroism (CD) measurement [11] (Table S1). However, the RMSD values showed that AlphaFold2 gave all permutants 3D models closely resembling the wild-type structure (Fig. 1c & 1d). Some models from RoseTTAFold had large RMSD values and were very different from the wild-type structure (Fig. 1c), but a part of them were actually foldable and active in experimental assays (Fig. 1d & S2). This data indicated that, although AlphaFold2 and RoseTTAFold generated computationally folded models for all DHFRcoli permutants, they were unable to classify foldable and unfoldable permutants. Therefore, whether AlphaFold2 and RoseTTAFold generate well-folded models or not is not an indicator of the actual foldability of a peptide sequence.

3.2. RMSD values of the DHFRcoli circular permutants correlate with their foldability

Highly flexible proteins might encounter frustration during folding, and permutation could introduce or eliminate this kind of frustration, leading to the change of folding kinetics [7]. An unfoldable protein is usually kinetically trapped by unsolvable frustration. We propose that if such frustration is introduced in to a DHFRcoli permutant, it might be possible to notice relatively large structural inconsistency in these models. This hypothesis could be true for AlphaFold2 and RoseTTAFold models, because frustrated residue-residue contact information might be less represented in their training data based on experimental structures from foldable proteins and then less accurately predicted. From the modeling in the last section, we noticed that the RoseTTAFold models of the DHFRcoli permutants were less accurate than the AlphaFold2 models (Fig. 1c, 1d & S2), so we only used the AlphaFold2 models for further analyses in this and the next sections. Indeed, we noticed that unfoldable permutants had larger RMSD values when aligned to the wild-type structure (Fig. 2a). On the contrary, the pLDDT values, a model quality evaluation score in AlphaFold2, were not distinguishable between these two groups (Fig. 2b). Since DHFRcoli-1 only had five glycine residues appended

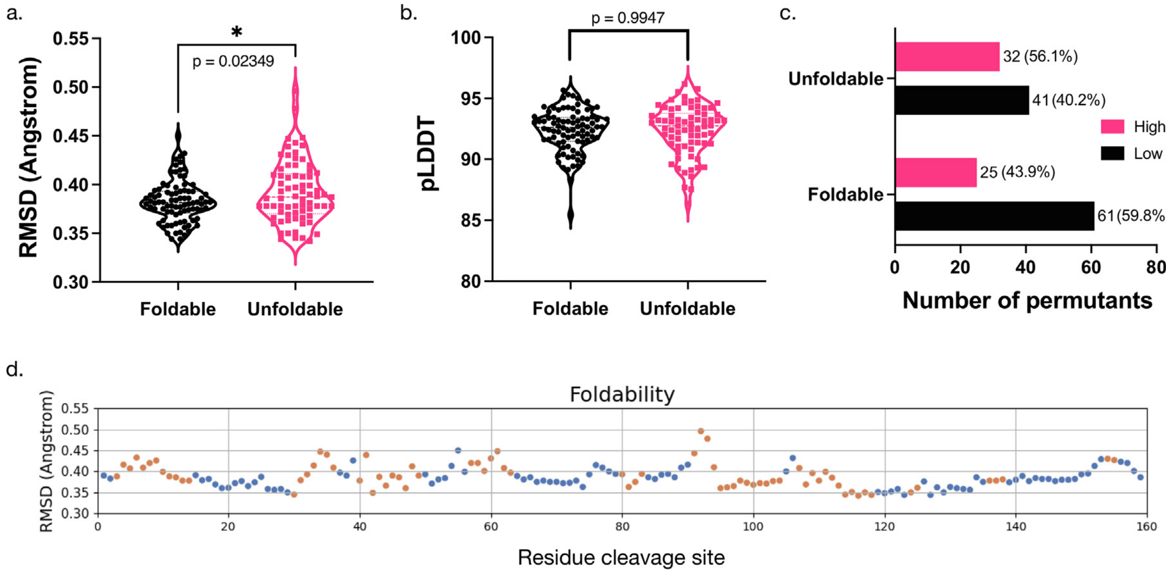

Fig. 2. The RMSD variations of the AlphaFold2 models of DHFRcoli circular permutants. (a) The comparison of the model RMSD values of experimentally foldable and unfoldable permutants. The statistical p value was calculated with the unpaired *t*-test. (b) No difference in the average pLDDT values of the models of the foldable and unfoldable permutants. The average pLDDT value was calculated from the pLDDT values of all alpha carbons in a model. (c) The amounts of the foldable and unfoldable permutants in the Low RMSD group and the High RMSD group. The numbers beside the bar are the permutant number and the percentage in the corresponding group. (d) Scatter plot of the RMSD values of the 159 DHFRcoli circular permutants. The blue dots indicate the experimentally foldable permutants, and the orange dots indicate the unfoldable ones. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to the C-terminus and minimally affecting the structure, we chose to use the RMSD value of DHFRcoli-1 as a threshold to split all permutants into two categories: the "High RMSD" group contains the permutants with RMSD values higher than that of DHFR-1, and the "Low RMSD" group contains the rest. Then it is noticeable that the Low RMSD group had more foldable permutants (59.8 %) than the High RMSD group (43.9 %) (Fig. 2c, 2d). Meanwhile, the latter contained more unfoldable permutants than the former (56.1 % v.s. 40.2 %). We calculated the standard deviations from the five models of each permutant, and the Low RMSD permutants also had a smaller average standard deviation (0.018 Å) than the High RMSD permutants (0.032 Å). A note is that the original paper [11] mentioned that some foldable permutants were refolded in vitro from inclusion bodies, but the paper did not provide enough information that could be used to tell if and how many permutants in the High RMSD group were among them.

