



Protein structure prediction in the deep learning era

Zhenling Peng^{1,a}, Wenkai Wang^{2,a}, Renmin Han^{1,a},
Fa Zhang³ and Jianyi Yang¹

Abstract

Significant advances have been achieved in protein structure prediction, especially with the recent development of the AlphaFold2 and the RoseTTAFold systems. This article reviews the progress in deep learning-based protein structure prediction methods in the past two years. First, we divide the representative methods into two categories: the two-step approach and the end-to-end approach. Then, we show that the two-step approach is possible to achieve similar accuracy to the state-of-the-art end-to-end approach AlphaFold2. Compared to the end-to-end approach, the two-step approach requires fewer computing resources. We conclude that it is valuable to keep developing both approaches. Finally, a few outstanding challenges in function-orientated protein structure prediction are pointed out for future development.

Addresses

¹ Ministry of Education Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao 266237, China

² School of Mathematical Sciences, Nankai University, Tianjin 300071, China

³ School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China

Corresponding authors: Yang, Jianyi (yangji@sdu.edu.cn); Zhang, Fa (zhangfa@ict.ac.cn)

^a Co-first authors.

Current Opinion in Structural Biology 2022, 77:102495

This review comes from a themed issue on **Folding and Binding**

Edited by **Annalisa Pastore** and **Eugene Shakhnovich**

For a complete overview see the [Issue](#) and the [Editorial](#)

Available online xxx

<https://doi.org/10.1016/j.sbi.2022.102495>

0959-440X/© 2022 Elsevier Ltd. All rights reserved.

Introduction

Protein structure prediction aims to predict the 3D structure from amino acid sequence, which is regarded as one of the grand challenges in computational biology [1]. The progress in protein structure prediction is very slow until the last decade. Especially, the deep learning-based AlphaFold2 system increases the accuracy of protein structure prediction to an unprecedented level

[2]. AlphaFold2 is thus well accepted as one of the milestones in the field [3].

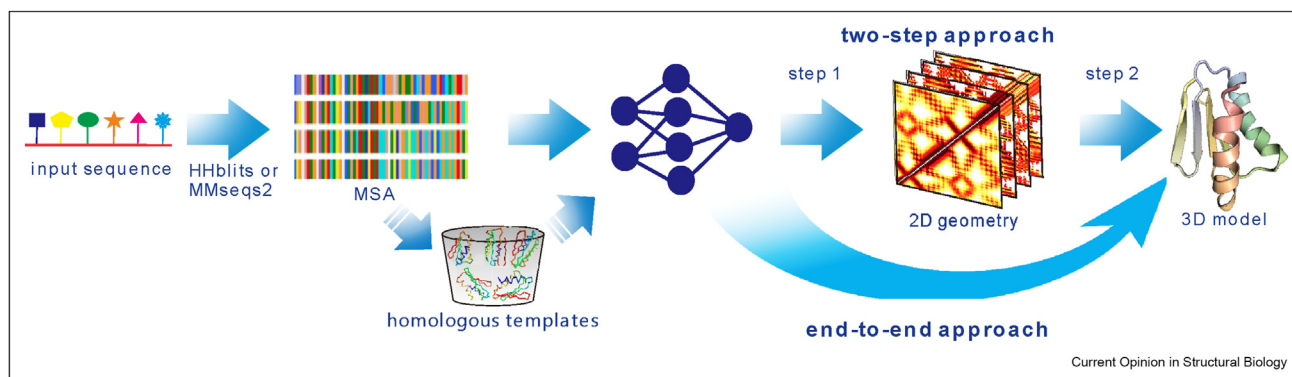
There are three essential factors contributing to the breakthrough in protein structure prediction. The first one is the availability of big biological data, including experimentally determined protein sequences and structures. The second one is the significant advance in the development of computer hardware, including Graphics Processing Unit (GPU) and Tensor Processing Unit (TPU). The last one is the rapid progress of deep learning algorithm, such as residual convolutional network (ResNet) [4] and attention-based transformer [5].

