



# OPEN Conformational ensembles for protein structure prediction

Jiaan Yang<sup>1,2,11</sup>✉, Wen Xiang Cheng<sup>1,11</sup>, Peng Zhang<sup>1,10,11</sup>, Gang Wu<sup>3</sup>, Si Tong Sheng<sup>4</sup>, Junjie Yang<sup>5</sup>, Suwen Zhao<sup>6</sup>, Qiyue Hu<sup>7</sup>, Wenxin Ji<sup>8</sup> & Qiong Shi<sup>9,11</sup>

Acquisition of conformational ensembles for a protein is a challenging task, which is actually involving to the solution for protein folding problem and the study of intrinsically disordered protein. Despite AlphaFold with artificial intelligence acquired unprecedented accuracy to predict structures, its result is limited to a single state of conformation and it cannot provide multiple conformations to display protein intrinsic disorder. To overcome the barrier, a FiveFold approach was developed with a single sequence method. It applied the protein folding shape code (PFSC) uniformly to expose local folds of five amino acid residues, formed the protein folding variation matrix (PFVM) to reveal local folding variations along sequence, obtained a massive number of folding conformations in PFSC strings, and then an ensemble of multiple conformational protein structures is constructed. The P53\_HUMAN as a well-known protein and LEF1\_HUMAN and Q8GT36\_SPIOL as typical disordered proteins are taken as the benchmark to evaluate the predicted outcomes. The results demonstrated an effective algorithm and biological meaningful process well to predict protein multiple conformation structures.

**Keywords** Protein structure prediction, Protein conformation, Protein folding, Intrinsically disordered protein, Protein intrinsic disorder, AlphaFold

A native protein has flexible conformations because of lower free-energy barriers between different folding states. Although AlphaFold and other AI machine learning-based methods can predict protein structures in single status with accuracy against 3D structural data from experimental measurements, it has the limitation to obtain multiple conformation structures<sup>1,2</sup>. A native protein structure should be fully represented by an ensemble of multiple conformations in vary degrees<sup>3,4</sup>. Essentially, to overcome the limitation is into the challenge involving the solution for protein folding problem. A protein may take various pathways folding into a stable state when it emerges from a ribosome, also it may repeatedly unfold and refold during its lifetime, or take an altered conformation under different conditions<sup>5</sup>. According to the Levinthal paradox, a protein can be folded into an astronomical number of structural conformations<sup>6</sup>.

Generally, a protein itself may well curve into a folded state, and also its conformation may be changed under different environments and interactions. Several forces drive a protein itself to fold into 3D structure<sup>7,8</sup>. First, the hydrogen bond is the primary force to form secondary fragment, such as  $\alpha$ -helix and  $\beta$ -strand in protein structure. Second, the hydrophobic interactions make the residues tightly packed in protein, which particularly lead hydrophobic residues predominantly in the core while the polar residues usually trend on the surface. Third, the electrostatic interactions between amino acids make attracting or repelling residue contact. Fourth, the constraints of preferred dihedral angles of neighboring backbone bond influence the protein folding with space orientation. However, a protein structure is not static, the folded proteins are in constant motion, which is crucial for biological activity. In vitro and cell studies, the free-energy barriers,  $\Delta G^\ddagger$ , for folding conformation change is quite small, which is only on the order of  $\sim 5$  kcal/mol<sup>3,4,9</sup>. So, it is not supervised that a protein has folding flexibility, and its conformation may be changed in various physiological conditions, such as solvent, temperature, ligand and protein interaction, etc.<sup>10</sup>. The protein always folds into a structure that is thermodynamically stable under physiological conditions<sup>11</sup>. It is apparently that protein conformation has

<sup>1</sup>Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, Guangdong, China. <sup>2</sup>Micro Biotech, Ltd., Shanghai 200123, China. <sup>3</sup>School of Basic Medicine, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, Hubei, China. <sup>4</sup>HYK High-Throughput Biotechnology Institute, Shenzhen 518057, Guangdong, China. <sup>5</sup>Wuhan International Biohub Cooperation, Wuhan 430075, Hubei, China. <sup>6</sup>iHuman Institute, ShanghaiTech University, Shanghai 201210, China. <sup>7</sup>Beyang Therapeutics Co. Ltd, Shanghai 201210, China. <sup>8</sup>National Facility for Protein Science in Shanghai, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China. <sup>9</sup>Laboratory of Aquatic Genomics, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen 518057, China. <sup>10</sup>Biomedical Engineering, Shenzhen University of Advanced Technology, Shenzhen 518060, China. <sup>11</sup>Jiaan Yang, Wen Xiang Cheng, Peng Zhang, and Qiong Shi are co-authors. ✉email: jiaanyang@yahoo.com

flexibility which is able to fluctuate between different structural states by itself or under different conditions. From perspective view of intrinsically disordered proteins (IDP) or intrinsically disordered range (IDR), many proteins fail to form a stable structure for entire protein or partial ranges in structure, which may have impact on biological functions or diseases<sup>12,13</sup>.

The acquisition of protein structure from its amino acid sequence is a challenge in structural biology for longstanding. The protein structures are investigated by both experimental measurements and computational predictions. The experimental techniques required the protein structure state to exist for a long enough time scale for observing and detection while the computational dynamics simulations are generally limited to modelling events with very short timescales on the order of about a microsecond<sup>14</sup>. X-ray crystallography and Cryo-electron microscopy (Cryo-EM) are efficient methods to determine some of protein folding 3D structure<sup>15</sup>. Protein nuclear magnetic resonance (NMR) is the primary method to collect multiple folding states for relative small proteins in solvent<sup>16</sup>. So far, over 225,000 structures are available in the Protein Data Bank (PDB)<sup>17</sup>, which provided rich information for protein folding study. Also, other methods are also able to provide the information for protein folding, such as fluorescence spectroscopy<sup>18</sup>, circular dichroism spectroscopy<sup>18</sup> and single-molecule force spectroscopy<sup>19</sup> etc. However, the progress of experimental measurements for protein 3D structures cannot keep up with the pace of rapid increase number of protein sequences determined by genetic code. So, the computational approaches to predict protein structure according to amino acid sequence are important. Since 1994, the Critical Assessment of Techniques for Protein Structure Prediction (CASP) has organized and promoted the community for protein structure prediction, and the quality of methods has been assessed<sup>20,21</sup>. Over hundreds of computational prediction methods are developed, which are categorized into two categories. One is template-based modelling, which is based on homologous proteins to predict similar structures<sup>22</sup>. Another is template-free modelling, which is based on force fields to obtain protein structure with a minimum in Gibbs free energy<sup>23–25</sup>. Some effective computational methods have developed, such as I-TASSER (Iterative Threading ASSEMBLY Refinement)<sup>26</sup> with the detection of structure templates from the PDB, and HHpred<sup>27</sup> with sequence searches for protein homology information. Recently, with deep learning in artificial intelligence (AI), AlphaFold approach, including AlphaFold 2 and AlphaFold 3, successfully predicted protein structures with higher accuracy than any other methods<sup>20,21,28</sup>. AlphaFold first handled the residues as nodes and the connection of residues for protein structure. Then, with neural network system, it trained residue-to-residue and atom-atom using an internal confidence measure based on protein 3D structures from PDB for unknown structures. The protein structure was refined by evolutionarily related multiple sequence alignment (MSA). With iterating process, AlphaFold obtained protein structure with higher accuracy, and its database contained over 200 million of predicted protein 3D structures. Although AlphaFold well predicted protein structure in single state<sup>29</sup>, it has not delivered a comprehensive solution for the protein folding problem<sup>30</sup>. In other words, despite the advances in protein structure prediction recently, the goal to acquire complete ensemble of protein folding conformations has not been approached yet.

The shortcomings and limitations of AlphaFold approach have been revealed in several aspects<sup>31–34</sup>. First, the predicted result from AlphaFold only represents a single state of protein structure. So the result cannot fully embody a native protein structure with conformational heterogeneity. Second, although the deep learning neural network technology helped to solve the complex protein prediction problem, but it cannot directly provide understanding how the protein is folded in biophysical principles. Third, AlphaFold is not single sequence method, which heavily depends on multiple sequence alignment (MSA) to predict protein structure. So, its predicted result is lower confidence if the number of homologous sequences in database is less. Also, it is not able to specify the subtle changes in sequence, such as residue mutation or species differentiation how to impact the change of folding conformation. Although some models have been proposed, such as RoseTTAFold, ESMFold, OpenFold and EigenFold etc.<sup>34</sup>, the correlation with protein folding flexibility still does not have the solution<sup>34</sup>.

Here, a new approach, FiveFold, has been developed trying to answer the protein folding problem and to provide comprehensive complete conformations for protein 3D structure prediction. The FiveFold approach is a single sequence method which contains four modules. First, a set of 27 of protein folding shape code (PFSC) in alphabet letters is established to cover complete folding space for five amino acid residues, which greatly simplify the complex protein folding object<sup>35</sup>. For any given protein 3D structure, its conformation can be expressed by a PFSC string and a PDB-PFSC database is created. Second, a database is generated which contained contains all possible folding patterns in PFSC letter for any combination of five amino acid residues. Third, the protein folding variation matrix (PFVM) assembled all possible local folding variants in each column with PFSC letter along sequence in matrix, which directly displays the fluctuation of folding conformations for the entire protein<sup>36</sup>. Also, it well revealed how the protein folding features were related to the order of amino acids in sequence. With optimization of various combinations of PFSC letters, a massive number of PFSC strings can be produced from PFVM, which represent a massive number of conformations. Fourth, an ensemble of multiple conformation 3D structures are constructed by high-throughput homogenous conformation search screening PDB-PFSC database with using these PFSC strings for protein structure prediction. The FiveFold approach has at least three significant features for protein structure prediction. First, the information for complete folding conformations is held in PFVM. An astronomical number of conformations for each protein is able to be explicitly obtained from its PFVM, which provides an unambiguous answer to Levinthal paradox. Of course, a set of most probable conformations in a limited number can be acquired, and then an ensemble of protein 3D structures is able to be predicted. Second, the FiveFold approach is a single sequence method, which a PFVM depended on its amino acid sequence, i.e. different sequence has different folding variations and features. Thus, any residue mutation in sequence or same protein for differentiation species has different PFVM, and then has different conformational structures. Third, the FiveFold provided a biophysically understanding algorithm to predict an ensemble of multiple conformation protein structures. Therefore, the FiveFold offered a meaningful approach which can overcome the shortcomings and limitations of current computational protein structure prediction.