3.3. The correlation between RMSD values and foldability weakly depends on structural elements

Both experimental and theoretical data support the formation of folding nuclei during protein folding [15]. Based on the cleavage sites of unfoldable circular permutants of DHFRcoli, Iwakura et al. assigned the 73 corresponding DHFRcoli residues to ten "Folding Elements" that might play important roles in the formation of folding nuclei [11]. Accordingly, the other 86 residues do not belong to Folding Elements (Fig. 3a). Since secondary structure units (alpha helices and beta sheets) are the main structural elements of folded proteins, we asked if the High RMSD group contains more residues within secondary structure units than the Low RMSD group.

Among the 159 residues of DHFRcoli, 79 (49.7 %) residues form secondary structure units (alpha helices and beta sheets), and 80 (50.3 %) residues belong to random coils. As shown in Fig. 3b, the 79 permutants with the cutting site in secondary structure units, 25 (31.6 %) were foldable and 54 (68.4 %) were unfoldable. The 80 permutants with the cutting site outside of secondary structure units, 61 (76.2 %) were foldable and 19 (23.8 %) were unfoldable. Therefore, bond cleavage within secondary structure units resulted in unfoldable permutants with higher probability. When the cutting site was within secondary structure units, there were more foldable permutants in the Low RMSD group than that in the High RMSD group (15 v.s. 10). Similarly, when the cutting site was outside of secondary structure units, there were also more foldable permutants in the Low RMSD group than that in the High RMSD group (46 v.s. 15). When the secondary structure rule and the model RMSD rule were combined, the percentage of foldable permutants in the Low RMSD group (34.1 %, or 15/44) was higher than the percentage of the foldable permutants in the High RMSD group (28.6 %, or 10/35). When the cutting site was outside of secondary structure units, these percentages of foldable permutants were 79.3 % (46/58) and 68.2 % (15/22) respectively. Therefore, no matter where the cutting site is, the permutants in the Low RMSD group has higher foldability than those in the High RMSD group.

3.4. Alanine insertion confirms the conclusions from circular permutation

To further investigate the folding of DHFRcoli, Shiba et al. [14] systematically constructed 158 alanine insertion mutants of DHFRcoli. They obtained the precipitant ratios of all mutants by comparing the fluorescence intensities of the protein bands on denatured SDS-PAGE gels. A mutant with a precipitant ratio (protein in supernatant/total protein) <60 % was defined as foldable by Shiba et al. [14]. Although this definition might include proteins forming molten globules, we adopted their definition in this work for consistency. In their design, an alanine residue was inserted between one pair of neighbored residues to construct an alanine insertion mutant. For example, the mutant DHFRcoli-1A2 contains an inserted alanine residue between the first and the second residues. When the n^{th} residue is an alanine, (n-1)An and nA(n + 1) have the same sequence. Therefore, the actual number of the mutants was 145 since DHFRcoli contains 13 alanine residues. Again, we computed the 3D structural models of these 145 DHFRcoli mutants with both AlphaFold2 and RoseTTAFold. When the n^{th} residue is an alanine residue, the same computational and experimental model was used for the mutants (n-1)An and nA

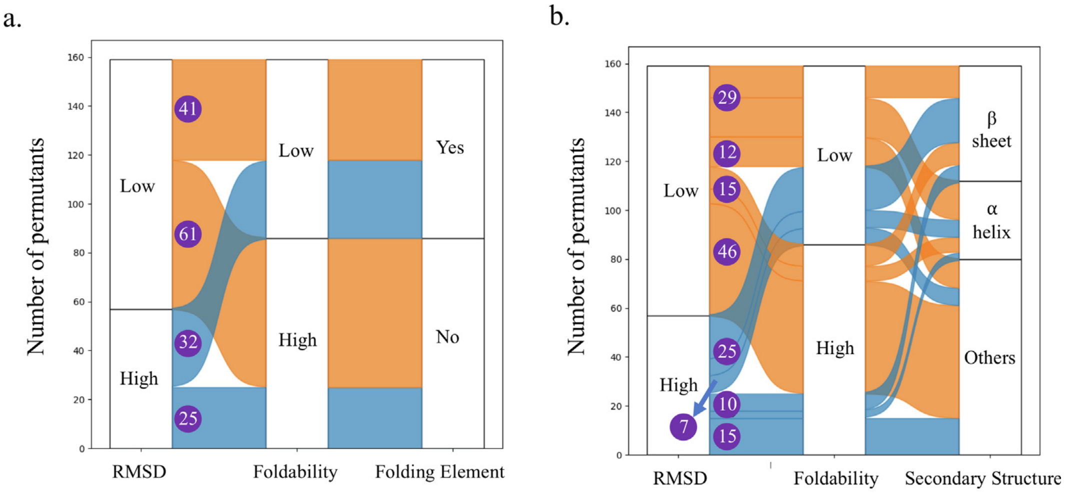

Fig. 3. The correlation between the structural variations of the AlphaFold2 models of the 159 DHFRcolil circular permutants, their experimental foldability, and structural elements. (a) Iwakura et al. assigned the DHFRcoli residues into "Folding Elements" if the corresponding circular permutant had low foldability [11]. (b) The correlation between RMSD values of the models, the foldability of the permutants, and the secondary structure units of the residues. The numbers in purple circles are the numbers of the permutants in different categories. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(n + 1). For simplicity, we refer them as 158 alanine-insertion mutants in this paper.