This article presents a brief review of the representative deep learning-based protein structure prediction methods developed in the past two years. A more comprehensive review can be found in the studies by Pearce et al. [6,7]. [Figure 1](#) shows the major steps involved in deep learning-based protein structure prediction methods. Starting from the amino acid sequence of a protein, homologous sequences are first collected with sequence alignment tools such as HHblits [8] and MMseqs2 [9], to construct a multiple sequence alignment (MSA). Either the raw MSA or MSA-derived co-evolution features are then fed into a deep neural network for structure prediction. Homologous structure templates (when available) can be also included in the network. Depending on how the structure is predicted, the methods can be classified into two categories: two-step approach and end-to-end approach ([Table 1](#)).

Two-step approach

The two-step approach divides the task of protein structure prediction into inter-residue 2D geometry prediction with deep learning (step 1) and 3D structure realization (step 2). Correspondingly, there are two key components here. One is what 2D geometry to predict in the first step, and the other one is how to convert the predicted 2D geometry into 3D structure in the second step. Representative methods include RaptorX-Contact [10,11], AlphaFold [12], trRosetta [13,14], trRosettaX [15], ProFOLD [16], tFold [17], DMPfold [18], D-I-TASSER [19], MULTICOM [20], and so on.

Figure 1



Major steps in deep learning-based protein structure prediction methods. Homologous templates are optional, which is reflected by dashed arrows. The two-step and the end-to-end approaches are indicated by straight and curved arrows, respectively.

For the first component, the inter-residue contact map is used in the earlier works [11,21], which is later extended to the distance map [10]. trRosetta takes one step further by predicting both distance and orientations. The application of the ResNet in RaptorX-Contact successfully doubled the precision of predicted contacts [11]. This improvement is mostly because the contact map is predicted globally rather than locally. That is to say, the network predicts the whole contact map run rather than a few selected elements [11] in each run. After this work, ResNet and its variants are then used extensively in almost all subsequent methods to predict 2D geometry [12–20] (see Table 1).

The second component is to build 3D structure model from the predicted 2D geometry. With native 2D geometry, it is straightforward to construct 3D structure using the distance geometry-based CNS package [22], by converting the 2D geometry into constraints. However, what we have in hand is not experimental but predicted 2D geometry, which usually contains noises and conflicted contacts/distances (e.g., not satisfying the triangle inequality). As a result, we need to specify an appropriate set of constraints, making it very inconvenient to use. In addition, it becomes very slow for CNS to construct reasonable 3D structure models for large proteins (e.g., with more than 500 amino acids). In D-I-TASSER, the constraints are used to guide the Replica Exchange Monte Carlo (REMC) simulations [23] to generate 3D structure models, which are usually more accurate than CNS at the expense of more running time. In AlphaFold, gradient descent energy minimization with L-BFGS [24] is applied; but its implementations are largely unknown. An efficient solution is given in trRosetta, which converts predicted distance and orientations into smooth energy functions and builds 3D structure models quickly with the quasi-Newton method L-BFGS with PyRosetta. trRosetta was shown

to outperform all previously described methods in benchmark tests on the CASP13 and the CAMEO-derived datasets [13]. In addition, both web server and open-source codes are provided for trRosetta, which greatly benefits subsequent research in the field [15–17,25,26].

End-to-end approach

The end-to-end approach predicts 3D structure directly within one unified network in one step. Compared to the two-step approach, end-to-end structure prediction is attractive but more difficult to implement. For example, AlQuraishi developed an end-to-end differentiable model (named RGN) with the recurrent geometric network, which takes the input of amino acid sequence and position-specific scoring matrix (PSSM, derived from MSA) and outputs 3D structure [27]. As shown in a recent study that does not use co-evolution information [26], RGN is not comparable to other state-of-the-art two-step methods.

The first working version of the end-to-end approach is AlphaFold2 from DeepMind [2], which outperforms all existing methods by a large margin according to the CASP14 experiment. AlphaFold2's ablation study shows that its success can be attributed to the combination of a few key components (Figure 2). The first one is the attention-based network (termed as Evoformer, 48 blocks), which takes the input of raw MSA and structure templates and outputs two representations (MSA and pair). From the input, the MSA and pair representations are first initialized and then exchanged and updated iteratively through attention and triangle update. The second one is its attention-based structure module termed Invariant Point Attention (IPA, 8 blocks), which takes the output from Evoformer and outputs 3D structure. Unlike the two-step approach, this network (module) is connected with the first network, enabling

Table 1

The representative protein structure prediction methods developed in the past two years (i.e., since 2020).

