Analysis of Gradient Descent Based System of AlphaFold's Protein Structure Prediction

Arunima Mandal, Fatema Tabassum Liza
Department of Computer Science
Florida State University
Emails: {am19cf, fl19g}@my.fsu.edu

***Abstract*—The primary structure of a protein is its amino acid sequence which drives the folding and intra-molecular bonding of the linear amino acid chain which ultimately determines the protein's unique three-dimensional shape. The three-dimensional shape of a protein can be determined by the Protein structure prediction from its amino acid sequence. This protein problem is of fundamental importance as the structure of the protein mostly determines its function. However, experimentally protein structures can be difficult to determine. In this report, a system has been described that was implemented and entered by the group A7D in CASP13. This particular system uses large genomic datasets to predict protein structure called AlphaFold. The 3D models of proteins that are generated by AlphaFold are far more accurate than anything before-marking significant progress on one of the core challenges in biology.**

***Index Terms*—Protein structure prediction, Distance prediction, AlphaFold, Deep learning, Machine Learning**

I. Introduction

The team focused especially on the problem of modelling target shapes from scratch (FM - Free Modelling), without using previously solved proteins as templates (TBM - Template Based Modelling). The submissions by A7D at CASP13 were made using three free-modeling(FM) methods which combine the predictions of the three neural networks. Among these methods, we studied mainly the third system which involves Gradient Descent as it shows more promising results than the other two. This proposed system contains the following steps 3 neural networks which predicts -

- Extracting MSA features from the sequence
- Feeding the features into three neural networks for - Distance histogram (Distogram) prediction, Torsion angle prediction and Background distogram
- Optimizing using Gradient Descent

After applying gradient descent repeatedly, finally the desired 3D coordinates of the protein structure is produced. Figure 1 depicts the overall process.

The efficiency of this AlphaFold highly depends on the distance prediction - where we have to train a neural network to make accurate predictions of the distances between pairs of residues, which conveys more information than contact predictions. Using this information, a potential of mean force has been constructed to accurately describe the shape of a protein. The neural network also predicts the torsion angles distribution. The resulting potential can be optimized by a simple gradient descent algorithm with respect to torsion angles to generate structures without complex sampling procedures. In the CASP13 FM assessors' ranking by summed z-scores, this system scored highest with 68.3 vs 48.2 for the next closest group (an average GDT_TD of 61.4).