Similar to the circular permutants, the distribution of the backbone RMSD values of the models of these 158 alanine-insertion mutants showed that AlphaFold2 generated near-native 3D models (Fig. 4a & Table S2). However, the unfoldable mutants had a larger mean RMSD value than the foldable mutants (Fig. 4b). When we divided the models into two groups based on their RMSD values relative to the RMSD value of DHFRcoli-1A2 as in the previous sections, the mutants in the Low RMSD group had higher foldability than the mutants in the High RMSD group (Fig. 4c). The mutants in the High RMSD groups formed local clusters as well (Fig. 4d). Among these 158 mutants, there were 79 (50.0 %) mutants with an alanine residue inserted either within secondary structural elements or outside of secondary structural elements respectively. As shown in Fig. 4e, among the 79 mutants with the insertion site within secondary structure units, 23 (29.1 %) were foldable and 56 (70.9 %) were unfoldable. The 79 mutants with the insertion site outside of secondary structure units, 64 (81.0 %) were foldable and 15 (19.0 %) were unfoldable. Therefore, alanine insertion within secondary structure units resulted in unfoldable mutants with higher probability. When the insertion site was within secondary structure units, there were significantly more foldable mutants in the Low RMSD group than that in the High RMSD group (20 v.s. 3). Similarly, when the insertion site was outside of secondary structure units, there were also more foldable mutants in the Low RMSD group than that in the High RMSD group (59 v.s. 5). When the secondary structure rule and the RMSD rule were combined, the percentage of foldable mutants in the Low RMSD group (40.0 %, or 20/50) was higher than the percentage of the foldable mutants in the High RMSD group (10.3 %, or 3/29). When the insertion site was outside of secondary structure units, these percentages were 86.8 % (59/68) and 45.4 % (5/11). Therefore, no matter where the insertion site is, the mutants in the Low RMSD group has higher foldability than those in the High RMSD group. These data indicated that the conclusions from the circular permutants above are also valid for alanine insertion mutants. In addition, there was a positive correlation between the RMSD values of the circular permutants and the alanine insertion mutants (Fig. 4f), confirming the experimental data showing that alanine insertion and circular permutation were different but comparable on DHFRcoli [14].

Different from the circular permutants, RoseTTAFold generated models resembling the wild-type structure for all alanine insertion mutants (Fig. 5a & Table S2), likely due to the fact that the insertion of one alanine residue had limited consequence on multiple sequence alignment. The conclusions from AlphaFold2 held true for RoseTTAFold, including: (1) The unfoldable mutants had a larger mean RMSD value than the foldable mutants (Fig. 5b); (2) The mutants in the Low RMSD group had higher foldability than the mutants in the High RMSD group (Fig. 5c); (3) The mutants in the High RMSD groups formed local clusters (Fig. 5d). As shown in Fig. 5e, when the insertion site was outside of secondary structure units, the percentage of foldable mutants in the Low RMSD group (88.2 %, or 15/17) was higher than the percentage of the foldable mutants in the High RMSD group (79.0 %, or 49/62). When the insertion site was within secondary structure units, no mutant in the Low RMSD group was foldable due to the very limited sample size (4), and the foldable mutants in the High RMSD group were only 30.7 % (23/75). Additionally, the comparison of AlphaFold2 models and RoseTTAFold models showed that their model qualities (evaluated by their RMSD values aligned to the wild-type structure) were correlated but not identical, and the AlphaFold2 models had lower RMSD values than the RoseTTAFold models (Fig. 5f). Despite of these discrepancies, we could see that both AlphaFold2 models and RoseTTAFold models are good enough to reveal a good correlation between model RMSD values and protein foldability.

3.5. RMSD values of the computational models quantitatively correlate with protein foldability

We then asked to what extent the RMSD values of computational models quantitatively correlate with protein foldability. In the alanine insertion study, Shiba et al. [14] reported the precipitant ratios of all mutants from E. coli cell lysates. We found that the RMSD values of the computational models quantitatively correlate with the precipitant ratios with a Spearman r being 0.5421 ($p < 0.0001$) for the AlphaFold2 models and a Spearman r being 0.3554 ($p < 0.0001$) for the RoseTTAFold models (Fig. 6), indicating that RMSD values could be a predictor of protein foldability. For foldable circular permutants, Iwakura et al. [11] obtained their in vitro enzymatic activity data (k_{cat} and K_M) and conformational stability data (deltaG and m-Value). Interestingly, there were no clear correlations between the RMSD values and these data (Fig-