Category	Method	2D geometry	Network	3D realization	URL of web server and/or standalone package
Two-step approach	AlphaFold [12]	Distance	ResNet	L-BFGS	N/A
	trRosetta [13,14]	Distance and Orientations	ResNet	L-BFGS in PyRosetta	https://yanglab.nankai.edu.cn/trRosetta/
	trRosettaX [15]	Distance	multi-scale ResNet	REMC	https://yanglab.nankai.edu.cn/trRosetta/download/
	D-I-TASSER [19]	Contact	ResNet	CNS	https://zhanggroup.org/D-I-TASSER/
	MULTICOM [20]	Distance	ResNet	L-BFGS in PyRosetta	http://sysbio.met.missouri.edu/multicom_cluster/ https://github.com/multicom-toolbox/multicom
End-to-end approach	ProFOLD [16]	Distance	ResNet	L-BFGS in PyRosetta	http://protein.ict.ac.cn/ProFOLD
	iFold [17]	Distance and Orientations	ResNet	L-BFGS in PyRosetta	https://github.com/fusong-ju/ProFOLD
	RoseTTAFold (pyRosetta) [25]	Distance and Orientations	Three-track transformer	L-BFGS in PyRosetta	https://drug.ai.tencent.com/ https://robetta.bakerlab.org/
	AlphaFold2 [2]	Distance	Evoformer	IPA, AMBER	https://github.com/RosettaCommons/RoseTTAFold/
	ColabFold [28]	Distance and Orientations	Three-track transformer	SE(3)-transformer	https://github.com/deepmind/alphafold/ https://github.com/sokrypton/ColabFold/ https://robetta.bakerlab.org/
	RoseTTAFold (e2e) [25]	Distance	Three-track transformer	Learnable MDS	https://github.com/RosettaCommons/RoseTTAFold/ https://github.com/psipred/DMPfold2
	DMPfold2 [30]	Distance	ResNet		

direct reasoning about the spatial and evolutionary relationships. The last one is the recycling iterations (kind of structure refinement), which is especially useful in improving the modeling for hard targets with more number of iterations. Of course, a few other factors, such as self-distillation, enriched MSA from metagenome database, MSA clustering, structure relaxation, and extensive training with the TPU facility also help.

Unlike AlphaFold, AlphaFold2 is available to the public with open source codes for inference (though the training codes are not released yet). The methodology and the source codes are well documented with details, which is especially helpful for computational researchers. Meanwhile, with the Colab resource, the ColabFold web server is available, which integrates AlphaFold2, RoseTTAFold, and MMseqs2 for fast and accurate structure modeling [28]. This is especially useful for researchers from the experimental community. To make it even more convenient, AlphaFold2 was applied to predict >23,000 structures for the human proteome [29]. The first release (November 2021) of the AlphaFold DB covers more than 360,000 predicted structures for 21 model organisms. The database now contains structure models for more than 200 million proteins from the protein universe (as of July 2022).

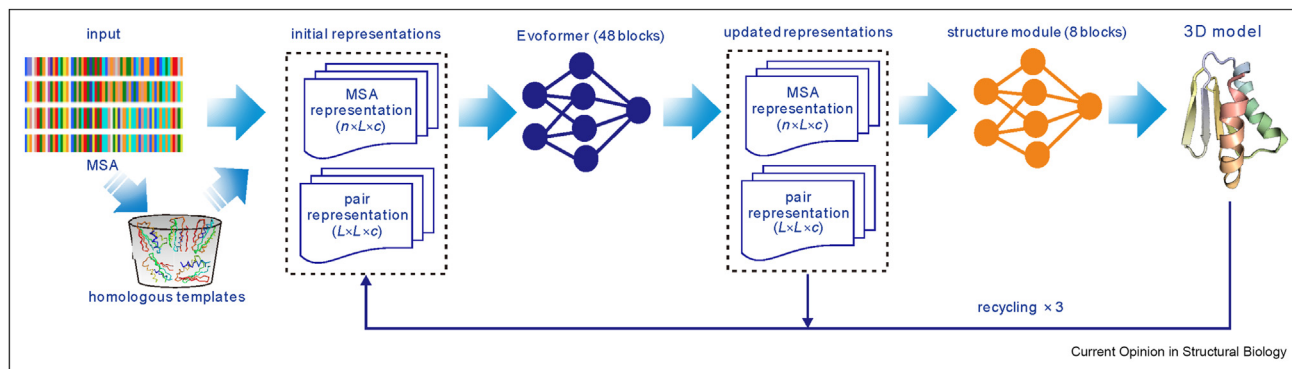
Inspired by AlphaFold2, the Baker group developed RoseTTAFold using a three-track network [25]. RoseTTAFold is less accurate than AlphaFold2 but uses fewer computing resources. DMPfold2 is another end-to-end method for ultrafast structure prediction, though less accurate than other state-of-the-art methods [30]. We believe that more end-to-end methods will be developed based on these advances.