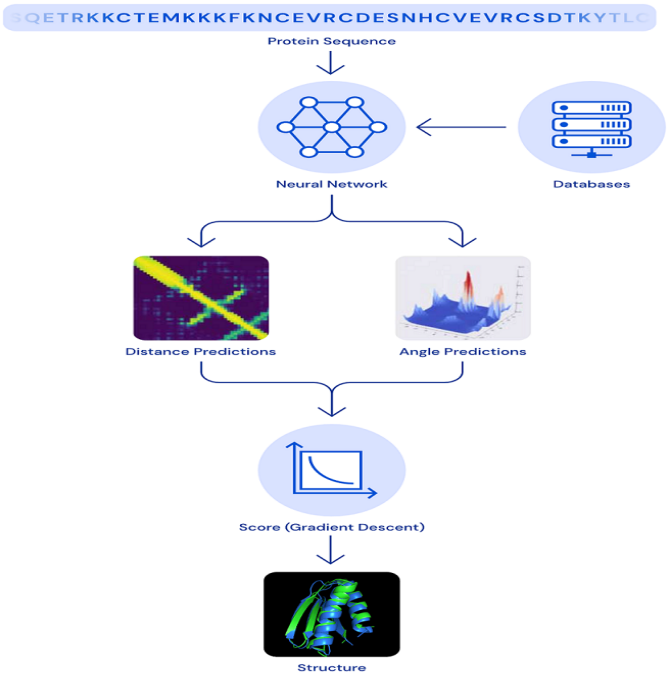

Fig. 1. Gradient Descent based AlphaFold process

II. Related Works

So far, the most-successful FM approaches [1]–[3] have relied on fragment assembly. In these approaches, a structure is created through a stochastic sampling process - such as simulated annealing [4] - that minimizes a statistical potential that is derived from summary statistics extracted from structures in the Protein Data Bank (PDB) [5]. In recent years, the accuracy of structure predictions has improved through the use of evolutionary covariation data [6] that are found in sets of related sequences. Several methods [7], [8], including neural networks [9], [10], have been used to predict the probability that a pair of residues is in contact based on features computed from MSAs. Other studies [11], [12] have used predictions of the distance between residues, particularly

for distance geometry approaches [13], [14]. Neural network distance predictions without covariation features were used to make the evolutionary pairwise distance dependent statistical potential[25], which was used to rank structure hypotheses.

The papers [9], [15] that we have based our project on have described the methodology and results of distance-based contact prediction, threading, and folding methods implemented on three RaptorX servers, which are built upon the powerful deep convolutional residual neural network(ResNet) method initiated for contact prediction.

III. Methodology

The first step of the approach is to extract MSA features from the given protein sequence. Then those features are fed into the 3 neural networks for predicting distance and torsion predictions. The team proposed 3 different systems. We will cover here only the Gradient descent based system instead of Simulated annealing and fragment assembly, which yielded the optimal solution [Figure 2].

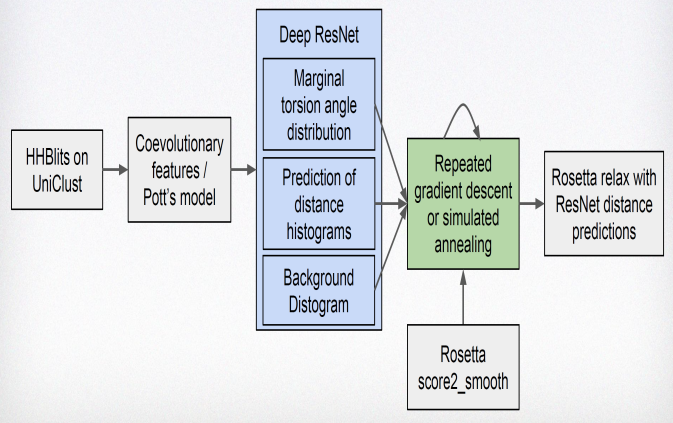

Fig. 2. The bird-eye view of the proposed approach

A. Distance histogram (Distogram) prediction

In this step, a neural network is trained to predict the distances d_{ij} between the β-carbon atoms of pairs of residues. While in some cases, these distances are highly constrained by secondary structure or clear co-evolutionary signals, in most cases, they will be uncertain. In order to model this uncertainty, the network predicts discrete probability distributions $P\Big(d_{ij}|S, MSA(S)\Big)$, given a sequence S and its multiple sequence alignment $MSA(S)$. These distance distributions are modeled with a softmax distribution for distances in the range 2 to 22 Å split into 64 equal bins. [16]

The architecture of the network is a deep two-dimensional dilated convolutional residual network. Previously, a two-dimensional residual network was used that was preceded by one-dimensional embedding layers for contact prediction [17]. The proposed network is two-dimensional throughout and uses 220 residual blocks with dilated convolutions [18]. Each residual block, consists of a sequence of neural network layers that interleave-

- three batchnorm layers
- two 1×1 projection layers
- a 3×3 dilated convolution layer
- and exponential linear unit (ELU)40 non-linearities

All the successive layers cycle through dilations of 1, 2, 4, 8 pixels to allow propagation of information quickly across the cropped region [Figure 3]. For the final layer, a position specific bias was used, such that the biases were indexed by residue-offset (capped at 32) and bin number. [16] The network is trained with stochastic gradient descent using a cross-entropy loss.