## Materials and methods

The FiveFold approach is based on all possible folding patterns for free connected five points as well as five amino acid residues in protein, and then the conformational ensembles for protein structures are predicted. The approach is composed of four modules, such as PFSC, 5AAPFSC database, PFVM and construction of ensemble of protein structures. Also, the software, web server and supporting database are already established.

### Four modules

#### PFSC

Any protein conformation is able to be completely described by a PFSC (protein folding shape code) alphabetic string. Mathematically, starting from a model of five points with successive connection without any constrain, a set of folding shapes for five of amino acid residues is obtained completely to cover folding space, which are represented by 27 letters including “\$” as PFSC. The PFSC well character protein folding features, including alpha-helix, beta-strand, irregular folds and mixture in various degrees. For any protein with given 3D structure, its PFSC string is able fully to describe the folding conformation without gap along sequence, covering secondary structure fragments as well as tertiary structure fragments. The conversion procedure from a given protein 3D structure to PFSC is described by blue arrows in Fig. 1. Thus, a PFSC string, as the fingerprint of protein folding, corresponds to one protein conformation<sup>35</sup>. The conformations for all given structures in PDB were converted into PFSC string and stored into PDB-PFSC database.

#### 5AAPFSC

5AAPFSC database assembled all possible folding patterns for each five amino acid fragment in PFSC letters. Based on 20 amino acids, 3,200,000 permutations of five amino acids exist. All folding shapes for each set of five amino acids are collected from protein database or computed by MD (molecular dynamics) simulation. About two-thirds of permutations of five amino acids are available in PDB, so their folding patterns are able to be collected. The remnant one-third of permutations of five amino acids were computed by MD simulation with CHARMM (Chemistry at Harvard Macromolecular Mechanics)<sup>37</sup>. Then, all folding patterns for five amino acids are converted into the PFSC letters and stored in 5AAPFSC database. The procedure for database setup is described by green arrows in Fig. 1.

#### PFVM

The local folding variations for any protein can be fully presented by its protein folding variation matrix (PFVM), and the multiple conformations for entire proteins can be produced. Any set of five amino acid residues has different folding variants and different number of folding patterns, and these folding features are represented by PFSC letters which are extracted from 5AAPFSC database. For a protein, five amino acid residues are continuously taken along sequence and each step moves forward only one residue beginning from N-terminus. Extracting all possible folding shapes from 5AAPFSC database, the PFSC letters for each of five amino acids are ranked in a column according to favorability, and then PFVM is formed along sequence from N-terminus to C-terminus. The procedure forming PFVM is described by first two of red arrows in Fig. 1. To take a PFSC letter from each column in PFVM, an astronomical number of folding conformations in PFSC strings for a protein can be explicitly obtained without ambiguity. However, in practice a limit number of most possible folding conformations are interested to be obtained. The order of PFSC letters in each column are ranked from higher possibility to lower. So, the most possible local folds appeared on top in matrix. One of the most possible conformations is the PFSC string at the first row in PFVM that is named as PFVM-01, and more possible conformations can be obtained with PFVM-01 by replacing PFSC letter in the same column. In order to obtain a limited number of optimum conformations, the PFSC letters at top rows in PFVM are considered for replacement in PFVM-01. Thus, a massive number of protein folding conformations in PFSC strings can be formed by the local folding variations in PFVM, which is represented by third red arrow in Fig. 1. It is apparently that any protein sequence has its proprietary PFVM, which provided its special local folding variations to construct a massive of PFSC strings for protein conformations<sup>36</sup>.

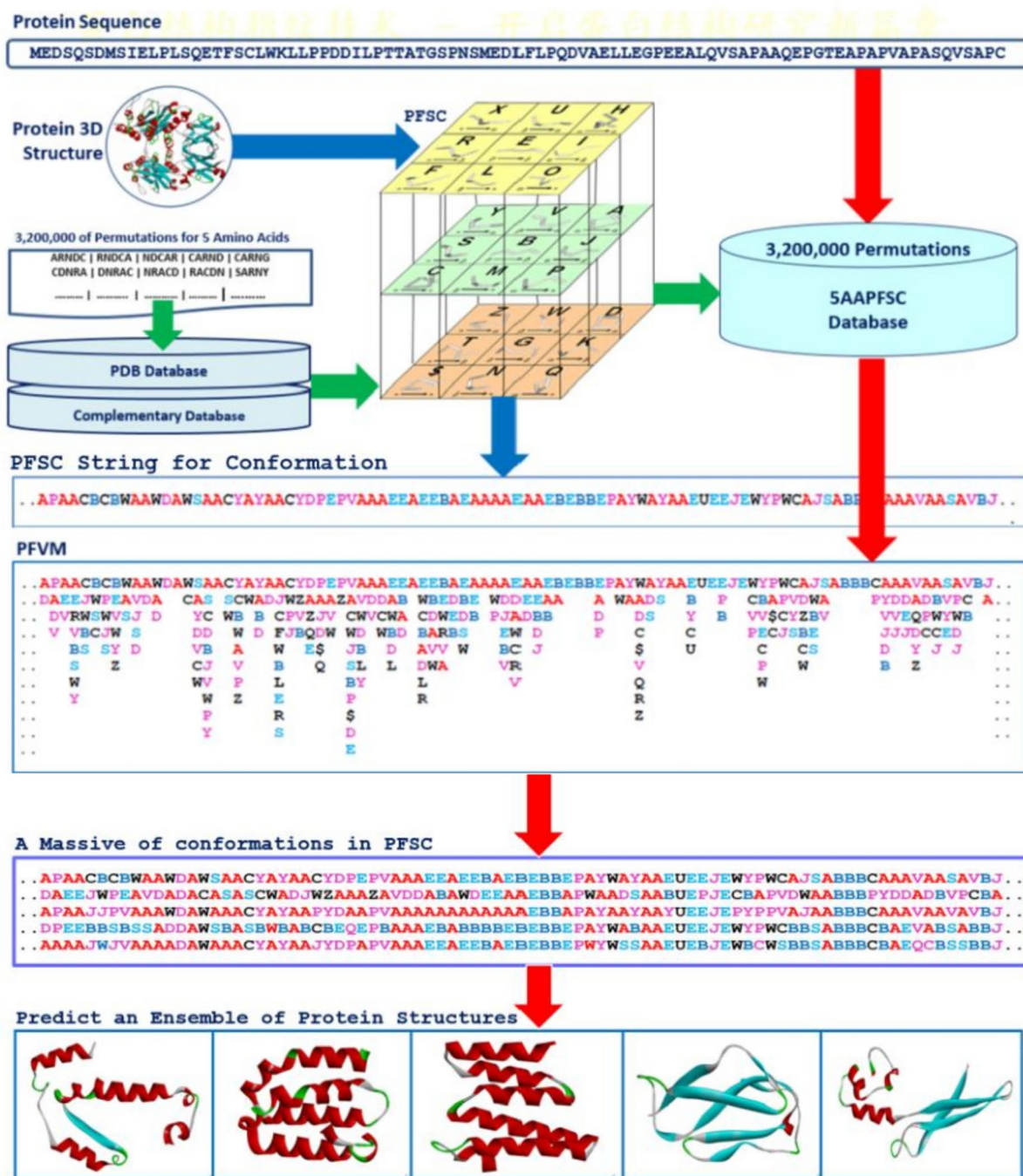
#### Construction of ensemble of protein 3D structures

It is a critical step how to construct an ensembles of protein 3D structures from PFSC string. According to folding patterns from each PFSC string, its similar folding 3D protein structure can be found by high-throughput screening PDB-PFSC database with conformational homologous search. Because of digital description, the process of high-throughput screening with homologous conformation search is very effectively. Generally, a similar given 3D protein structure as the searched result is accepted if the normalized score of Protein Folding Structural Alignment (PFSA)<sup>38</sup> is larger than 0.70. Due to larger protein is hard to meet the criteria, a protein is usually divided into multiple fragments with about 40–50 residues to satisfy the criteria, and then fragments are connected back for the whole protein. After that, the protein structure may be optimized by molecule dynamics simulation. Thus, the protein 3D structure is able to be constructed according each PFSC string, and the ensembles of protein 3D structures for multiple conformations are predicted according to multiple PFSC strings formed from its PFVM, which is presented by fourth red arrow in Fig. 1.

### Validation

The quantitative accuracy of FiveFold predicted protein structures is validated by comparison with given experimental structures. The protein structures are converted into the conformations with PFSC strings, and then the PFSC strings are aligned for comparison. A set of multiple conformational structures can be formed from its PFVM, which may cover wider range of flexible conformational space than given structures. If a protein





**Fig. 1.** FiveFold approach to predict the conformational ensembles of protein structures with PFSC and PFVM algorithms. The cubic contains a set of 27 PFSC as folding patterns. The blue arrows indicated the process how to obtain the PFSC string from a protein with known 3D structure. The green arrows indicated the process to construct 5AAMPFSC database, which contained all folding shapes in PFSC letters for 3,200,000 of permutations of five amino acids. The red arrows indicated the process to predict an ensemble of protein structures from a protein sequence with the PFVM. The PFSC letters with red and pink colors are for typical helix and alike-helix local folds; the blue and light blue colors for beta strand and alike-beta strand and block color for irregular folds<sup>36</sup>.

has given structure, the predicted structures should have at least one conformation well to align with the given structure with 1.00 for the score of Protein Folding Structural Alignment (PFSA).