Remarks on two-step and end-to-end approaches

From what is described above, it seems that the end-to-end approach outperforms the two-step approach significantly. Does this mean that shall we give up the two-step approach? In our opinion, the answer is ‘no’ based on the following observations.

First, we adapt a few components (i.e., Evoformer) from AlphaFold2 and incorporate them into the first step of trRosettaX. The second step of structure realization remains unchanged. We name the new version by trRosettaX2. Preliminary tests on the CASP14 targets (Figure 3a) show that the predicted structure models by trRosettaX2 are >10% more accurate than trRosettaX. In addition, trRosettaX2 outperforms RoseTTAFold and approaches AlphaFold2 (Figure 3b). The ~5% difference with AlphaFold2 may be explained by AlphaFold2’s structure module, larger network (48 vs 12 Evoformer blocks), extensive training on TPU, self-distillation, and recycling, which are not fully considered in trRosettaX2

Figure 2



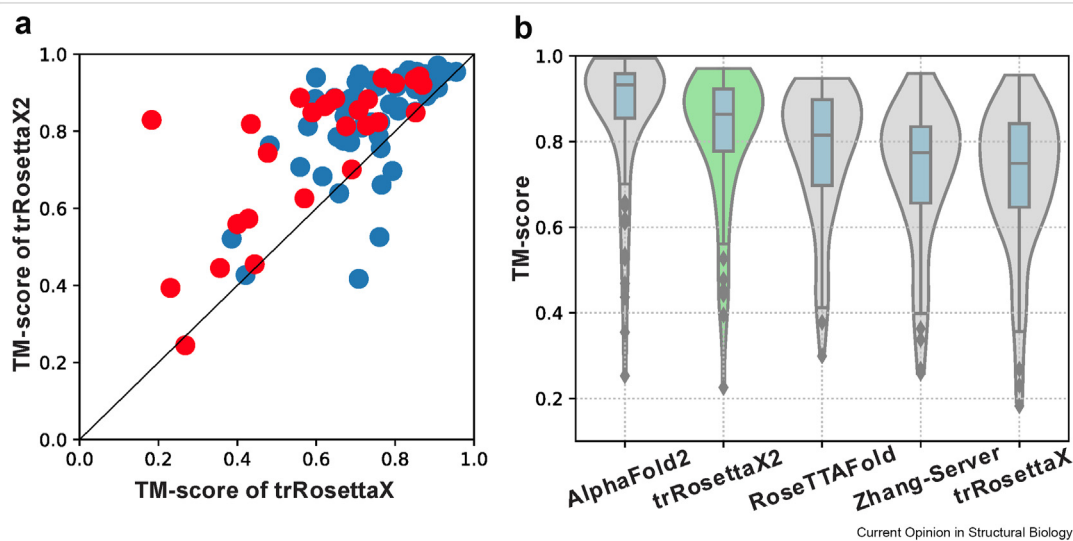
Key components of the AlphaFold2 system for protein structure prediction. The variables n , L , and c are the number of sequences in MSA, the length of the query protein, and the number of feature channels, respectively.

due to the limit of our computing resource. Thus, we conclude that with a more accurate 2D geometry prediction in the first step, the two-step approach is possible to achieve similar accuracy of the end-to-end approach.