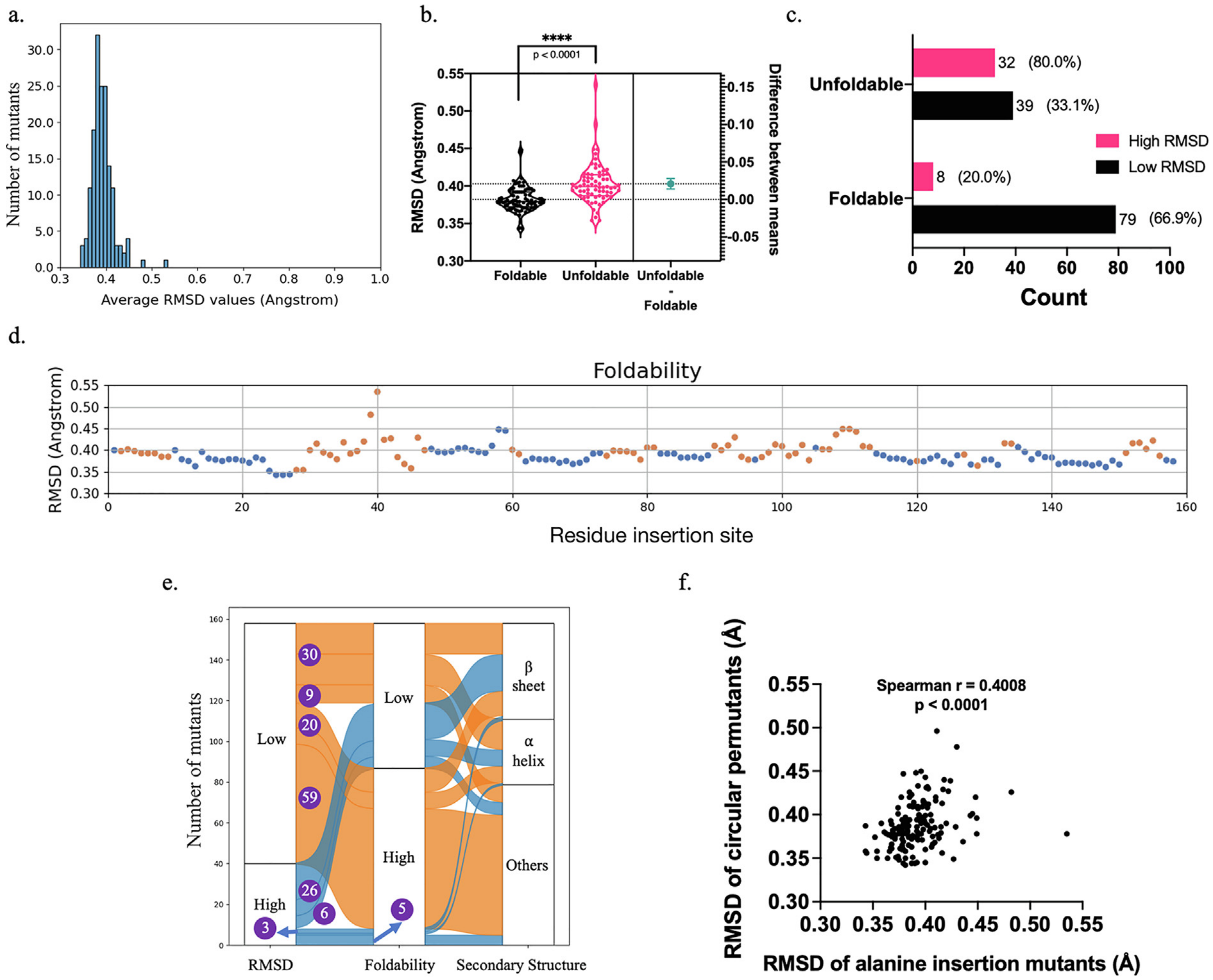

Fig. 4. The statistics of the AlphaFold2 models of 158 alanine insertion mutants of DHFRcoli. (a) The distribution of the RMSD values of the 158 AlphaFold2 models. (b) The comparison of the AlphaFold2 models' RMSD values of experimentally foldable and unfoldable mutants. The statistical p value was calculated with the unpaired *t*-test. (c) The amounts of the foldable and unfoldable mutants in the Low RMSD group and the High RMSD group. The numbers beside the bar are the mutant number and the percentage in the corresponding group. (d) Scatter plot of the RMSD values of the 158 DHFRcoli alanine insertion mutants. The blue dots indicate the experimentally foldable mutants, and the orange dots indicate the unfoldable ones. (e) The correlation between RMSD values of the models, the foldability of the alanine insertion mutants, and the secondary structure units of the residues. The numbers in purple circles are the numbers of the alanine insertion mutants in different categories. (f) The correlation between the RMSD values of the AlphaFold2 models of the circular mutants and the alanine insertion mutants. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ure S3a). On the contrary, even only when foldable alanine mutants were included, a positive correlation between their RMSD values and the precipitant ratios was noticed for AlphaFold2 models (Figure S3b). These data indicated that the RMSD values of the predicted models are correlated with proteins' native foldability but not their in vitro activities. Nonetheless, this would need further investigations since circular permutation and alanine insertion are comparable but not exactly equivalent [14].