#### Computational efficiency and feasibility

The computational overhead is low and the performance is efficient because the most routines in FiveFold are carried out based on digital alphabetic description for protein folding conformations. The software code was



developed by Java (jdk1.8.0) and the database is stored in XML format. The task to predict conformational ensembles of protein structures was carried out on PC (i7-1355U, 1.70 GHz, 64-bit operating system) with Windows 11 platform. For example, the prediction of 10 conformational structures for a protein with 1000 amino acids in sequence takes a short period of time. Generally, from a sequence to obtain its PFVM takes about 5 s; from PFVM to form 10 PFSC strings takes about 10 s; the high-throughput homologous conformation search screening for 10 PFSC strings takes about 20 min; to construct 10 of protein 3D structures takes about 5 min. So, it takes less than 30 min to generate 10 conformational structures from its protein sequence.

Results

The conformational ensembles of protein 3D structures are predicted by FiveFold approach with using PFVM and PFSC algorithm. Based on a sequence its specific PFVM is obtained, and then a massive number of PFSC strings is formed by combination of PFSC letters from PFVM, which represents the multiple folding conformations. According to these PFSC strings, an ensemble of 3D structures is constructed to characterize the most possible multiple conformations for a native protein. P53\_HUMAN (P04637) is a structural well-known multiple domain protein with over two hundreds of given 3D structural data in PDB, and LEF1\_HUMAN (Q9UJU2) and Q8GT36\_SPIOL (Q8GT36) are two typical intrinsically disordered proteins. These proteins are good benchmarks to evaluate the predicted results of an ensemble of multiple conformation structures. Here, with FiveFold, an ensemble of multiple conformation structures are predicted for each protein, which are compared with given protein structures as well as AlphaFold predicted structure.

P53\_HUMAN


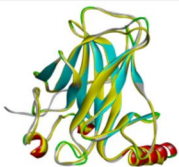
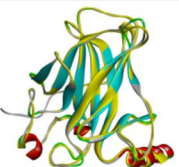
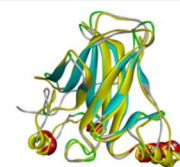

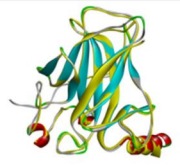


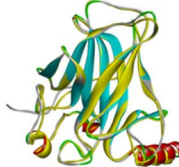
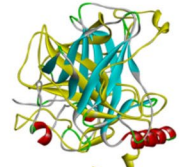
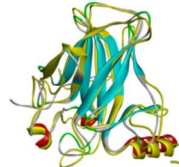
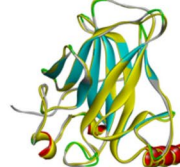
The P53\_HUMAN (P04637) is composed of 393 amino acid residues which is a regulatory protein and prevents cancer formation. The P53\_HUMAN is a protein with four domains. As a well-studied protein, the P53\_HUMAN has about 260 of given 3D structural data in PDB, including over 140 for DNA-binding domain (100–288) and 3 for the entire protein. These structures provided rich folding information to evaluate the predicted results. Generally, the protein structure of a domain may swing around certain conformation while the fragments between domains and both terminuses have more folding flexibility. With FiveFold, the structure prediction will first focus on the DNA-binding domain, and then on the entire P53\_HUMAN protein.

Domain structure prediction

As P53\_HUMAN has many given 3D structures, it is important to understand the conformations of given 3D structures before protein structure prediction. Although a domain in protein may have a stable structure, the factors of environment and measurement methods indeed affect the folding conformation for a domain. The information about conditions, environment and measurement methods for 12 of P53 DNA-binding domain structures plus AlphaFold prediction are listed in Table 1. These protein structures were measured by different means crossing time span from 1995 to 2022. It is apparently that some of protein structures interact with ions, some with protein or DNA. So, it is not surprising that P53 DNA-binding domain may have different folding conformations in some degrees. Each structure is separately paired with 2PCX for comparison, and the 3D structural superposition and the overlay similarity scores are displayed in Fig. 2. Although the values of scores exposed which pair may have higher or lower similarity, but it is hard visually to distinguish or illustrate the difference between each pair of conformations. However, the PFSC is able well to reveal the conformation difference. Each conformation of given 3D structure is first expressed by a PFSC string. And each PFSC string is separately aligned with 2PCX in the C section of Table 2 for comparison. It obviously displayed the conformational similarity and dissimilarity, and the scores of protein folding structural alignment (PFSA) in italics are listed in Fig. 2. First, both similarity scores for most structures have parallel trend to evaluate the protein conformation similarity. Second, the PFSC alignment is better than the structure superposition to expose the local folding

| PDB ID    | Condition and environment   | Method                  | Year |
|-----------|---|-------------------------|------|
| 2PCX      | Zn <sup>+2</sup>  | X-Ray Diffraction       | 2007 |
| 1GZH      | P53-binding protein 1, SO <sub>4</sub> <sup>-2</sup> , Zn <sup>+2</sup>   | X-Ray Diffraction       | 2002 |
| 1TSR      | DNA, Zn <sup>+2</sup>   | X-Ray Diffraction       | 1995 |
| 1YCS      | Apoptosis-stimulating of p53 protein 2,   | X-Ray Diffraction       | 1997 |
| 2FEJ      | Above body temperature, Zn <sup>+2</sup>  | Solution NMR            | 2005 |
| 2MEJ      | Bcl-2-like protein 1, Zn <sup>+2</sup>  | Solution NMR            | 2013 |
| 2YBG      | N(6)-ACETYLLYSINE, Zn <sup>+2</sup>   | X-Ray Diffraction       | 2011 |
| 5BUA      | DNA, N(6)-ACETYLLYSINE, Zn <sup>+2</sup>  | X-Ray Diffraction       | 2015 |
| 5LGY      | DNA, , N(6)-ACETYLLYSINE, Zn <sup>+2</sup>  | X-Ray Diffraction       | 2016 |
| 2ACO      | DNA, Zn <sup>+2</sup>   | X-Ray Diffraction       | 2005 |
| AlphaFold |   | Prediction by AlphaFold | 2022 |
| 6XRE      | RPB1, RPB2, RPB3, RPB4, RPB5, RPB7, RPB9, RPB11-a, RPABC2, , RPABC3, BRAB4, RPABC5, Mg <sup>+2</sup> , Zn <sup>+2</sup> | Electron Microscopy     | 2020 |
| 8F2I      |   | Electron Microscopy     | 2022 |

**Table 1.** The different condition, environment and measurement method for 3D structures of DNA-binding domain regions (100–288) of P53\_HUMAN (P04637) protein. There are 13 protein structures timing span 1995–2022, which 8 were measured by X-Ray diffraction, 2 by solution NMR, 2 by electron microscopy in PDB and one from AlphaFold database. Also, the differences of condition and environment for each structure are listed.

|   |   |   |  |   |   |
|---|---|---|--|---|---|
|  |  |  |  |  |  |
| <b>1GZH-A</b><br>0.796<br><i>0.746</i>  | <b>1TSR-A</b><br>0.806<br><i>0.802</i>  | <b>1YCS-A</b><br>0.814<br><i>0.802</i>  | <b>2FEJ-1</b><br>0.645<br><i>0.769</i>   | <b>2MEJ-1</b><br>0.762<br><i>0.744</i>  | <b>2YBG-A</b><br>0.796<br><i>0.740</i>  |
|  |  |  |  |  |  |
| <b>5BUA-A</b><br>0.844<br><i>0.839</i>  | <b>5LGY-A</b><br>0.849<br><i>0.832</i>  | <b>2AC0-A</b><br>0.839<br><i>0.864</i>  | <b>6XRE-M</b><br>0.388<br><i>0.693</i>   | <b>8F2I-A</b><br>0.643<br><i>0.686</i>  | <b>AlphaFold</b><br>0.876<br><i>0.851</i>   |

**Fig. 2.** Pair comparisons between 2PCX-A structure and each given 3D structure of DNA-binding domain (100–288) of P53\_HUMAN (P04637). The image of superposition of each pair of structures is displayed, the colorful structure is for 2PCX-A and yellow for the compared structure. The PDB-ID and chain name of the compared structure is listed under image. The overlap similarity score is displayed under PDB-ID, which was obtained by overlaying the molecule structures with alignment by a combination of 50% steric component and 50% electrostatic component in software of Discovery Studio v24.1.0. Also, with PFSC string alignment, the score of protein folding structural alignment (PFSA) in italics is listed in bottom.

similarity and difference. The PFSC letters with red and pink colors are for typical helix and alike-helix local folds; the blue and light blue colors for beta strand and alike-beta strand. So, with quick scan on color of PFSC code, the secondary structure fragments are generally aligned while the difference of local folds are remarked by yellow color on background. Also, with using PFSC, the AlphaFold structure conformation (D section in Table 2) is able to be compared with given protein structures. Thus, the PFSC alignment better revealed the folding similarity and difference for given P53\_HUMAN structures.