Second, the two-step approach has its merit that is less obvious in the end-to-end approach. The first is in the model training. In general, the two-step approach takes fewer computing resources to train than the end-to-end approach, due to the reduced complexity of tasks. For example, trRosettaX2 took only one A100 GPU card

(40 GB) to train for about 10 days. In comparison, AlphaFold2 took 128 TPU v3 cores to train for about two weeks. The second is in the inference and applications. For the two-step approach, the major running time is in the second step of 3D structure generation, which is not always needed in some applications, such as in protein design [31]. In general, thousands of proteins are designed in a project while only a small portion of them are foldable. It is enough to filter out those designs that are not foldable using the predicted 2D geometry in the first step only, which is very fast. In comparison, the end-to-end approach like AlphaFold2 can also output 2D

Figure 3



Performance of the representative protein structure prediction methods on 91 CASP14 targets. a, comparison between the two-step approach trRosettaX and its improved version trRosettaX2. The red and blue points refer to the FM + FM/TBM targets and the TBM targets, respectively. b, comparison between the two-step and the end-to-end methods. End-to-end: AlphaFold2; two-step: trRosettaX, trRosettaX2, RoseTTAFold (pyRosetta version), Zhang-Server (i.e., D-I-TASSER).

geometry but with lower speed and more computing resources [15].

To summarize, the end-to-end approach is appealing and has achieved tremendous success. Nevertheless, it is worthwhile of keeping developing the two-step approach, which usually takes fewer computing resources and is thus friendly for labs with limited hardware conditions.

Outstanding challenges in protein structure prediction

AlphaFold2 is recognized as one of the milestones in protein structure prediction, which “will change everything” [3]. However, the protein structure prediction problem has not been solved [32]. There are still many outstanding challenges [33], especially in biological function-oriented protein structure prediction, which are outlined below.

- (1) The existing methods rely on the availability of MSA. There are many orphan proteins without any sequence homologies in the existing sequence databases (such as some proteins from viruses). The modeling accuracy for such proteins is in general poor. For example, in the predicted structure models for the human proteome, about 42% of residues are predicted with low accuracy (with pLDDT < 70) [29]. Some of such residues/proteins are disordered, which is also challenging for experimental determination. For other low-accuracy residues/proteins, lacking high-quality MSA is the key. Thus, it is valuable to develop new methods to improve the accuracy for orphan proteins. A few groups are trying to develop single-sequence methods using pre-trained language models [34–37]. However, these methods need to be explained with caution. Though MSA is not used directly, the pre-trained language models may contain implicit statistics from MSA [38]. Thus, it is mandatory to test them on orphan proteins rigorously to verify their performance.
- (2) Dynamic structures and folding pathway. Protein function is closely related to dynamic structures and folding pathway [39], such as alternative conformations and disordered proteins. Usually conventional molecular dynamics is applied for studying protein dynamics, but with limited success. Due to the lack of enough experimental data for dynamics, we do not anticipate significant progress shortly. In a recent study, Outeiral *et al.* demonstrated that the state-of-the-art structure prediction methods failed to generate meaningful folding pathways [40]. Nevertheless, we believe that more new methods will be developed in the future, with the rapid development of deep learning algorithms and the progress in experimental techniques.
- (3) Protein-ligand/drug interactions. The improved accuracy of predicted structures was anticipated to have implications to pharmaceutical research. However, scientists in the field of drug discovery are not so optimistic [41]. The first concern is the uncertainty about the accuracy of the active site and the side chains. The second concern is that protein structures are dynamic and thus some may undergo significant conformation changes upon binding to ligand/drug. This makes molecular docking an inappropriate solution to predicting protein-ligand/drug complex structure. An ideal way is to predict protein and ligand/drug structures simultaneously to better reflect their interactions. Based on the advance from AlphaFold2, there would be some new methods in this direction.
- (4) Protein–protein interactions. Conventional methods apply molecular docking to construct protein–protein complex structures, with the input of monomer structures [42,43]. Like protein-ligand interactions, the protein structure conformation may be changed upon binding to other proteins. In addition, accurate docking depends on the availability of high-accuracy monomer structure models, which is not always true. To address these issues, the Baker and the Elofsson groups demonstrate that the ‘fold-and-dock’ approach seems promising to predict accurate protein–protein complex structure from sequence inputs [44–46]. Based on a combination of RoseTTAFold and AlphaFold2, the Baker group generated structure models for 106 previously unidentified core eukaryotic protein complexes [47]. Recently, AlphaFold-Multimer [48] was developed following a similar architecture to AlphaFold2, which outperforms the docking approach [43] and AlphaFold2-based approach [45]. Nevertheless, the accuracy of AlphaFold-Multimer and its variant AF2Complex [49] is not stable, making proteome-wide predictions much more difficult than monomers. It is even more challenging to assemble large protein complexes [50]. Probably, the interplay between experimental data and deep learning algorithms may be an effective way to solve the problem of structure assembly for multimers in the future [51].
- (5) Effects of missense mutations. Precision prediction of the effects of missense mutations is a challenging task, partly due to the lack of large-scale and high-quality experimental data [52]. The structure-based prediction approaches might be improved using AI-predicted structures. However, it was shown that no significant correlation was observed between the mutation effects and the modeling confidence scores [53]. This is probably because the structure prediction algorithms were not designed/trained for this purpose. More sophisticated algorithms, such as self-distillation and deep transfer