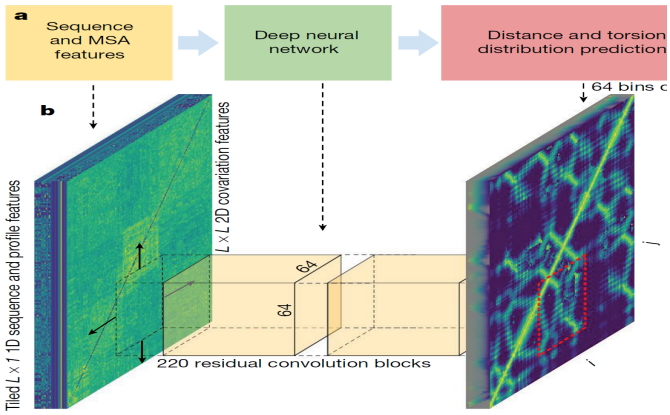

Fig. 3. The neural network predicts the entire $L \times L$ distogram based on MSA features, accumulating separate predictions for 64×64-residue regions.

B. Cropped distogram

To avoid overfitting and constrain memory usage, the network was always trained and tested on 64×64 regions of the distance matrix, that is, the pairwise distances between 64 consecutive residues and another group of 64 consecutive residues. The entire distance matrix [Figure 4] was split into non-overlapping 64×64 crops for each training domain. By training off-diagonal crops, the interaction between residues that are further apart than 64 residues could be modelled [16]. As it is known that contact prediction needs only a limited context window [19], each crop consisted of the distance matrix that represented the related positions of two 64-residue fragments. These regions are tiled together to produce distance predictions for the entire protein.

Randomizing the offset of the crops each time a domain is used in training leads to a form of data augmentation, which results into a thousands of different training examples from a single protein. To improve accuracy further, predictions from an ensemble of four separate models, trained independently with slightly different hyper-parameters, are averaged together.

C. Torsion angle prediction

To predict the marginal Ramachandran distributions, $P(\phi_i, \psi_i|S, MSA(S))$, the one-dimensional pooled activations are used independently for each residue, as a discrete probability distribution approximated to 10° (1,296 bins). In

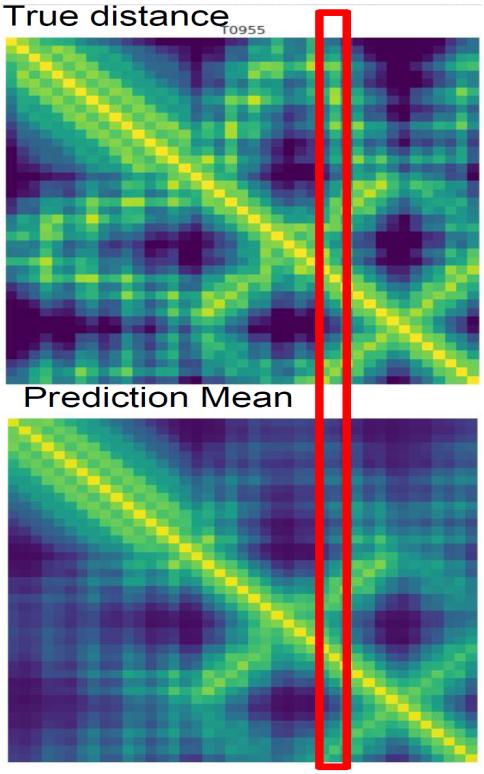

Fig. 4. Native vs predicted inter-residue distances for target T0955 and highlighting residue 29

practice during CASP13 the team used distograms from a network that was trained to predict distograms, secondary structure and ASA. Torsion predictions were taken from a second similar network trained to predict distograms, secondary structure, ASA and torsions, as the former had been more thoroughly validated. [16]