An ensemble of protein multiple conformation 3D structures for P53 DNA-binding domain is able to be constructed by FiveFold approach. The PFVM for DNA-binding domain regions (100–288) of P53\_HUMAN (Section B of Table 2) is obtained according to its amino acid sequence in section A. The local folding variations of P53 are fluctuated in both folding patterns and numbers along sequence, which are explicitly displayed by PFSC at each column in PFVM. The first row in PFVM is a complete PFSC string, which is named as PFVM-01 representing one of most possible conformations. If any PFSC letter in PFVM-01 is replaced by a letter in same column, then total  $1.33 \times 10^{150}$  PFSC strings will be formed, i.e. astronomical number of folding conformations can be fashioned from PFVM. However, most biology researchers are interesting in a limit number of stable conformations. Based on PFVM-01, a set of most probable conformations with limit number is formed by the optimized process to replace the PFSC letters in same column in PFVM. For examples, with using PFSC letters in the second and third rows for replacement, a set of PFSC strings is formed. Also, with using PFSC vector feature, the optimized coupling PFSC string is able to be obtained. The section E of Table 2 listed 18 PFSC strings, which are some of possible multiple conformations of P53\_HUMAN. It is obvious that 12 PFSC strings for given structures at the sections of C, a PFSC string from AlphaFold at the section D and 18 PFSC strings from FiveFold prediction at the section E in Table 2 are able to be aligned for comparison, and the local folding similarity and differences are exposed. Subsequently, 18 protein 3D structures can be constructed according to each PFSC string from FiveFold prediction with high-throughput screening PDB-PFSC database to search structures or fragments with conformation homology. Finally, 18 of protein 3D structures is able to be predicted. The superposition between each of the predicted protein 3D structure and the chain A of given P53\_HUMAN protein structure 2PCX is displayed in Fig. 3, and the overlay similarity scores and the scores of PFSA in italics are listed. It is noticed that the values of overlay similarity scores in Fig. 3 for predicted structures are more dispersed than values in Fig. 2, which indicated the FiveFold predicted structures with wider flexibility than the given structures. Also, one conformation of the FiveFold predicted structures (PFVM-13) in Finger 3 is well matched with the given structure of 2PCX. It is not surprised that the PFVM contains the local folding variations which embrace all local folding patterns for given structures<sup>36</sup>. At the section C of Table 2, if the PFSC letter in each column for given structures is different from 2PCX, the background is remarked by yellow. In each column, the PFSC letter for 2PCX itself and other PFSC letters with yellow color are embraced by the PFSC letters of same column of PFVM in the section B. So, any conformation of protein given structure can be formed from PFVM. The Fig. 4 showed that each of 6 FiveFold predicted structures was well matched with its corresponding given structure. The results demonstrated that the FiveFold had a capability, which does not only can predict the multiple conformational structures in flexibility, but also can predict structure in accuracy to compare with



[illegible]

**Table 2.** The PFVM and PFSC for DNA-binding domain region (100–288) of P53\_HUMAN (P04637).

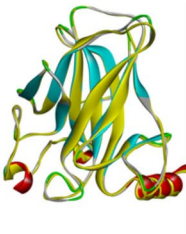


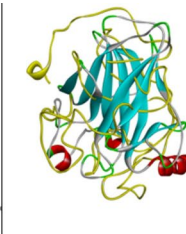
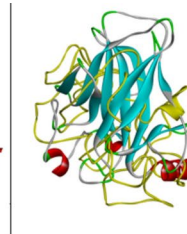
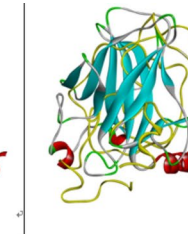

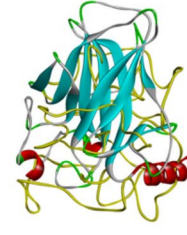
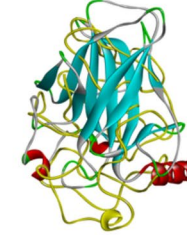
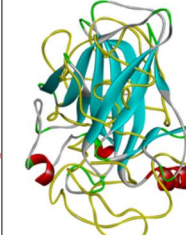
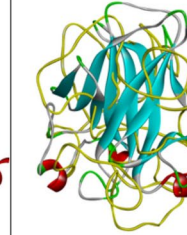
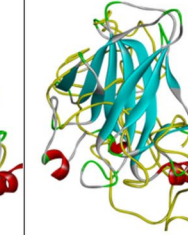
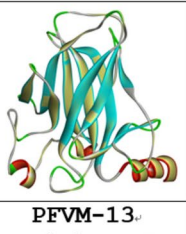
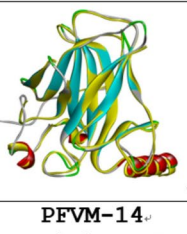

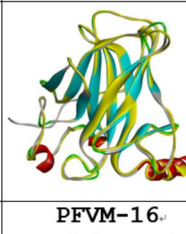
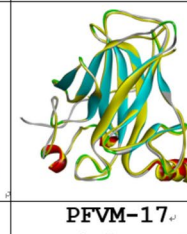
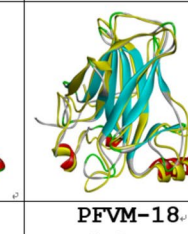
The protein sequence is listed in section A; the PFVM in section B; the conformations in PFSC string for given 3D structures in section C; the conformation of AlphaFold in PFSC string in section D and a set of conformations in PFSC string predicted by FiveFold in section E. The PFSC letters with red and pink colors are for typical helix and alike-helix local folds; the blue and light blue colors for beta strand and alike-beta strand and block color for irregular folds.

given structure. Thus, an ensemble of multiple conformation protein P53 DNA-binding domain 3D structures with dispersed folding conformations is obtained while the structure with accuracy is simultaneously predicted.

### Structure prediction for entire protein

The 53 protein is composed of four domains, such as transactivation domain1 (6–30), transactivation domain 2 (35–59), DNA-binding domain (100–288) and tetramerisation domain (319–357). The ranges of these domains are marked by yellow color for sequence background in Table 3. The conformations of any domain may be fluctuated around a stable scaffold. As the P53\_HUMAN has multiple domains in structure, higher disordered folds may happen at fragments between domains or N-terminal and C-terminal. For examples, the ranges between domains of 1–5, 31–34, 60–99 and 289–318 in PFVM do not have stable secondary structure fragment. Due to the 3D structure for entire P53\_HUMAN protein being harder to obtain by experimental measurement or computational prediction, the limited number of structural data is available. Three given protein 3D structures (8F2I, 8F2H and 6XRE) are found in PDB which covered almost full range of P53\_HUMAN protein, and a predicted structure in AlphaFold database. More of multiple conformations for P53\_HUMAN structure can be obtained by FiveFold approach. The PFVM with full range of P53\_HUMAN is displayed in the section under its sequence in Table 3. The PFSC conformations of given 3D structures and AlphaFold predicted structure as well as two FiveFold structural conformations (PFVM-01 and PFVM-02) formed from PFVM are listed at the bottom in Table 3. With PFSC colors, such as the red and pink for alpha helix, the blue and light blue for beta strand and black for irregular folding shape, it is apparently that the conformations of predicted P53 protein structures are overall aligned with given structures by secondary structure fragments. Two of FiveFold predicted 3D structures for P53 can be constructed by the PFSC strings (PFVM-01 and PFVM-02). Furthermore, the superposition between each 3D structure and 8F2I-A is visually displayed in Fig. 5. As multiple domains in P53 protein with several intrinsic disordered fragments, the structural difference in space orientation is not surprised. So, the structure superposition for protein comparison cannot obtain higher score. However, the PFSC alignment for protein comparison makes more sense because it avoids the difference in space orientation for protein fragments. With comparison of given structure of 8F2I-A, the overlap similarity score (with combination of 50% steric component and 50% electrostatic component) and the PFSA (folding structural alignment) score are listed under structures in Fig. 5. It is showed that the two FiveFold predicted structures have higher similar score



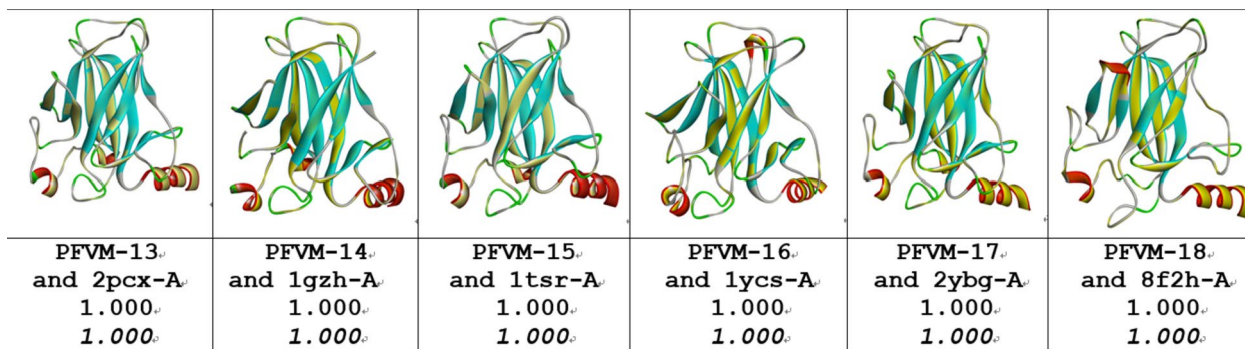
|  |  |  |   |  |  |
|--|--|--|---|--|--|
|   |   |   |   |   |   |
| PFVM-01<br>and 2pcx-A<br>0.867<br><i>0.879</i>                                     | PFVM-02<br>and 2pcx-A<br>0.405<br><i>0.754</i>                                     | PFVM-03<br>and 2pcx-A<br>0.388<br><i>0.744</i>                                     | PFVM-04<br>and 2pcx-A<br>0.349<br><i>0.709</i>                                      | PFVM-05<br>and 2pcx-A<br>0.398<br><i>0.729</i>                                       | PFVM-06<br>and 2pcx-A<br>0.344<br><i>0.648</i>                                       |
|   |   |   |   |   |   |
| PFVM-07<br>and 2pcx-A<br>0.353<br><i>0.618</i>                                     | PFVM-08<br>and 2pcx-A<br>0.425<br><i>0.711</i>                                     | PFVM-09<br>and 2pcx-A<br>0.387<br><i>0.696</i>                                     | PFVM-10<br>and 2pcx-A<br>0.385<br><i>0.709</i>                                      | PFVM-11<br>and 2pcx-A<br>0.302<br><i>0.601</i>                                       | PFVM-12<br>and 2pcx-A<br>0.324<br><i>0.590</i>                                       |
|  |  |  |  |  |  |
| PFVM-13<br>and 2pcx-A<br>1.000<br><i>1.000</i>                                     | PFVM-14<br>and 2pcx-A<br>0.796<br><i>0.746</i>                                     | PFVM-15<br>and 2pcx-A<br>0.806<br><i>0.802</i>                                     | PFVM-16<br>and 2pcx-A<br>0.814<br><i>0.802</i>                                      | PFVM-17<br>and 2pcx-A<br>0.796<br><i>0.740</i>                                       | PFVM-18<br>and 2pcx-A<br>0.643<br><i>0.686</i>                                       |