3.6. Model difference between AlphaFold2 and RoseTTAFold partially predicts foldability

In these two DFHRcoli examples, the wild-type structures were reasonable references for calculating the RMSD values of the computational models. When it comes to the evaluation of the foldability of a protein without a wild-type reference (de novo protein design, for example), is it possible to get any indicators of foldability from the computational models of AlphaFold2 and RoseTTAFold? A possible indicator is the RMSD value between the AlphaFold2 model and the RoseTTAFold model (defined as inter-model RMSD hereafter), since we supposed that foldable proteins might have more evolutionary information for better and more consistent models. To answer this question, we calculated the inter-model RMSD values of the alanine insertion mutants of DHFRcoli (Table S2). Then we noticed that there was indeed a positive correlation between foldability and the inter-model RMSD values (Fig. 7a & Table S2). Without a wild-type structure as the reference, we grouped the models according to the average inter-model RMSD value. In the group with inter-model RMSD values lower than the average, 63.0 % alanine inserts are foldable, which is much higher than 43.9 %, the percentage of the foldable mutants in the group with higher RMSD values (Fig. 7b). This result also holds true when the median inter-model RMSD was used as the reference (Fig. 7b).

Then we asked if the inter-model RMSD value between AlphaFold2 model and RoseTTAFold model could be useful in de novo protein design. We obtained the sequence of the 129 de novo designs from the recent work of the Baker group, among which 27 were assigned as correctly folded [6]. We noticed that, using either the average RMSD or the median RMSD as the reference, the percentage of correctly folded designs in the lower inter-model RMSD group was much higher (Fig. 7c). Therefore, the