learning, are needed to take the advantage of AI-predicted structures.

Conclusions

With the accumulation of big biological data, the progress in deep learning algorithms, and advance in computer hardware, the last decade has witnessed breakthroughs in protein structure prediction. In this article, we reviewed the representative protein structure prediction methods in the past two years. Specifically, we classify existing methods into two groups, i.e., the two-step approach and the end-to-end approach. The end-to-end approach AlphaFold2 has substantially increased the accuracy of protein structure prediction. Based on the experimental assessment of the trRosettaX and its improved version trRosettaX2, we show that the two-step approach may have the potential to achieve comparable accuracy to AlphaFold2, but with a significantly lower requirement of computing resources. We conclude that it is valuable to keep developing both groups of approaches. We point out a few outstanding challenges in function-orientated protein structure prediction, which may be the direction for future development.

Conflict of interest statement

Nothing declared.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (NSFC T2225007, T2222012, 11871290, 61873185, and 61932018).

References

Papers of particular interest, published within the period of review, have been highlighted as:

- * of special interest
- ** of outstanding interest

1. Dill KA, MacCallum JL: **The protein-folding problem, 50 years on.** *Science* 2012, **338**:1042–1046.
2. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, **Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, et al.: Highly accurate protein structure prediction with AlphaFold.** *Nature* 2021, **596**:583–589.
This paper describes the AlphaFold2 methodology in detail. This is the first working method for end-to-end protein structure prediction with high accuracy. AlphaFold2 represents a milestone in the field.
3. Callaway E: **'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures.** *Nature* 2020, **588**:203–204.
4. He K, Zhang X, Ren S, Sun J: **Deep residual learning for image recognition.** In *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2016:770–778.
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I: **Attention is all you need.** *Adv Neural Inf Process Syst* 2017:30.
6. Pearce R, Zhang Y: **Toward the solution of the protein structure prediction problem.** *J Biol Chem* 2021, **297**:100870.