**Fig. 3.** Comparisons between the given 2PCX-A structure and 18 of FiveFold predicted structure (yellow) for DNA-binding domain (100–288) of P53\_HUMAN. The images of superposition of each pair of structures are displayed. The conformation name for FiveFold and given structure are listed under image. The overlap similarity score is displayed under structure name, which was obtained by overlaying the molecule structures with alignment by a combination of 50% steric component and 50% electrostatic component in software of Discovery Studio v24.1.0. Also, with PFSC string alignment, the score of protein folding structural alignment (PFSA) in italics is listed.

of PFSA than AlphaFold structure, which indicated the similarity in conformation despite of space orientation difference. Therefore, it demonstrated the multiple conformational structures for entire P53\_HUMAN protein with multiple domains can be predicted by FiveFold approach.

**LEF1\_HUMAN**

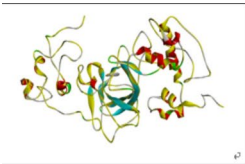
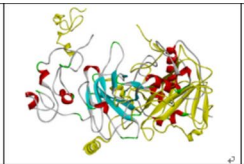
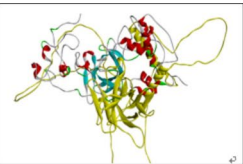
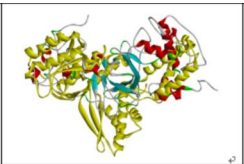
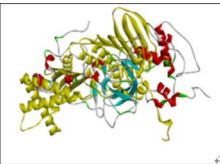
The LEF1\_HUMAN (Q9UJU2) is an intrinsically disordered protein, which is composed of 399 amino acid residues with two disorder regions, i.e. CTNNB1 binding domain (9–213) and high mobility group domain (298–368). The disordered information is available in database, such as the InterPro (integrated resource of protein families, domains and functional sites) or the protein disorder MobiDB<sup>39,40</sup>. The LEF1\_HUMAN protein contains about 88% of disordered structure which is displayed in Fig. 6, it does not have any 3D structure data available in PDB, but a predicted structure can be obtained from AlphaFold database. Although, the AlphaFold offered a single state of 3D structural data, it indicated the LEF1\_HUMAN as an intrinsically disordered protein. With FiveFold approach, the multiple conformational structures of LEF1\_HUMAN can be predicted by its PFVM. The sequence of LEF1\_HUMAN is listed in top section of Table 4, its PFVM is displayed in middle section, the conformations in PFSC for AlphaFold structure and 10 conformations obtained from PFVM are listed in bottom section. The folding similarity and dissimilarity are well exposed by alignment of these PFSC strings. It is obvious that the conformations formed from PFVM have more alpha-helical fragments than AlphaFold perdition. Also, most of the secondary structural fragments are aligned locally. According to each PFSC string formed from PFVM, its 3D structure can be constructed by FiveFold approach. Therefore,



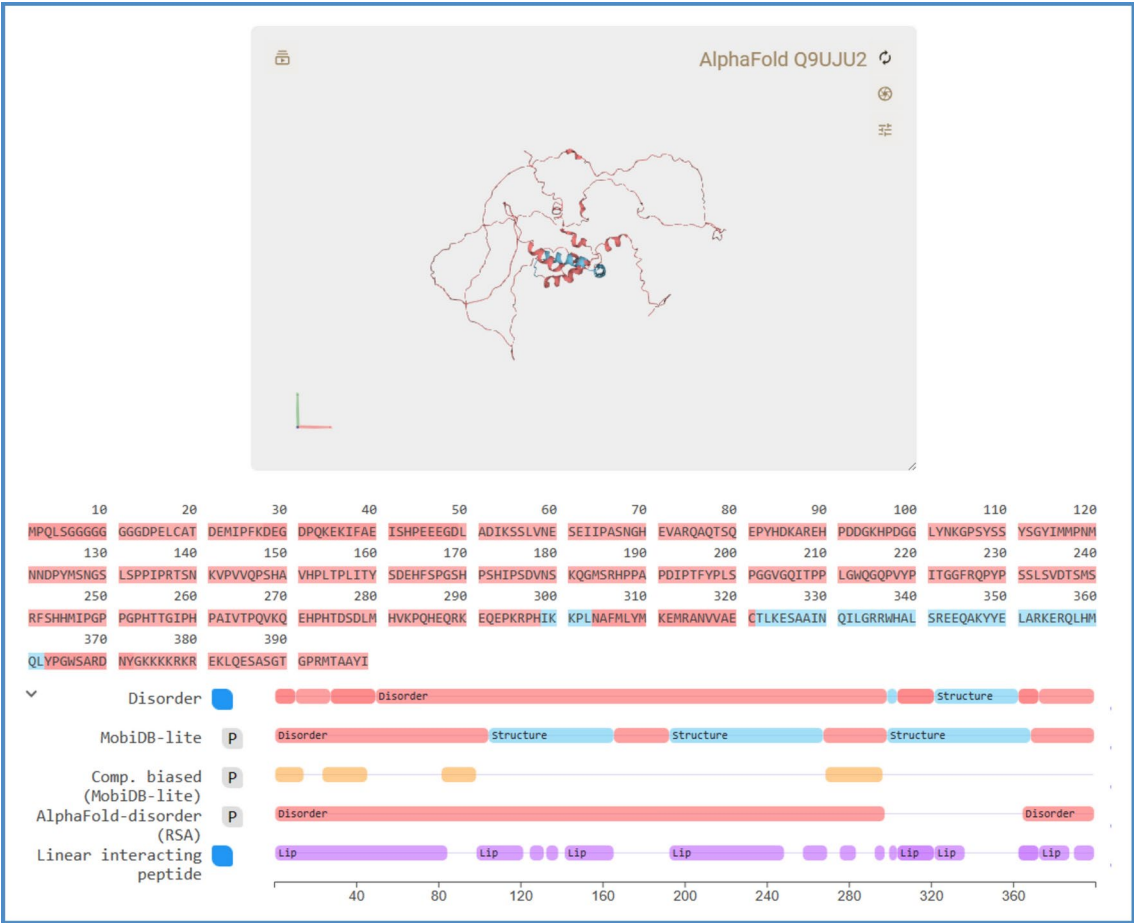
**Fig. 4.** Comparisons between the FiveFold predicted structure (yellow) and each of 6 given structures for DNA-binding domain (100–288) of P53\_HUMAN. The images of superposition of each pair of structures are displayed. The conformation name for FiveFold and given structure are listed under image. The overlap similarity score is displayed under structure name, which was obtained by overlaying the molecule structures with alignment by a combination of 50% steric component and 50% electrostatic component in software of Discovery Studio v24.1.0. Also, with PFSC string alignment, the score of protein folding structural alignment (PFSA) in *italics* is listed.

[illegible]



|   |   |  |   |   |
|---|---|--|---|---|
|  |  |  |  |  |
| <b>8F2H-A<br/>and 8F2H-A</b>  | <b>6xre-M<br/>and 8F2H-A</b>  | <b>AlphaFold<br/>and 8F2H-A</b>  | <b>PFVM-01<br/>and 8F2H-A</b>   | <b>PFVM-02<br/>and 8F2H-A</b>   |
| 1.000   | 0.243   | 0.301  | 0.204   | 0.304   |
| <i>1.000</i>  | <i>0.227</i>  | <i>0.536</i>   | <i>0.571</i>  | <i>0.525</i>  |

**Fig. 5.** Pair comparisons between 8F2I-A and two given 3D structures, AlphaFold prediction and two FiveFold predicted structures for entire of P53\_HUMAN protein. The image of superposition of each pair of structures is displayed, the colorful structure is for 8F2I-A and yellow color for compared structures. The structure names of the compared structures are listed under image. The overlap similarity score is displayed under structure name, which was obtained by overlaying the molecule structures with alignment by a combination of 50% steric component and 50% electrostatic component in software of Discovery Studio v24.1.0. Also, with PFSC string alignment, the score of protein folding structural alignment (PFSA) in italics is listed.