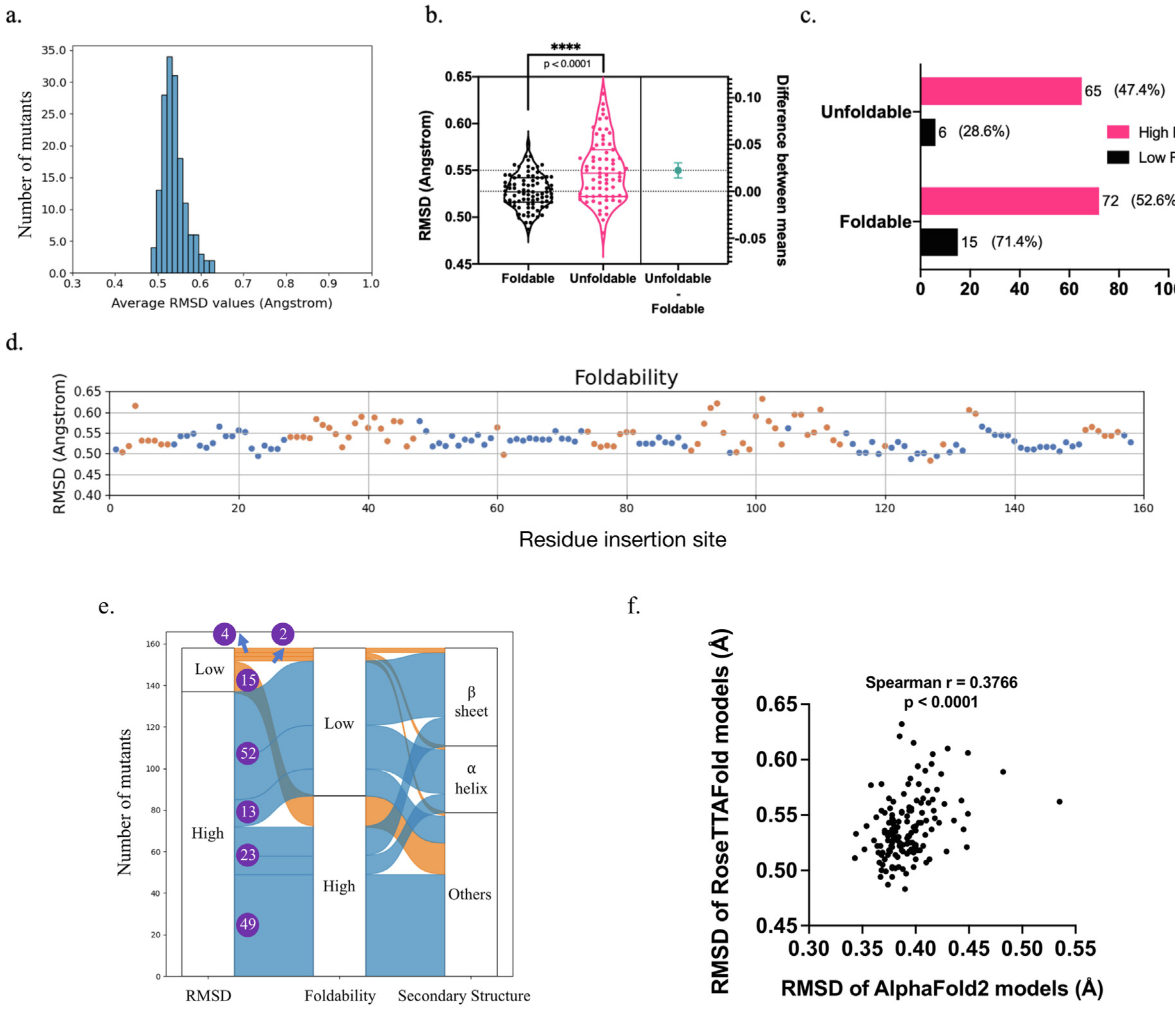

Fig. 5. The statistics of the RoseTTAFold models of 158 alanine insertion mutants of DHFRcoli. (a) The distribution of the RMSD values of the 158 RoseTTAFold models. (b) The comparison of the RoseTTAFold models' RMSD values of experimentally foldable and unfoldable mutants. The statistical p value was calculated with the unpaired *t*-test. (c) The amounts of the foldable and unfoldable mutants in the Low RMSD group and the High RMSD group. The numbers beside the bar are the mutant number and the percentage in the corresponding group. (d) Scatter plot of the RMSD values of the 158 DHFRcoli alanine insertion mutants. The blue dots indicate the experimentally foldable mutants, and the orange dots indicate the unfoldable ones. (e) The correlation between RMSD values of the models, the foldability of the mutants, and the secondary structure units of the residues. The numbers in purple circles are the numbers of the mutants in different categories. (f) The correlation between RMSD values of the AlphaFold2 models and the RoseTTAFold models of the DHFRcoli alanine insertion mutants. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

inter-model RMSD value between AlphaFold2 model and RoseTTAFold could be a useful parameter for evaluating the protein foldability in de novo design.

Since unfoldable proteins tend to precipitate, we supposed that the inter-model RMSD value could also be used to evaluate the solubility of proteins. So we tested this idea on a dataset containing 1664 proteins of E. coli reported by Niwa et al. [16]. As reported by Niwa et al., the solubility of these proteins is bimodally distributed (Fig. 7d & Table S3). Similarly, we noticed a correlation between the solubility and the inter-model RMSD values of these E. coli

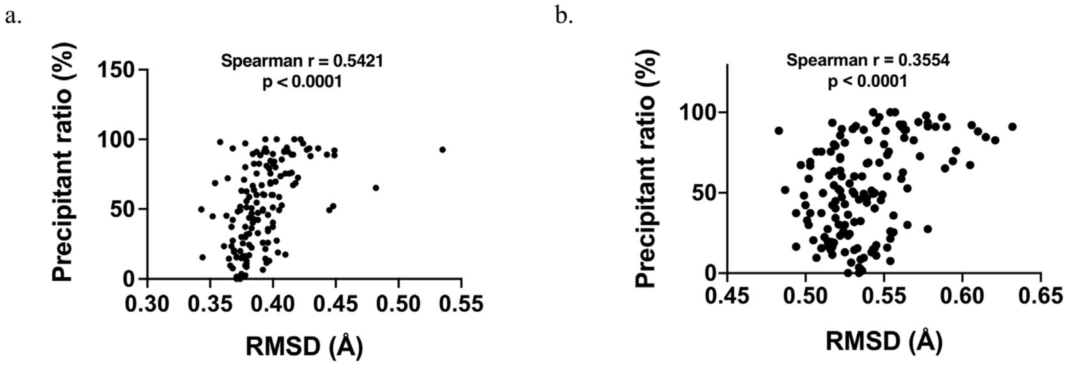

Fig. 6. The quantitative correlation between the RMSD values of computational models and protein foldability. (a) Plot for the AlphaFold2 models of the DHFRcoli alanine insertion mutants. (b) Plot for the RoseTTAFold models of the DHFRcoli alanine insertion mutants. The precipitant ratios were obtained from [14].