7. Pearce R, Zhang Y: **Deep learning techniques have significantly impacted protein structure prediction and protein design.** *Curr Opin Struct Biol* 2021, **68**:194–207.
8. Remmert M, Biegert A, Hauser A, Söding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.** *Nat Methods* 2012, **9**:173–175.
9. Steinegger M, Söding J: **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.** *Nat Biotechnol* 2017, **35**:1026–1028.
10. Xu J: **Distance-based protein folding powered by deep learning.** *Proc Natl Acad Sci USA* 2019, **116**:16856.
11. Wang S, Sun S, Li Z, Zhang R, Xu J: **Accurate de novo prediction of protein contact map by ultra-deep learning model.** *PLoS Comput Biol* 2017, **13**:e1005324.
12. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Židek A, Nelson AWR, Bridgland A, et al.: **Improved protein structure prediction using potentials from deep learning.** *Nature* 2020, **577**:706–710.
This paper describes the first version of AlphaFold. It belongs to the two-step approach. It is the champion in CASP13, demonstrating the power of deep learning in increasing the accuracy of structure prediction. However, it is a pity that no working source codes or web server was available for this method.
13. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D: **Improved protein structure prediction using predicted inter-residue orientations.** *Proc Natl Acad Sci USA* 2020, **117**:1496.
This paper describes the trRosetta method for de novo structure prediction. With both inter-residue distance and orientations, trRosetta generates more accurate predictions than other state-of-the-art methods, including AlphaFold. Both source codes or web server are provided for trRosetta. It is a very useful platform for structure prediction due to its fast speed and high accuracy.
14. Du Z, Su H, Wang W, Ye L, Wei H, Peng Z, Anishchenko I, Baker D, Yang J: **The trRosetta server for fast and accurate protein structure prediction.** *Nat Protoc* 2021, **16**:5634–5651.
15. Su H, Wang W, Du Z, Peng Z, Gao SH, Cheng MM, Yang J: **Improved protein structure prediction using a new multi-scale network and homologous templates.** *Adv Sci* 2021, **8**:e2102592.
This paper describes an improved version of trRosetta (trRosettaX) by using a new network and the automated inclusion of homologous templates.
16. Ju F, Zhu J, Shao B, Kong L, Liu TY, Zheng WM, Bu D: **CopulaNet: learning residue co-evolution directly from multiple sequence alignment for protein structure prediction.** *Nat Commun* 2021, **12**:2535.
This paper proposes a new method to predict inter-residue distance from MSA directly rather than from MSA-derived co-evolution features.
17. Shen T, Wu J, Lan H, Zheng L, Pei J, Wang S, Liu W, Huang J: **When homologous sequences meet structural decoys: accurate contact prediction by tFold in CASP14-(tFold for CASP14 contact prediction).** *Proteins* 2021, **89**:1901–1910.
18. Greener JG, Kandathil SM, Jones DT: **Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints.** *Nat Commun* 2019, **10**:3977.
19. Zheng W, Li Y, Zhang C, Zhou X, Pearce R, Bell EW, Huang X, Zhang Y: **Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14.** *Proteins: Struct, Funct, Bioinf* 2021, **89**:1734–1751.
20. Hou J, Wu T, Guo Z, Quadir F, Cheng J: **The MULTICOM protein structure prediction server empowered by deep learning and contact distance prediction.** In *Protein structure prediction*. Edited by Kihara D, US: Springer; 2020:13–26. 10.1007/978-1-0716-0708-4_2.
21. Wu Q, Peng Z, Anishchenko I, Cong Q, Baker D, Yang J: **Protein contact prediction using metagenome sequence data and residual neural networks.** *Bioinformatics* 2020, **36**:41–48.
22. Brunger AT: **Version 1.2 of the crystallography and NMR system.** *Nat Protoc* 2007, **2**:2728–2733.
23. Swendsen RH, Wang JS: **Replica Monte Carlo simulation of spin glasses.** *Phys Rev Lett* 1986, **57**:2607–2609.