D. Distance Potential

In this step, a reference distribution has been predicted. For that, a similar model is trained on the same dataset. The reference distribution is not conditioned on the sequence, but to account for the atoms between which we are predicting distances, the team provided a binary feature $\delta_{\alpha\beta}$ to indicate whether the residue is a glycine (C_α atom) or not (C_β atom) and the overall length of the protein. A distance potential is created from the negative log likelihood of the distances, summed over all pairs of residues i, j. [15]

$$\mathrm{V}_{distance}(x) = -\sum_{i,j,i\neq j} logP\Big(d_{ij}|S, MSA(S)\Big)$$

With a reference state, this becomes the log-likelihood ratio of the distances under the full conditional model and under the background model. [15]

$$V_{distance}(x) = \sum_{i,j,i\neq j} -logP\Big(d_{ij}|S, MSA(S)\Big) - logP(d_{ij}|length)$$

Torsion distributions are modelled as a negative log likelihood under the predicted torsion distributions. As we have marginal distribution predictions, each of which can be multimodal, it can be difficult to jointly optimize the torsions. To unify all of the probability mass, at the cost of modelling fidelity of multimodal distributions, a unimodal von Mises distribution was fitted to the marginal predictions. This potential was summed over all residues i. [16]

Finally, to prevent steric clashes, a van der Waals term was introduced through the use of Rosetta's V_{score2_smooth}. Extended Data Figure 3c (top) shows the effect on the accuracy of the structure prediction of different terms in the potential.

E. Gradient Descent

Gradient descent finds the optimal energy positions for the folding of the proteins [Figure 5]. To realize structures that minimize the constructed potential, the group created a differentiable model of ideal protein backbone geometry, giving backbone atom coordinates as a function of the torsion angles $(\phi, \psi) : x = G(\phi, \psi)$. The complete potential to be minimized is then the sum of the distance, torsion and score2_smooth. As every term in V_{total} is differentiable with respect to the torsion angles, given an initial set of torsions ϕ, ψ, [16] which can be sampled from the predicted torsion marginals, we can minimize V_{total} using a gradient descent algorithm, such as L-BFGS [20]. The optimized structure is dependent on the initial conditions, so the optimization is repeated multiple times with different initializations. Repeated optimization (5000 repeats) from initial torsion angles sampled from predicted torsion angle distributions was found to converge quickly. [15]

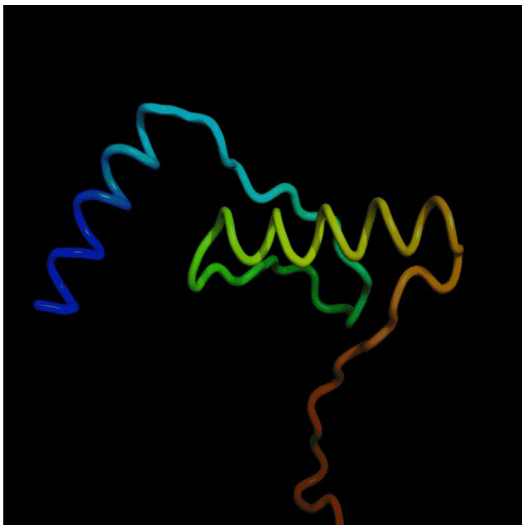

Fig. 5. Gradient Descent method predicting a 3D structure for target T1008

IV. Evaluation

The AlphaFold team followed the following described setup for running their code.

A. Experimental Setup

The following tools and dataset versions were used for the CASP system and for subsequent experiments-

- PDB 15 March 2018
- CATH 16 March 2018
- HHpred web server
- Uniclust30 2017-10
- SST web server(March 2019)
- BioPython v.1.65
- Rosetta v.3.5
- PyMol 2.2.0 for structure visualization
- TM-align 20160521
- HHblits based on v.3.0-beta.3 (three iterations, $E = 1 \times 10 - 3$)
- PSI-BLAST v.2.6.0 nr dataset (as of 15 December 2017) (three iterations, $E = 1 \times 10 - 3$)

B. Dataset

All the neural network models are trained on structures extracted from the PDB. They extracted non-redundant domains by utilizing the CATH7 352018-03-16). This gives 31,247 domains, which are split into train, and test sets (29,427 and 1,820 proteins respectively) keeping all domains from the same homologous super-family (H-level in the CATH classification) in the same partition. MSAs were generated using HHblits, 22 and from these profiles, 1D features were extracted. Potts models were fitted on the MSAs using pseudo-likelihood to generate 2D co-evolutionary features. PSIBLAST profiles were also used in the distance prediction network [16].

