A Review of Protein Structure Prediction using Deep Learning

Meredita Susanty[12*], *Tati Erawati* Rajab[2], and *Rukman* Hertadi[2]

[1]Universitas Pertamina, Jl Teuku Nyak Arief Simprug, Kebayoran Lama Jakarta Selatan 12220, Indonesia
[2]Institut Teknologi Bandung, Jl. Ganesa No.10 Lb. Siliwangi Coblong Bandung, Jawa Barat 40132, Indonesia

Abstract. Proteins are macromolecules composed of 20 types of amino acids in a specific order. Understanding how proteins fold is vital because its 3-dimensional structure determines the function of a protein. Prediction of protein structure based on amino acid strands and evolutionary information becomes the basis for other studies such as predicting the function, property or behaviour of a protein and modifying or designing new proteins to perform certain desired functions. Machine learning advances, particularly deep learning, are igniting a paradigm shift in scientific study. In this review, we summarize recent work in applying deep learning techniques to tackle problems in protein structural prediction. We discuss various deep learning approaches used to predict protein structure and future achievements and challenges. This review is expected to help provide perspectives on problems in biochemistry that can take advantage of the deep learning approach. Some of the unanswered challenges with current computational approaches are predicting the location and precision orientation of protein side chains, predicting protein interactions with DNA, RNA and other small molecules and predicting the structure of protein complexes.

1 Introduction

Proteins are an essential part of living things that trigger cells to perform different functions. Unlike DNA that never changes, a group of protein known as the proteome changes allows organisms to grow and develop. Only a few proteins work alone; most interact and form relationships with other proteins. The interactions between these proteins also change over time. Evolutionary relationships between proteins occur because organisms have to maintain certain functions as they evolve. The analogy is like a social network between humans. With about 10 billion protein molecules, a cell has a complex network of proteins. This protein network determines the health of a cell, which also affects the health of the organism in which the cell is located [1].

Proteins are linear chains of amino acids linked by covalent bonds. Amino acids are named in the form of a code with a length of 25 characters consisting of the alphabet. The rules for naming amino acids are 20 characters for standard amino acids, 2 for non-standard amino acids selenocysteine and pyrrolysine, 2 for ambiguous amino acids, and 1 for unknown amino acids [2], [3]. Apart from being encoded as a strand, proteins also have a 3D molecular structure. The various levels of protein are primary (chain of amino acids), secondary (local features), and tertiary (global features). Proteins are usually composed of several large domain proteins, strands of which are evolutionarily preserved and have well-defined folds and functions.

Knowledge of protein structure is fundamental to become the basis for other researches such as understanding certain diseases, developing new catalysts or developing drugs that are more effective and have lower side effects. Knowledge of protein structure provides an understanding of the function and workings of proteins, allowing researchers to influence, control, or modify proteins.

One way to find out the structure of a protein is to use an experimental approach, such as performing sequencing to determine the primary structure of a protein using mass spectrometry [1], or the use of technologies such as X-Ray diffraction crystallography and Nuclear Magnetic Resonance (NMR) Spectroscopy [4] to determine the tertiary structure. However, finding the tertiary structure of proteins through experiments requires a large amount of money and time. The computational approach becomes an alternative way to overcome this limitation.

The researchers undertook the initiative to hold a biennial event known as the Critical Assessment of Protein Structure Prediction (CASP) to track the development of protein structure prediction. CASP is a competition between research groups trying to predict how proteins fold. The protein structure used in CASP is a structure that has been measured experimentally but has not been published so that the competition participants do not know the 3D structure of the protein [5]. The participants' prediction results were compared with the actual protein form obtained from the experimental results, and the similarity was measured using the Global Distance Test (GDT) metric in percentage form. The deep learning approach to predict protein structure has made rapid progress in the last 2 CASP periods [6]. In previous years the GDT value in

* Corresponding author: meredita.susanty@universitaspertamina.ac.id

CASP did not reach 40%, but in the last 2 CASP periods (CASP 13 and CASP 14), one group (DeepMind) managed to achieve GDT extraordinary value of 60% in 2018 using the model known as AlphaFold. At CASP14, the group updated its model, AlphaFold2 and achieved a GDT of around 90%. However, some areas where AlphaFold2 does not perform well, namely in predicting protein complexes (oligomers) in which several amino acids interact. AlphaFold2 only predicts individual proteins, whereas many individual proteins combine to form protein complexes to function. AlphaFold2 has not been able to predict how proteins interact with DNA, RNA and small molecules and determine the exact location of side chains.