24. Liu DC, Nosedal J: **On the limited memory BFGS method for large scale optimization.** *Math Program* 1989, **45**:503–528.
25. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD: **Accurate prediction of protein structures and interactions using a three-track neural network.** *Science* 2021, **373**:871–876.
- This paper describes the RoseTTAFold method, which was inspired by AlphaFold2. A three-track network was proposed to improve the accuracy. It has two versions: the pyRosetta version (similar to trRosetta) and the end-to-end version. SE(3)-transformer was used to build 3D structure in the end-to-end version.
26. Xu J, McPartlon M, Li J: **Improved protein structure prediction by deep learning irrespective of co-evolution information.** *Nat Mach Intell* 2021, **3**:601–609.
27. AlQuraishi M: **End-to-End differentiable learning of protein structure.** *Cell Syst* 2019, **8**:292–301. e293.
28. Mirdita M, Schutze K, Moriawaki Y, Heo L, Ovchinnikov S, Steinegger M: **ColabFold: making protein folding accessible to all.** *Nat Methods* 2022, **19**:679–682.
29. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Zidek A, Bridgland A, Cowie A, Meyer C, Laydon A, *et al.*: **Highly accurate protein structure prediction for the human proteome.** *Nature* 2021, **596**:590–596.
- This paper describes the application of AlphaFold2 for the human proteome. It predicts more than 23,000 structures for the human proteome. 58% of them are predicted with high accuracy (pLDDT>70). This application increases the impact of AlphaFold2 further.
30. Kandathil SM, Greener JG, Lau AM, Jones DT: **Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins.** *Proc Natl Acad Sci U S A* 2022:119.
31. Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, Bafna K, Norm C, Kang A, Bera AK, *et al.*: **De novo protein design by deep network hallucination.** *Nature* 2021, **600**:547–552.
32. Moore PB, Hendrickson WA, Henderson R, Brunger AT: **The protein-folding problem: not yet solved.** *Science* 2022, **375**:507.
33. Jones DT, Thornton JM: **The impact of AlphaFold2 one year on.** *Nat Methods* 2022, **19**:15–20.
34. Chowdhury R, Bouatta N, Biswas S, Rochereau C, Church GM, Sorger PK, AlQuraishi M: **Single-sequence protein structure prediction using language models from deep learning.** *bioRxiv* 2021, 10.1101/2021.08.02.454840:2021.2008.2002.454840.
35. Singh J, Litfin T, Singh J, Paliwal K, Zhou Y: **SPOT-Contact-LM: improving single-sequence-based prediction of protein contact map using a transformer language model.** *Bioinformatics* 2022, **38**:1888–1894.
36. Wang W, Peng Z, Yang J: **Single-sequence protein structure prediction using supervised transformer protein language models.** *bioRxiv* 2022, 10.1101/2022.01.15.476476:2022.2001.2015.476476.
37. Wu R, Ding F, Wang R, Shen R, Zhang X, Luo S, Su C, Wu Z, Xie Q, Berger B, Ma J, Peng J: **High-resolution de novo structure prediction from primary sequence.** *bioRxiv* 2022, <https://doi.org/10.1101/2022.07.21.500999>.
38. Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A: **Transformer protein language models are unsupervised structure learners.** *bioRxiv* 2020, <https://doi.org/10.1101/2020.12.15.422761>.
39. Englander SW, Mayne L: **The nature of protein folding pathways.** *Proc Natl Acad Sci U S A* 2014, **111**:15873–15880.
40. Outeiral C, Nissley DA, Deane CM: **Current structure predictors are not learning the physics of protein folding.** *Bioinformatics* 2022, **38**:1881–1887.
41. Mullard A: **What does AlphaFold mean for drug discovery?** *Nat Rev Drug Discov* 2021.
42. Yan Y, Tao H, He J, Huang SY: **The HDock server for integrated protein-protein docking.** *Nat Protoc* 2020, **15**:1829–1852.
43. Kozakov D, Hall DR, Xia B, Porter KA, Padhorna D, Yueh C, Beglov D, Vajda S: **The ClusPro web server for protein-protein docking.** *Nat Protoc* 2017, **12**:255–278.
44. Baek M, Anishchenko I, Park H, Humphreys IR, Baker D: **Protein oligomer modeling guided by predicted interchain contacts in CASP14.** *Proteins* 2021, **89**:1824–1833.
45. Bryant P, Pozzati G, Elofsson A: **Improved prediction of protein-protein interactions using AlphaFold2.** *Nat Commun* 2022, **13**:1265.
46. Pozzati G, Zhu W, Bassot C, Lamb J, Kundrotas P, Elofsson A: **Limits and potential of combined folding and docking.** *Bioinformatics* 2021. 10.1093/bioinformatics/btab760.
47. Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J, Ness TJ, Banjade S, Bagde SR, *et al.*: **Computed structures of core eukaryotic protein complexes.** *Science* 2021, **374**. eabm4805.
48. Evans R, O'Neill M, Pritzel A, Antropova N, Senior AW, Green T, Zidek A, Bates R, Blackwell S, Yim J: **Protein complex prediction with AlphaFold-Multimer.** *bioRxiv* 2021.
- This paper describes the AlphaFold-Multimer for multimer structure prediction. It has a similar architecture to AlphaFold2. Benchmark tests show that it outperforms other methods significantly.
49. Gao M, Nakajima An D, Parks JM: **Skolnick J: AF2Complex predicts direct physical interactions in multimeric proteins with deep learning.** *Nat Commun* 2022, **13**:1744.
50. Bryant P, Pozzati G, Zhu W, Shenoy A, Kundrotas P, Elofsson A: **Predicting the structure of large protein complexes using alphafold and sequential assembly.** *bioRxiv* 2022.
51. Fontana P, Dong Y, Pi X, Tong AB, Hecksel CW, Wang L, Fu TM, Bustamante C, Wu H: **Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold.** *Science* 2022, **376**. eabm9326.
52. Gelman S, Fahlberg SA, Heinzelman P, Romero PA, Gitter A: **Neural networks to learn protein sequence-function relationships from deep mutational scanning data.** *Proc Natl Acad Sci U S A* 2021:118.
53. Buel GR, Walters KJ: **Can AlphaFold2 predict the impact of missense mutations on structure?** *Nat Struct Mol Biol* 2022, **29**:1–2.