C. Implementation

We got the source code of AlphaFold implementation provided by the paper [16] and it is a free source code. We downloaded that and ran it. The weights and data file of the program is about 44 GB so we are unable to send that with the attachment of this submission. If convenient, kindly download the weights from the link http://bit.ly/alphafold-casp13-weights and the data from http://bit.ly/alphafold-casp13-data. These input files can be found from the link of the given source code https://github.com/deepmind/deepmind-research. Though there was definite directions to run the code, we faced difficulty in successful compilation of the code. The given scripts had to be modified in some places to run successfully. Then we tried to understand the implementation of the paper in the program and tried to relate the code associated with it. As for the coding part, we analyzed the inter-relations of the file dependencies and observed the internal flow of the implementation by tweaking some parts against the output. Also, we tried to improve the off-diagonal contact distances. We have to further understand the scope of this improvement and understand the dependencies of the code sections better to turn it to a successful attempt. We will continue our endeavour in the pursuit of this in future.

D. Result

The output we got from running the code, was the distograms and contact map (T1019s2.rr file). The outputs are in the Alphafold_output folder of our submission. In the **pasted** folder, the contact map (T1019s2.rr file) can be observed as in Figure 6:

```
PFRMAT RR
TARGET T1019s2
AUTHOR DM-ORIGAMI-TEAM
METHOD dm-contacts-resnet
MODEL 1
KVEPVGNAYGHWTKHGKEFPEYQNAKQYVDAAHNFMTNPPPGTLTKTRPNGDTLYYNPVTNVFASKDINGVPRTMFKPEKGIEYWNKQ
1 2 0 8 0.924252
1 3 0 8 0.699320
1 4 0 8 0.183992
1 5 0 8 0.127204
1 6 0 8 0.092013
1 7 0 8 0.107418
1 8 0 8 0.029173
1 9 0 8 0.038669
1 10 0 8 0.032726
1 11 0 8 0.025141
1 12 0 8 0.019511
1 13 0 8 0.020452
1 14 0 8 0.020101
1 15 0 8 0.017305
1 16 0 8 0.014301
1 17 0 8 0.010839
1 18 0 8 0.012423
```

Fig. 6. Contact map of target T1019s2

Data in the .rr file format are inserted between MODEL and END records of the submission file. The prediction should start with the sequence of the predicted target splitted (if necessary) in several rows. The sequence should be followed by the list of contacts in the five-column format:

i j d1 d2 p

The format description from the PDB:

- Indices i and j of the two residues in contact should be provided such that $i \leq j$, i.e. only half of the contact map is supplied.
- The numbers d1 and d2 indicate the distance limits defining a contact. In CASP, a pair of residues is defined to be in contact when the distance between their C-beta atoms (C-alpha in case of glycine) is less then 8 Angstroms. Therefore, typically $d_1 = 0$ and $d_2 = 8$. These parameters are currently dumb and left in the format only for the consistency with previous CASPs.
- The real number p indicates probability of the two residues being in contact, and should be in the range 0.0 - 1.0. Values larger than 0.5 identify the pairs of residues that are predicted to be more likely in contact than not. In binary (two-class) evaluations, the probability value of 0.5 will be considered as the cutoff separating contacts from non-contacts.
- Any pair NOT listed is assumed to be predicted as not in contact.