2 Protein Structure Prediction

In general, computational approaches are divided into two broad categories; (1) based on physical principles and (2) based on evolutionary principles as shown in Fig 1. Table 1 shows the development of protein structure prediction using various computational machine learning and deep learning methods for both categories. The physics-based approach simulates the folding process of the amino acid chain using molecular dynamics based on the potential energy of the force field in a particular time or fragment assembly using the energy function to form an energy stable 3D structure. However, molecular dynamics are only effective for small proteins, while fragment assembly has good accuracy if it has protein similarity information [7].

An approach based on evolutionary principles is based on the assumption that all living things came from a common ancestor and then evolved due to adaptation to the environment. In this adaptation, there is a change in the protein structure so that it can function optimally. When there is a change in structure, the essential amino acids do not change. Only the essential amino acid supporting amino acids change to suit the biophysical environment in which the protein must function, known as homology. However, this approach requires homologous sequence information so that it is difficult to determine the structure for a completely new protein. This approach is also difficult to investigate the impact of mutations on function.

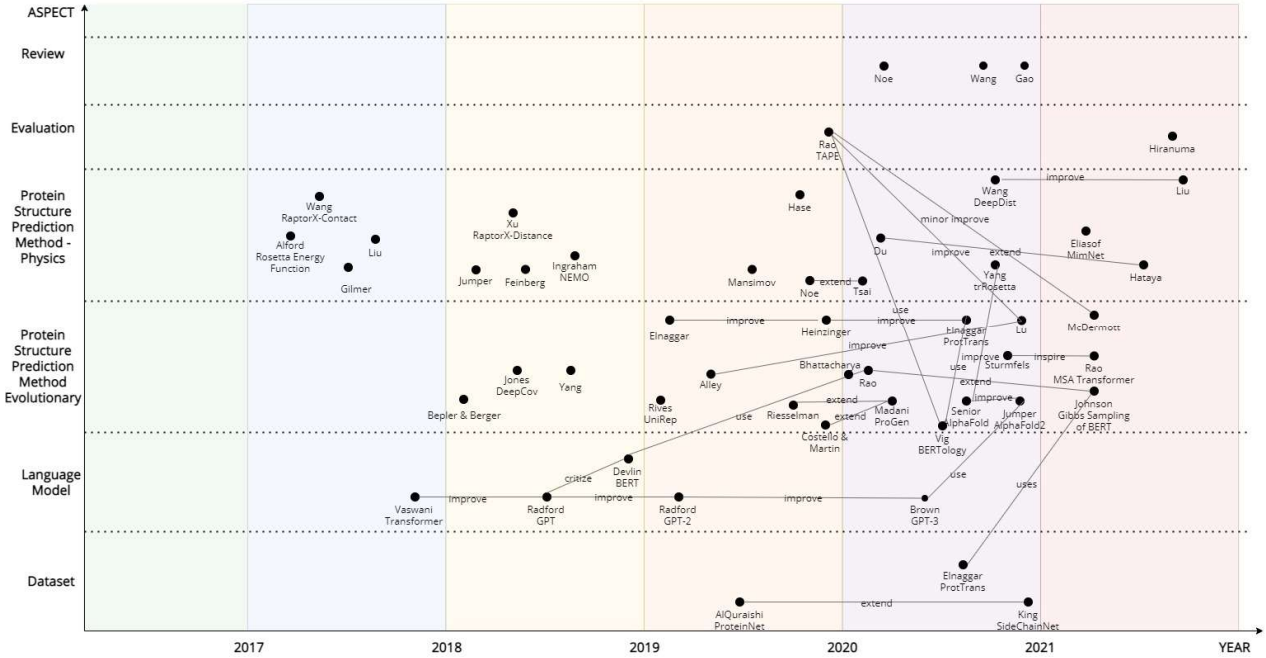

Fig. 1. Protein Structure Prediction using Computational Approach and Related Publication

2.1 Physic based Approach

Physics-based approaches to protein folding typically involve designing energy functions that guide protein dynamics in the conformation landscape from the unfolded to the folded state. Various approaches were carried out in the last few decades to design energy functions [8], [9] also using the first principle atomistic force field [10]–[16], which was then simplified using a coarse-grained approach protein modelling [17], [18]. In this context, artificial neural networks can help design energy functions to account for multibody terms that are not easy to model analytically.