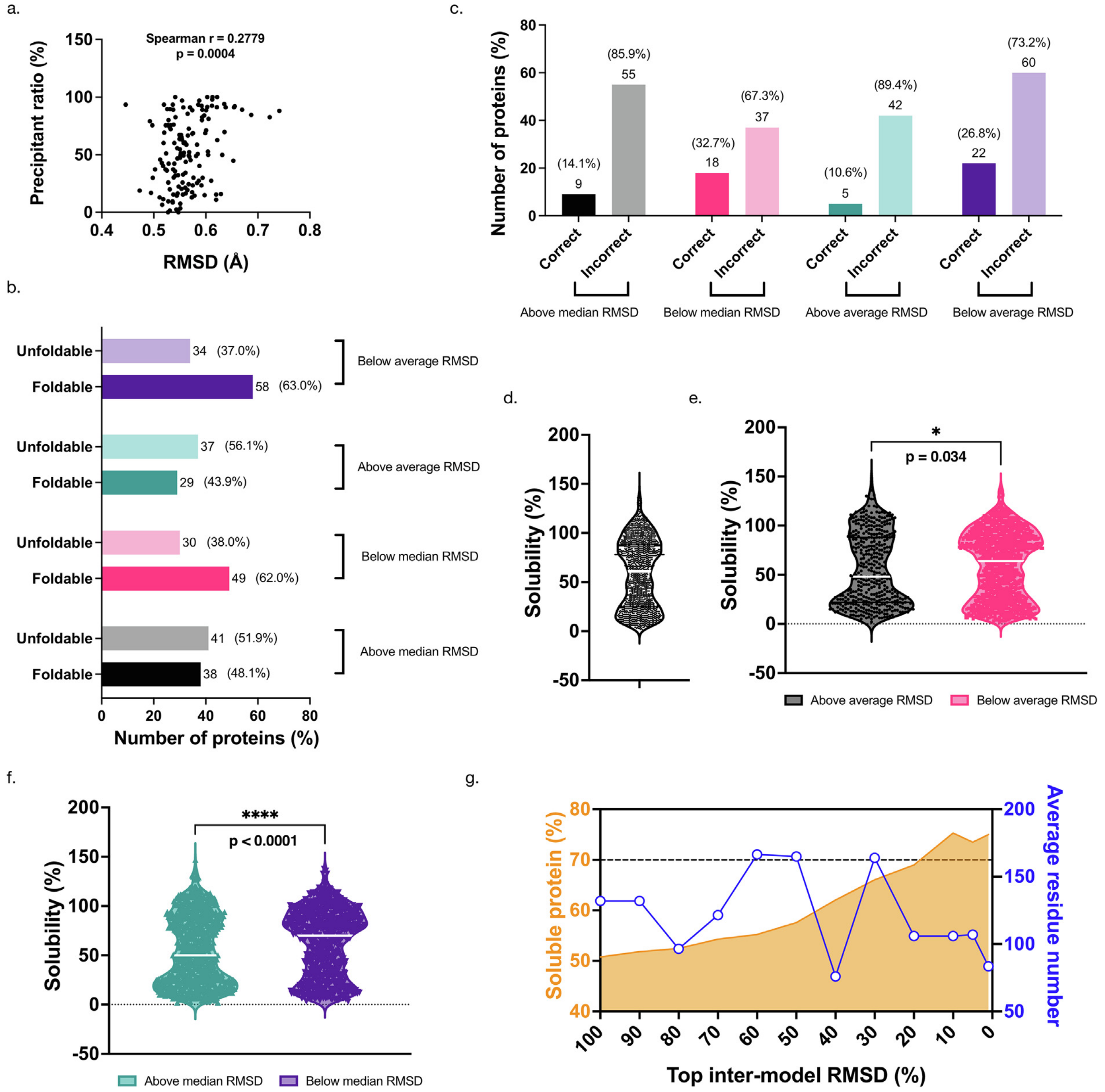

Fig. 7. Model difference between AlphaFold2 and RoseTTAFold partially predicts protein foldability. (a) There is a positive correlation between the precipitation ratios of the DHFRcoli alanine insertion mutants and their model RMSD values between AlphaFold2 and RoseTTAFold (inter-model RMSD). (b) The DHFRcoli alanine insertion mutants with lower inter-model RMSD values have better foldability. The numbers of the models are shown beside to the bars, with the percentages shown in the parentheses. (c) The classification of the correctly folded proteins of the de novo designs by the Baker group [6] and the inter-model RMSD values. More correctly folded designs have lower inter-model RMSD values. The numbers of the models are shown above the bars, with the percentages shown in the parentheses. (d) The distribution of the experimental solubility of the 1664 E. coli proteins reported in [16]. (e) The E. coli proteins with lower inter-model RMSD values have higher solubility when using the average inter-model RMSD value as the cutoff. The median of each category is shown as the white line, while the first and third quartiles are shown in colored lines. (f) The E. coli proteins with lower inter-model RMSD values have higher solubility when using the median inter-model RMSD value as the cutoff. The median of each category is shown as the white line, while the first and third quartiles are shown in colored lines. (g) The percentages and the average residue numbers of soluble proteins when different inter-model RMSD cutoff values were defined for the 1664 E. coli proteins.

proteins when the average inter-model RMSD value was used as the cutoff (Fig. 7e). The proteins with lower inter-model RMSD values had a solubility distribution biased toward the high solubility end, whereas the proteins with higher inter-model RMSD values biased toward the low solubility end (Fig. 7e). Similar but even more significant trends were noticed when the median RMSD value was used as the cutoff (Fig. 7f). When different inter-model RMSD cutoff values were set, the percentage of protein with high solubility (solubility ≥60 %) steadily increased from 50 % to 75 % (Fig. 7g), but this trend was not dependent on the size of the proteins.

4. Discussion

The protein folding problem is a "holy grail" problem in biology. To predict the natural 3D structure from a sequence of amino acids has been a tantalizing but extremely challenging task in the last 50 years. With the assembly of the biannual Critical Assessment of Structure (CASP) meetings, global efforts have pushed forward the progress in the computational prediction of protein structure [17]. At the recent CASP14 meeting, artificial intelligence (AI) based AlphaFold2 [4] and RoseTTAFold [5] achieved high-accuracy prediction of protein structures comparable to experimental data. However, in a recent example, only a small fraction of computationally designed proteins could be expressed as well-folded proteins [6].

A possible reason of this caveat of the current AI-based protein structure prediction tools is that from the training of PDB data, only the protein folding code could be extracted. That is, neither AlphaFold2 nor RoseTTAFold could directly obtain the kinetics information of protein folding. Meanwhile, the folding of a protein is not only determined by whether it could be theoretically folded, but also affected by its folding kinetics [2]. If a peptide chain could not be folded in a suitable time scale, its folding would be trapped by frustration and fails to form folded structures. Therefore, it is not surprising that the current AI-based prediction tools could not determine the biological foldability of a protein (Figs. 1 & 4).