**Fig. 6.** The intrinsically disordered information for LEF1\_HUMAN protein from MobiDB. The 3D structure predicted by AlphaFold is displayed on top. The disorder ranges are marked by red color with various methods.

an ensemble of multiple conformation 3D structures for LEF1\_HUMAN protein are able to be obtained. Each of ten of predicted structures predicted by FiveFold is directly compared with AlphaFold structure which is displayed in Fig. 7. The images apparently showed that most of structural superimpositions are widely dispersed while one of FiveFold structures (PFVM-08) has higher protein folding structural alignment (PFSA) score. It indicated higher similarity in folding conformation despite the difference of fragment space orientations. The



[illegible]

(Continue)

[illegible]

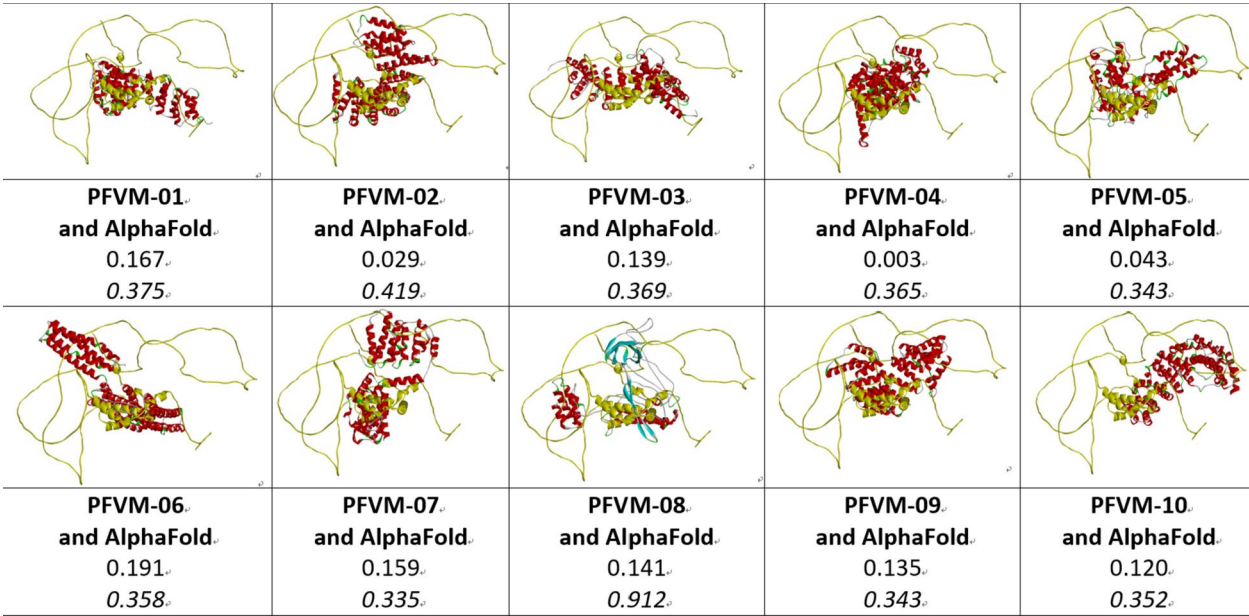
**Table 4.** The PFVM and PFSC conformations for LEF1\_HUMAN protein.

The protein sequence is listed on top section; the PFVM on middle section; a conformation of AlphaFold in PFSC string and ten of PFSC conformations formed from PFVM on bottom section. The PFSC letters with red and pink colors are for typical helix and alike-helix local folds; the blue and light blue colors for beta strand and alike-beta strand and block color for irregular folds.

results showed that the FiveFold is able to predict more of different conformation structures for the intrinsic disordered LEF1 HUMAN protein.

## Q8GT36\_SPIOL

The Q8GT36\_SPIOL (Q8GT36) is a Spinach specific protein in the photosynthetic thylakoid membrane with 103 amino acids in sequence. This protein with TSP9 as its gene name is a characteristic sample to illustrate the intrinsically disordered protein with largely flexible conformations in NMR structures<sup>41</sup>. The PFVM of Q8GT36\_SPIOL is obtained according to its sequence, then a set of possible conformations in PFSC is formed and an ensemble of 3D structures are predicted. The sequence of Q8GT36\_SPIOL protein is displayed in section A in Table 5, its PFVM in section B and 22 possible conformations in PFSC string built from PFVM at section C while the conformations of 20 models of NMR structures (PDB ID = 2FFT) expressed by PFSC strings in section D. According to the PFSC strings at section C, 10 conformational 3D structures are predicted and displayed in Fig. 8, which are compared with 20 models of NMR structures. With 3D structure comparison, it is obvious that both NMR structures and FiveFold predicted structures for the Q8GT36\_SPIOL protein certainly exhibited flexible conformations. However, the visual observation is hard specifically to illustrate how the conformations



**Fig. 7.** Structure superposition between AlphaFold structure and each of ten of FiveFold predicted structures. The image of AlphaFold structure is remarked by yellow color. Each of FiveFold predicted structure is constructed according its PFSC string in Table 4. The overlap similarity score is displayed under structure name, which was obtained by overlaying the molecule structures with alignment by a combination of 50% steric component and 50% electrostatic component in software of Discovery Studio v24.1.0. Also, with PFSC string alignment, the score of protein folding structural alignment (PFSA) in italics is listed.

are similar and different. Particularly, it is hard to compare more than two protein structures together. The PFSC alignment is easily to overcome these obstructs. With PFSC alignment of 22 FiveFold predicted conformations from PFVM in section C and 20 conformations of NMR structures in section D, the local folding similarity and dissimilarity are well uncovered. First, the conformation in range 35–43 has all alpha-helical folding features for both FiveFold predicted structures and NMR structures, and the same folding feature is also exposed at the first row PFSC in PFVM. Second, the conformation in range 78–82 of NMR structures have alpha-helical folding feature, but the FiveFold predicted structures from PFVM trend beta strand feature. Third, the conformations in other ranges certainly have various different folding features. These factors showed that the local folding similarity and difference for Q8GT36\_SPIOL protein are well exposed by PFSC alignment. Thus, the FiveFold provided a useful tool to predict an ensemble of multiple conformational structures for intrinsically disordered proteins, the PFVM directly reveal the local folding variations and the PFSC alignment well expose the folding similarity and difference for all structures along sequence.

Discussion

The FiveFold approach in protein structure fingerprint technology is a single sequence method, which depends on the order of amino acids in a sequence to predict the conformational ensembles of protein structures. Any protein sequence has its exclusive PFVM which represents its local folding characteristic along sequence. Based on its PFVM, the protein multiple conformations are formed and represented by PFSC strings, and then an ensemble of multiple conformation structures can be constructed. Below discussion is about issues how the FiveFold may impact on the protein structure prediction.

Accuracy vs. flexibility

The process of protein structure prediction is a computational approach to acquire three-dimensional structure according to its amino acid sequence. Many of molecular dynamics simulation approaches as well as AlphaFold deep learning-based system have achieved remarkable success with higher accuracy in the competition of CASP (Critical Assessment of Structure Prediction). It is indeed a big progress that the acquisition with high accuracy is archived for prediction of complex protein structure. However, the accuracy is only relative to a single state among protein flexible conformations. Most of native protein structures are conformation flexible or do not stay in a static state, i.e. it may fold and unfold itself in process for biological function, or alter the conformation as its environment and condition change. Thus, except the accuracy, the acquisition of multiple folding conformations for flexibility is another criterion to evaluate the results of protein structure prediction. Obviously, it is a pair of contradictions simultaneously to achieve both accuracy and flexibility. The protein folding process is guided by the interplay between residue interactions as well as the influence of environment, seeking to minimize the energy of the system and achieve a stable conformation or keep intrinsic disorder. The structure of a native protein is represented by the most energetically favorable state under physiological conditions. The consensus is that the order of amino acid in sequence determined its more possible folding conformations. And so, different



[illegible]

**Table 5.** The PFVM and conformations in PFSC strings for Q8GT36 SPIOL protein.

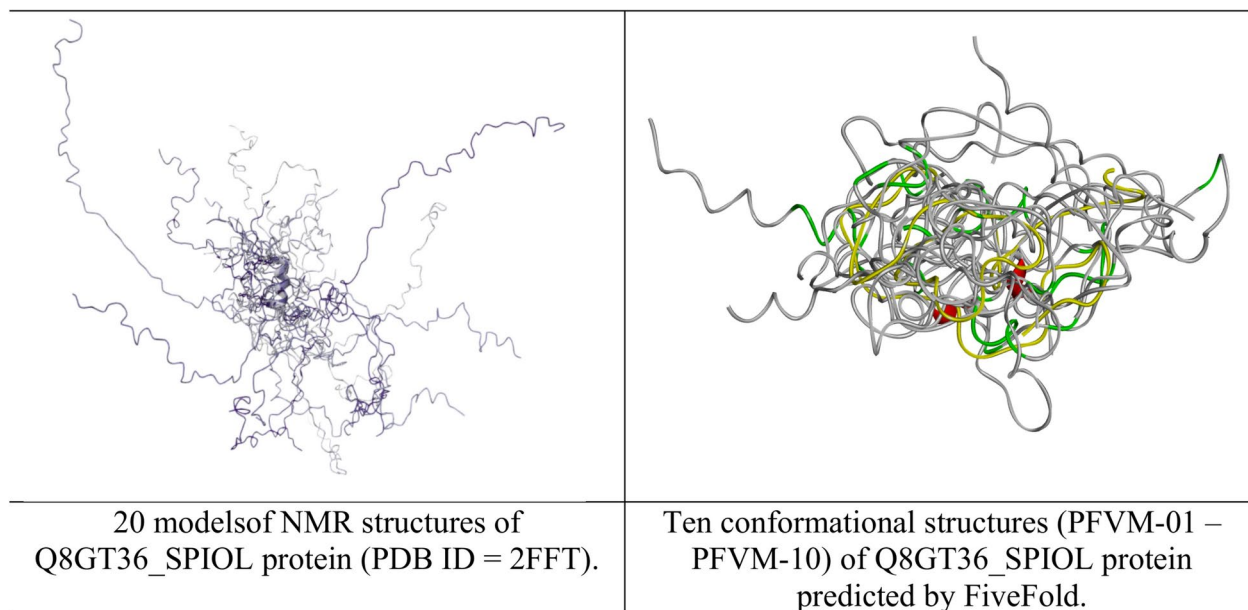
The sequence is listed in section A; its PFVM in section B; 22 conformations in PFSC string built from PFVM in section C and 20 conformations of NMR structures (PDB ID=2FFT) expressed in PFSC string in section D. The PFSC letters with red and pink colors are for typical helix and alike-helix local folds; the blue and light blue colors for beta-strand and alike-beta strand and block color for irregular folds.