In CASP13, the A7D team's AlphaFold achieved a state-of-the art performance. Before T0975, two systems based on simulated annealing and fragment assembly (and using 40-bin distance distributions) were used. From T0975 onward, newly trained 64-bin distogram predictions were used and structures were generated by the gradient descent system described here (three independent runs) as well as one of the fragment assembly systems (five independent runs). As we only focused on the Gradient Descent based system, so the figure 7 illustrates - to which extent the Gradient descent based system outperforms the other two.

Method	Mean GDT_TS for targets	
	Before T0975	From T0975
Fragment assembly with GDT-net	63.8	N/A
Fragment assembly with distance potential	62.4	63.4
Gradient descent on distance potential	N/A	64.4

Fig. 7. A7D CASP13 accuracies by method. Average GDT_TS scores of the A7D CASP13 submissions broken down by method. Since the methods used changed after T0975, we show the means for these two sets separately. Domains in which only one method was used have been excluded to make the numbers comparable.

Figure 8 shows the FM performance of the A7D system, showing the number of FM (FM + TBM/FM) domains (out of 43) solved to a given GDT_TS accuracy, which shows that the A7D system is particularly adept at producing 50-70 GDT_TS structures.

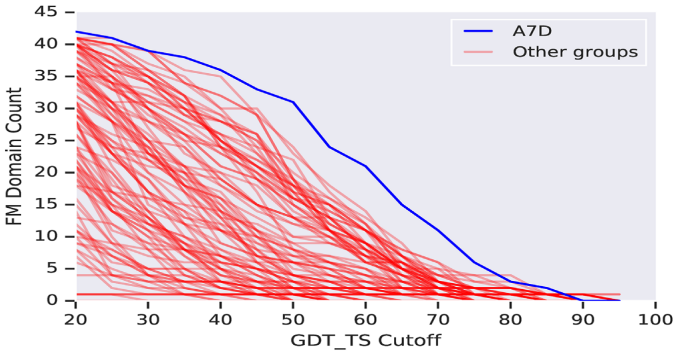

Fig. 8. Number of FM + FM/TBM domains (out of 43) solved to a GDT_TS threshold for all groups in CASP13

Figure 9 shows that, as expected, the system produces more accurate structures when the multiple sequence alignments are deeper, because of the distance predictor's dependence on co-evolutionary information. Since the system does not search for templates, the performance on TBM targets is often worse than that for FM targets with similar N_{eff}. Performance on TBM targets with few alignments can be much worse than for systems which explicitly use templates (eg, T0973-D1 which was over 40 GDT_TS worse than the best submission). Interestingly, the low-alignment designed protein T0955-D1 was solved to high accuracy (GDT_TS 88.4) despite having no alignments, presumably because of its short length and because the design process ensured it had highly typical structure.

V. Conclusion

The proposed approach is certainly a breakthrough compared to the previous models in the protein structure prediction domain, but there are still scopes for improvements, as among

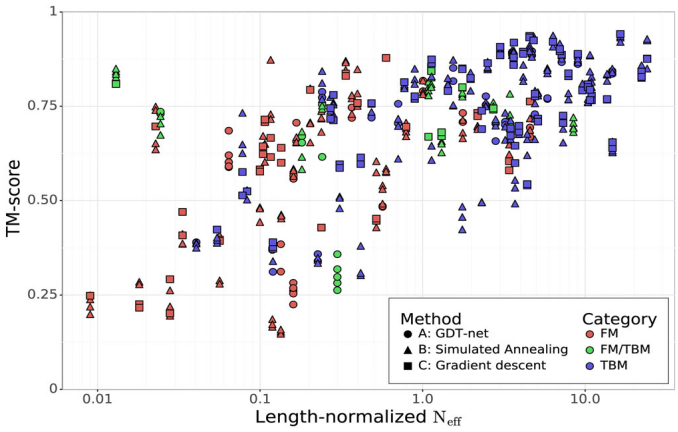

Fig. 9. The TM-score of the A7D submissions plotted against the length-normalized number of effective sequence alignments found (N_{eff}). Each domain decoy is colored by difficulty category, with a shape indicating the method by which it was generated