Nonetheless, considering that the structures in the PDB database are intrinsically foldable, it is possible that the network properties of AI-based methods contain some information of protein foldability. For example, the residue-residue contact information from multiple sequence alignment might contain foldability information, since unfoldable mutants have been discarded by evolution. Then a reasonable inference is that there is some foldability information hidden in the computationally folded models. The question is how that information could be extracted.

Circular permutation modifies the termini of a protein but does not change its sequence composition. Most circular permutants could fold into native structures, although their folding kinetics might change. The alanine insertion method is more conserved on perturbing protein sequence. Since the protein sequence is minimally changed, the computational tools would be hard to tell the difference on the foldability of different constructs. Our data showed that this is true, since AlphaFold2 and RoseTTAFold predicted near-native 3D models for both foldable and unfoldable DHFRcoli permutants and alanine-insertion mutants. However, it was interesting to notice that the RMSD values of the computational models contains some information on the foldability of a protein in our work. We think our work, along with the contributions from the other colleagues [18,19], will provide great help to the prediction of protein foldability.

A caveat of our work is that the experimental data only reflected the foldability of a peptide sequence kinetically accessible within a laboratory time scale. For example, Iwakura et al. [11] refolded some permutants from inclusion bodies in vitro. But this should not be a big problem, since a protein not foldable within a reasonable time scale might be not biologically useful.

Our work would be helpful to the design of protein circular permutants and novel proteins. Based on our findings, we hypothesize that including non-foldable protein sequences in the training data of neural networks would be useful for the AI-based prediction methods to predict protein foldability. In this regard, we believe that publishing unsuccessful protein design data is scientifically valuable [20]. Lastly, we hope that our work would be a hint to the establishment of AI-based prediction of protein folding kinetics in the future.

Author contributions

S.L. conceived the idea and did the computational work. K.W. and C.C. acquired the experimental data of the mutants. S.L. and K.W. analyzed and interpreted the data. S.L. wrote the manuscript. All authors reviewed and approved the submitted manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the support from the other members of our lab and Prof. Yongqi Huang for kind suggestions. S. L. was supported by the grants from National Natural Science Foundation of China (31971150), Department of Science and Technology, Hubei Provincial People's Government (2019CFA069), and Hubei University of Technology.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.08.034.

References

[1] Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–30.
[2] Dill KA, MacCallum JL. The protein-folding problem, 50 years on. Science 2012;338:1042–6.
[3] Dill KA, Ozkan SB, Shell MS, et al. The protein folding problem. Annu Rev Biophys 2008;37:289–316.
[4] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
[5] Baek M, Dimaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021;373:871–6.
[6] Anishchenko I, Pellock SJ, Chidyausiku TM, et al. De novo protein design by deep network hallucination. Nature 2021;600:547–52.
[7] Nobrega RP, Arora K, Kathuria SV, et al. Modulation of frustration in folding by sequence permutation. Proc Natl Acad Sci USA 2014;111:10562–7.
[8] Bliven S, Prlić A. Circular permutation in proteins. PLoS Comput Biol 2012;8:e1002445.
[9] Smith VF, Matthews CR. Testing the role of chain connectivity on the stability and structure of dihydrofolate reductase from E. coli: fragment complementation and circular permutation reveal stable, alternatively folded forms. Protein Sci 2001;10:116–28.
[10] Iwakura M, Nakamura T. Effects of the length of a glycine linker connecting the N-and C-termini of a circularly permuted dihydrofolate reductase. Protein Eng 1998;11:707–13.
[11] Iwakura M, Nakamura T, Yamane C, et al. Systematic circular permutation of an entire protein reveals essential folding elements. Nat Struct Biol 2000;7:580–5.
[12] Arai M, Maki K, Takahashi H, et al. Testing the relationship between foldability and the early folding events of dihydrofolate reductase from Escherichia coli. J Mol Biol 2003;328:273–88.
[13] He L, Tan P, Zhu L, et al. Circularly permuted LOV2 as a modular photoswitch for optogenetic engineering. Nat Chem Biol 2021:1–30.
[14] Shiba R, Umeyama M, Tsukasa S, et al. Systematic alanine insertion reveals the essential regions that encode structure formation and activity of dihydrofolate reductase. Biophysics (Nagoya-shi) 2011;7:1–10.
[15] Galzitskaya OV, Ivankov DN, Finkelstein AV. Folding nuclei in proteins. FEBS Lett 2001;489:113–8.
[16] Niwa T, Ying B-W, Saito K, et al. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. Proc Natl Acad Sci USA 2009;106:4201–6.
[17] Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15:285–9.
[18] Mezei M. On predicting foldability of a protein from its sequence. Proteins 2020;88:355–65.
[19] Kaushik R, Zhang KYJ. A protein sequence fitness function for identifying natural and nonnatural proteins. Proteins 2020;88:1271–84.
[20] Joober R, Schmitz N, Annable L, et al. Publication bias: what are the challenges and can they be overcome? J Psychiatry Neurosci 2012;37:149–52.
[21] Nakamura T, Iwakura M. Circular permutation analysis as a method for distinction of functional elements in the M20 loop of Escherichia coli dihydrofolate reductase. J Biol Chem 1999;274:19041–7.
[22] Mirdita M, Schütze K, Moriwaki Y, et al. ColabFold-Making protein folding accessible to all. 2021.
[23] Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.4. 2010.