proteins have different folding conformations while homology protein has similar folding conformation. The critical issue is how to predict protein structure with accuracy as well as how to obtain multiple conformations presenting flexibility. The FiveFold approach took a set of five amino acid residues as folding element, and all possible folds are described with PFSC alphabet digital letters. PFVM does not only well display the local folding fluctuations in pattern and number along sequence, but it is able exclusively to form an astronomical number of structural conformations for each individual sequence. Theoretically, the PFVM is able to provide comprehensive conformations against the protein folding problem. Practically, biological research is interesting to know a limited number of ensemble of folding conformation 3D structures. An ensemble of possible folding conformation 3D structures can be constructed by the optimized combinations using PFSC on top rows in PFVM with replacement of PFSC letters at same column. It is significant that the PFVM does not only construct the multiple conformations for flexible protein structures, but it also can predict the multiple protein structures with high accuracy against each given structure. For accuracy, the PFVM contains all necessary PFSC letters which is able to form the PFSC strings matching exact conformations for all given structures measured by experimental approaches. Therefore, according to each of these matching PFSC strings, the FiveFold is able to construct 3D structures with accuracy separately to match each given structure. Thus, the FiveFold is a useful tool for protein multiple conformation structure prediction while both accuracy and flexibility are accommodated.

## Deep learning vs. meaningful biophysical algorithm

The protein structure prediction traditionally involved with biophysical base approaches, such as comparative modelling, fold recognition or energy minimization under force fields, but they are facing either result reliability or costing expensive computationally. Recently the AI deep learning methodology of artificial neural networks, such as AlphaFold and etc. successfully predicted the protein structures with remarkable accuracy and





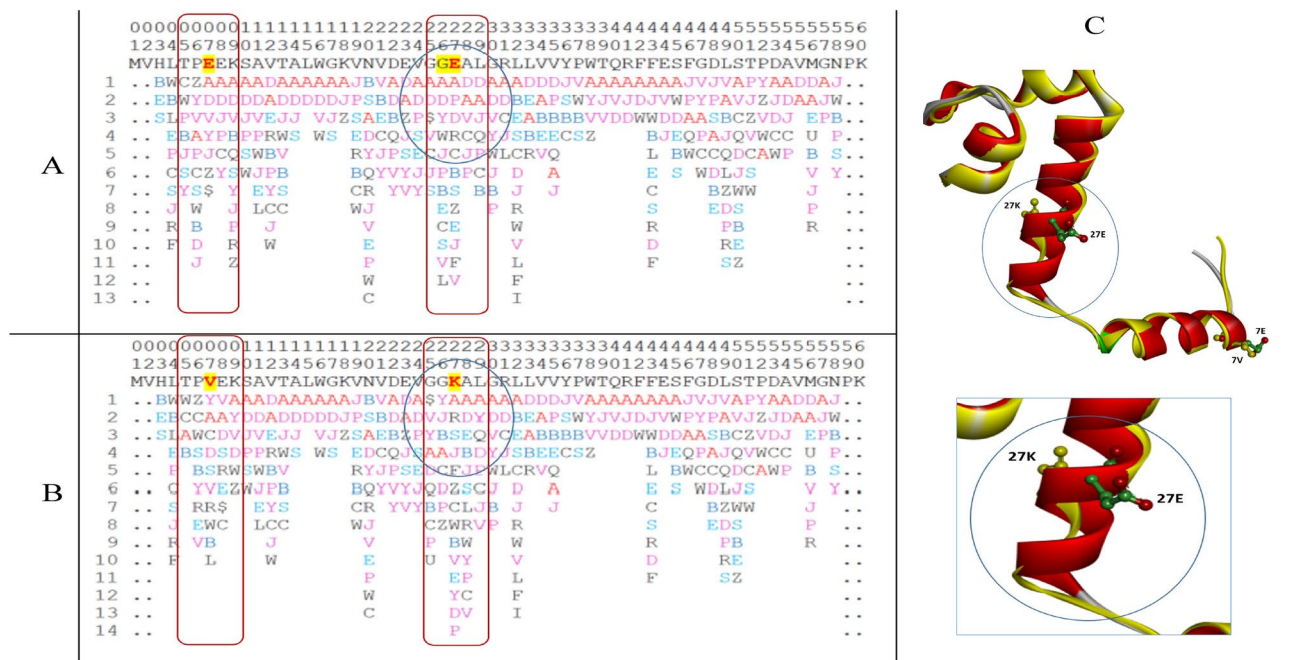
**Fig. 8.** The images of superposition for NMR structures and superposition for an ensemble of multiple conformational structures predicted by FiveFold for Q8GT36\_SPIOL protein.

effectiveness. The AlphaFold trained on large amount protein data to discover the relationship between structural feature and data to identify statistical patterns, but it does not directly extract them into a format readily to be understood. Although AI deep neural networks are really good for solving the complex protein problems, it does not directly expose the terms of theory or model for biological scientists. Also, another problem is that the AI deep learning method depended on the amount of high-quality data. If training data is lack, the reliability of prediction would be lower.

The FiveFold approach adopted a more meaningful biophysical algorithm to predict protein structure. First, the five of amino acid residues has two consecutive dihedral angles which is the smallest folding element in protein backbone. Second, a set of 27 PFSC covered complete folding space for five point connection without constrain, but less than 27 PFSC are actually needed for five amino acid residues because of constrain. Third, each PFSC is a folding shape which is represented by a digital letter for simplification, which makes the possibility easily handling a massive number of complex protein folds. Fourth, any set of five amino acids have different number of various folding patterns but could not exceeding 27, which avoided the infinite folding images. Fifth, based on a sequence, its PFVM can be obtained, each column lists the folding shapes in PFSC for five amino acids which well represented the local folding fluctuations along sequence. Sixth, more important is that each protein has its specific PFVM, and a massive number of PFSC strings are able to be formed to represent multiple conformations. Finally, the multiple conformation protein 3D structures can be constructed based on PFSC strings with homologous conformation search, and then an ensemble of protein 3D structures are predicted for any protein. Thus, the meaningful biophysical algorithm is permeated into each step in FiveFold approach for protein structure prediction. The unique features and capabilities between AlphaFold and FiveFold can be compared. In summary, AlphaFold depended on multiple sequence alignment, but FiveFold adopted single sequence. The AlphaFold provided a protein structure in a single state and indicated the regions of intrinsic disorder, the FiveFold is able to provide a set conformational structures for protein multiple states and is able to provide various folding patterns for intrinsic disordered regions. The AlphaFold applied AI technology to predict the complex protein structures, the FiveFold adopted digital alphabetic description for five residues to simplify the complex protein folding problem.

#### MSA vs. single sequence method

The multiple sequence alignment (MSA) is popularly applied by the AI deep learning process for protein structure prediction. It trains multiple sequences in protein database to obtain the predicted structure, so the results relied on plenty of homologous sequences<sup>34</sup>. Generally, it is true that the homologous sequences have similar folding conformation. However, the protein folding pattern is fundamentally determined by the order of amino acids in sequence. Thus, any difference in sequence may impact the protein folding conformation in various degrees. Especially it is significant to study protein mutation, to investigate the same protein for differentiation species or to design protein. Nevertheless, due to MSA application AlphaFold does not have the capability to distinguish the structural difference caused by few residue differences in mutation or species difference. In order avoid MSA process, some of single sequence methods, such as RGN2, ESMFold, OmegaFold and HelixFold-Single etc., have been developed with adoption of large language models (LLM) to predict protein structure. So far, however, it is still hard to reveal the structure difference caused by a few residue differences. Also, it does not provide knowledge for human-understandable form and the higher reliability in results. It is still a challenge to reveal



**Table 6.** The conformation structure difference affected by mutations in Hemoglobin subunit beta (HBB\_HUMAN) protein.

The PFVM before mutation is displayed in section A and the PFVM after mutation in section B. The differences of local folding features are marked in red rectangle and the folding features for E27K in oval. The 3D structures before and after mutation were predicted by FiveFold according to the PFSC string on the first row of PFVM, and superimposed and displayed in section C. The structure after mutation is yellow color, and the mutated residues are labelled around.

the protein structural difference caused by slight variance in sequence, which is important to impact protein biological function, to understand the cause of diseases or to design new protein for biomedicine.

The FiveFold approach is a single sequence method, which is only based on its specific sequence to predict protein structure and is able to distinguish the structure difference caused by individual residue replacement. As different proteins have different sequences, they have different PFVM. Thus, although the same protein for different biological species is higher homologous, the conformational difference can be directly revealed by PFVM. Also, a single residue mutation in sequence will impact local fold patterns in five columns, which can be explicitly exposed in PFVM. The mutations in Hemoglobin subunit beta (HBB\_HUMAN) protein caused different diseases, for example, the residue mutation of E7V caused Sickle cell disease and E27K caused Haemoglobin E. The residue mutations of E7V and E27K triggered the different local folding variations in PFVM, which are displayed in Table 6. The different local folding features are caused by mutations, and the folding change ranges are marked in red frames. With FiveFold, the 3D structures before and after mutation may be separately predicted by PFVM-01, which is the PFSC string on the first row of PFVM. It is obvious that the mutation E27K changed the residue physicochemical property from acidic to alkaline, and also the folding feature of typical alpha helix is deformed, which are marked in ovals in section A, B and C. The structure after mutation is colored with yellow in section C. It is obviously that the deformational helix around E27K twisted and caused the residue side chain to orientate different direction. Therefore, the PFVM is a significant single sequence approach with capability to reveal the folding difference in structure for both protein mutation and the same proteins of species differentiation.