the 43 targets in CASP13, 25 were correctly predicted by A7D [16]. Improving the distogram prediction can result into further improvment in speed, accuracy or even in overall performance. The main weakness of the approach is that it still relies heavily on coevolution. When few alignments are available, distance predictions tend to be uninformative and poor structures are generated. Since there is no explicit template lookup, performance on TBM targets with few homologous sequences was much worse than template-based methods [16]. Also, the group did not attempt to propagate the uncertainty about distance into an uncertainty in residue positions [16]. Nevertheless, a tool like AlphaFold might help rare disease researchers predict the shape of a protein of interest rapidly and economically. As scientists acquire more knowledge about the shapes of proteins and how they operate through simulations and models, this method may eventually help us contribute to efficient drug discovery, while also reducing the costs associated with experimentation.

References

[1] R. Das and D. Baker, "Macromolecular modeling with rosetta," *Annu. Rev. Biochem.*, vol. 77, pp. 363–382, 2008.

[2] D. T. Jones, "Predicting novel protein folds by using fragfold," *Proteins: Structure, Function, and Bioinformatics*, vol. 45, no. S5, pp. 127–132, 2001.

[3] C. Zhang, S. Mortuza, B. He, Y. Wang, and Y. Zhang, "Template-based and free modeling of i-tasser and quark pipelines using predicted contact maps in casp12," *Proteins: Structure, Function, and Bioinformatics*, vol. 86, pp. 136–151, 2018.

[4] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[6] D. Altschuh, A. Lesk, A. Bloomer, and A. Klug, "Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus," *Journal of molecular biology*, vol. 193, no. 4, pp. 693–707, 1987.

[7] S. Ovchinnikov, H. Kamisetty, and D. Baker, "Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information," *Elife*, vol. 3, p. e02030, 2014.

[8] S. Seemayer, M. Gruber, and J. Söding, "Ccmpred—fast and precise prediction of protein residue–residue contacts from correlated mutations," *Bioinformatics*, vol. 30, no. 21, pp. 3128–3130, 2014.

[9] M. J. Skwark, D. Raimondi, M. Michel, and A. Elofsson, "Improved contact predictions using the recognition of protein like contact patterns," *PLoS computational biology*, vol. 10, no. 11, 2014.

[10] D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner, "Metapsicov: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins," *Bioinformatics*, vol. 31, no. 7, pp. 999–1006, 2015.

[11] A. Aszódi and W. R. Taylor, "Estimating polypeptideα-carbon distances from multiple sequence alignments," *Journal of mathematical chemistry*, vol. 17, no. 2, pp. 167–184, 1995.

[12] F. Zhao and J. Xu, "A position-specific distance-dependent statistical potential for protein structure and functional study," *Structure*, vol. 20, no. 6, pp. 1118–1126, 2012.

[13] J. Xu and S. Wang, "Analysis of distance-based protein structure prediction by deep learning in casp13," *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1069–1081, 2019.

[14] A. Aszodi, M. Gradwell, and W. Taylor, "Global fold determination from a small number of distance restraints," *Journal of molecular biology*, vol. 251, no. 2, pp. 308–326, 1995.

[15] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland *et al.*, "Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (casp13)," *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1141–1148, 2019.

[16] ——, "Improved protein structure prediction using potentials from deep learning," *Nature*, pp. 1–5, 2020.

[17] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate de novo prediction of protein contact map by ultra-deep learning model," *PLoS computational biology*, vol. 13, no. 1, p. e1005324, 2017.

[18] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[19] D. T. Jones and S. M. Kandathil, "High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features," *Bioinformatics*, vol. 34, no. 19, pp. 3308–3315, 2018.

[20] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.