### Future development

It is significant to apply the FiveFold approach to predict protein structures and obtain multiple conformational structures. In PFVM, the protein local folding patterns are well represented by digital PFSC alphabetic letters in each column along sequence, which may benefit systematically to investigate the relationship between folding pattern changes and biological functions for complex proteins. However, there are many issues are needed to further advance and to associated with pharmaceutical and disease studies. First, a PFVM database may need to be generated for biological researchers to directly access. Also, the database can explicitly provide the folding patterns for study of intrinsically disordered protein. Second, with digital description, the PFVM may provide a useful means to explore the correlation between change of folding conformation and protein function or disease. It may become a systematical platform to associate with the cumulated rich experimental and clinical data. Third, to extract a limited number of optimum conformations from an astronomical number of possibilities of PFSC letter combinations from PFVM is a neural network process with multiple pathways, and hope an improved AI process to play better roles in future. Forth, the PFVM provided an effective platform with digital

alphabetic letters presenting the comprehensive local folding shapes along protein sequence. It is significant to apply the PFVM to further explore an astronomical number of conformations for the protein folding problem as well as to provide concrete folding patterns for intrinsically disordered proteins.

## Data availability

The protein sequences in this study are available in UniProt database: P53\_HUMAN: <https://www.uniprot.org/uniprotkb/P04637/entry> LEF1\_HUMAN: <https://www.uniprot.org/uniprotkb/Q9UJU2/entry> Q8GT36\_SPIOL: <https://www.uniprot.org/uniprotkb/Q8GT36/entry> The related protein 3D structures are available in PDB database: <https://www.rcsb.org/structure/2PCX> <https://www.rcsb.org/structure/1GZH> <https://www.rcsb.org/structure/1TSR> <https://www.rcsb.org/structure/1YCS> <https://www.rcsb.org/structure/2FEJ> <https://www.rcsb.org/structure/2MEJ> <https://www.rcsb.org/structure/5BUA> <https://www.rcsb.org/structure/5LGJ> <https://www.rcsb.org/structure/2AC0> <https://www.rcsb.org/structure/6XRE> <https://www.rcsb.org/structure/8F2I> Alphafold Protein Structure Database: <https://alphafold.com/search/text/P04637> <https://alphafold.com/search/text/Q9UJU2> The raw data in this study are available in supplementary files. Supplementary-01.rtf and supplementary-02.pdb are for P53\_HUMAN protein; Supplementary-03.rtf for LEF1\_HUMAN protein; Supplementary-04.rtf for Q8GT36\_SPIOL protein.

Received: 24 July 2024; Accepted: 19 December 2024

Published online: 12 March 2025

## References

- Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**(7873), 583–589 (2021).
- Stoddart, C. Structural biology: How proteins got their close-up. *Knowable Mag.* <https://doi.org/10.1146/knowable-022822-1.S2C> **ID247206999** (2022).
- Baker, D. Metastable states and folding free energy barriers. *Nat. Struct. Biol.* **5**, 1021–1024 (1998).
- Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, pathways, and the energy landscape of protein-folding: a synthesis. *Proteins* **21**, 167–195 (1995).
- Bai, Y. & Englander, S. W. Future directions in folding: The multi-state nature of protein structure. *Proteins* **24**(2), 145–151 (1996).
- Levinthal, C. How to fold gracefully. In *Mossbauer Spectroscopy in Biological Systems*, pp 22–24 (Allerton House, Monticello, 1969).
- Dill, K. A. Dominant forces in protein folding. *Biochemistry*. **29**(31), 7133–7155 (1990).
- Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science*. **338**(6110), 1042–1046 (2012).
- Chong, S. H. & Ham, S. Folding free energy landscape of ordered and intrinsically disordered proteins. *Sci. Rep.* **9**, 14927 (2019).
- Gruebele, M. The fast protein folding problem. *Annu. Rev. Phys. Chem.* **50**, 485–516 (1999).
- Dobson, C. M., Sali, A. & Karplus, M. Protein folding: a perspective from theory and experiment. *Angew. Chem. Int. Ed. Engl.* **37**, 868–893 (1998).
- Uversky, V. N. Intrinsically disordered proteins from A to Z. *Int. J. Biochem. Cell Biol.* **43**(8), 1090–1103 (2011).
- Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Ann. Rev. Biochem.* **83**, 553–584 (2014).
- Shaw, D. E. et al. Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346 (2010).
- Kendrew, J. C. et al. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*. **181**(4610), 662–666 (1958).
- Zhuravleva, A. & Korzhnev, D. M. Protein folding by NMR. *Progr. Nucl. Magn. Reson. Spectrosc.* **100**, 52–77 (2017).
- <https://www.rcsb.org/>
- Royer, C. A. Probing protein folding and conformational transitions with fluorescence. *Chem. Rev.* **106**(5), 1769–1784 (2006).
- Mashaghi, A., Kramer, G., Lamb, D. C., Mayer, M. P. & Tans, S. J. Chaperone action at the single-molecule level. *Chem. Rev.* **114**(1), 660–676 (2014).
- Sample, I. Google's DeepMind predicts 3D shapes of proteins. *Guardian*. **2**, 1–2 (2018).
- Heaven, W. D. DeepMind's protein-folding AI has solved a 50-year-old grand challenge of biology. *MIT Technol. Rev.* **30**, 1–2 (2020).
- Schmidt, T., Bergner, A. & Schwede, T. Modelling three-dimensional protein structures for applications in drug design. *Drug Discov. Today* **19**, 890–897 (2014).
- Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
- Kroese, D. P., Brereton, T., Taimre, T. & Botev, Z. I. Why the Monte Carlo method is so important today. *WIREs Comput. Stat.* **6**, 386–392 (2014).
- Shell, M. S., Ozkan, S. B., Voelz, V., Wu, G. A. & Dill, K. A. Blind test of physics-based prediction of protein structures. *Biophys. J.* **96**, 917–924 (2009).
- Zheng, W. et al. Folding non-homology proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Reports Methods* **1**, 100014 (2021).
- Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**(Web Server issue), W244–W248 (2005).
- Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* <https://doi.org/10.1038/s41586-024-07487-w> (2024).
- Varadi, M. et al. massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**(D1), D439–D444 (2022).
- Stephen Curry, No, DeepMind has not solved protein folding, Reciprocal Space (blog), 2 December 2020.
- Doerr, A. Protein design: the experts speak. *Nat. Biotechnol.* **42**, 175–178 (2024).
- Wayment-Steele, H. K. et al. Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).
- Chakravarty, D. & Porter, L. L. AlphaFold2 fails to predict protein fold switching. *Protein Sci.* **31**(6), e4353 (2022).
- Sala, D., Engelberger, F., Mchaourab, H. S. & Meiler, J. Modeling conformational states of proteins with AlphaFold. *Curr. Opin. Struct. Biol.* **81**, 102645 (2023).
- Yang, J. Comprehensive description of protein structures using protein folding shape code. *Proteins* **71**(3), 1497–1518 (2008).
- Yang, J. et al. Comprehensive folding variations for protein folding. *Proteins*. **90**(11), 1851–1872 (2022).
- Brooks, B. R. et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**(10), 1545–1614 (2009).



38. Yang, J. Complete Description of Protein Folding Shapes for Structural Comparison. In *Series: Protein Biochemistry, Synthesis, Structure and Cellular Functions: Protein Folding*, (ed. Walters, E.C.) 421–442 (Nova Science Publishers, New York, 2011) (ISBN: 978–1–61761–259–6).
39. <https://www.ebi.ac.uk/interpro/protein/UniProt/Q9UJU2/>
40. <https://mobidb.org/Q9UJU2>
41. Song, J., Lee, M. S., Carlberg, I., Vener, A. V. & Markley, J. L. Micelle-induced folding of spinach thylakoid soluble phosphoprotein of 9 kDa and its functional implications. *Biochemistry*. **45**(51), 15633–15643 (2006).

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (82230119, 82104497), the National Key R&D Program of China (2021YFC1712805), Guangdong Key R&D Program (2023B1111030002), Provincial Science and Technology Plan Projects in Guangdong Province (2022A0505020017), Shenzhen International Collaborative Project (GJHZ20210705141405015), SIAT Innovation Program for Excellent Young Researchers (2022), and Shanghai Frontiers Science Center for Biomacromolecules and Precision Medicine (S.Z.).

## Author contributions

J. Yang overall designed and managed the project and made the manuscript preparation. G. Wu involved input data preparation; W. X. Cheng involved output data collection and analysis; J. Yang, Q. Hu and Q. Shi involved data performance; S. Zhao and W. Ji involved study and discuss; S. T. Sheng deployed software code and maintain web server; P. Zhang tested and verified the code and final data.

## Funding

National Natural Science Foundation of China (82230119, 82104497); Key Technologies Research and Development Program (2021YFC1712805); Special Project for Research and Development in Key areas of Guangdong Province (2023B1111030002); Shenzhen International Collaborative Project (GJHZ20210705141405015); Shanghai Frontiers Science Center for Biomacromolecules and Precision Medicine (S.Z.).

## Declarations

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-84066-z>.

**Correspondence** and requests for materials should be addressed to J.Y.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025