An unsupervised approach to predicting the contact between residues is carried out by training the model on protein strands without any information about protein structure. The primary approach is to study the evolutionary boundaries between sets of similar protein strands using Markov Random Fields (Potts model) against the MSA underlying a protein strand, a technique known as Direct Coupling Analysis (DCA). Several studies have proposed using deep neural networks to replace shallow Markov Random Fields (MRF). Riesselman et al. trained the autoregressive model against MSA but ignored alignment and showed that protein function could be identified from unaligned strands [19]. In contrast to research by Rao et al. [20],

which used multiple MSA, Riesselman et al. [19] only used a set of related strands and did not use an end-to-end model to extract contact proteins.

Another model uses the Long Short Term Memory architecture with inputs in the form of amino acid sequences and PSSM and torsion angles to produce 3D structures [7]. The model built consists of 3 stages, computational, geometric and evaluation, by utilizing the results of Alquraishi's research [21] on the geometric stages.

Table 1. Computational Methods in Protein Prediction.

Contribution	Method
Protein representation	RNN [36], fine-tuned LSTM [47], BERT [35], ELMo [39], VAE [19]
Contact Prediction	Transformer [20],[37], factored attention [45], fine-tuned LSTM [47], CNN [42]
Profile Prediction	Transformer [51]
Distance Prediction	ResNet [59]
Secondary Structure Prediction	Fine-tuned LSTM[47], BERT[35]
Sidechain Dataset	ProteinNet + AMBER [57], orientasi β-carbon [28], CNN[56], Molekular Dinamik [8]
Tertiary Structure Prediction	Dilated Convolutional ResNet [25], Resnet [28], RNN [7], Markov Random Field [43], BERT [35], Transformer [23]
Model Interpretation	BERT, ALBERT, XLNet [48]
Model Comparison	Transformer-XL, XLNet, BERT, ALBERT [48], Transformer, LSTM, ResNet [49]
Accuracy Prediction	ResNet [60]

2.2 Evolutionary based Approach

Prediction of protein structure using a supervised approach using a deep neural network has made a breakthrough for predicting protein structure [22]–[24]. Early research using a supervised approach made use of co-evolutionary features [22], [25]–[29]. Furthermore, Multi Sequence Alignment (MSA) was used as input to predict protein structure using a supervised approach. [30], [31] studied a model that received MSA as input directly and then used 2D convolution [30] and Gated Recurrent Unit (GRU) [31] to process the input. A recent state-of-the-art protein prediction study, AlphaFold2 [23], used attention to process MSA-MSA using a supervised end-to-end model of protein structure.

Protein contact prediction has an important role in computational protein design [32]–[34] and is a key element of all currently state-of-the-art protein structure prediction methods [25], [26], [28]. Protein language modelling using an unsupervised approach has been carried out by several research groups [20], [35]–[39]. Some of these studies [20], [35], [37] used transformer architecture [40], [41]. Approaches using deep learning have been successful in predicting contact proteins [25], [26], [28], [29], [42] with inputs in the form of statistical covariance [29], [42], inferred coevolutionary parameters [25]–[28], sequences or evolutionary features [43]. Other studies have shown that the incorporation of co-evolutionary features is very important for model performance [44].

Since the advent of large-scale language models for natural language processing [40], [41], this approach has been used in other domains, namely protein structure prediction [35]–[37], [39], [45], [46]. The strands of amino acids that make up proteins are considered similar to the arrangement of words that make up sentences where a word can have a relationship with the word next to it and a word that is a bit far from a certain word position.

Previous research predicts protein contact using a language model using a supervised approach. Bepler and Berger combined unsupervised sequence pre-training with structural supervision to produce sequence embedding. They were the first to fine-tune a pre-trained model using a Long Short-Term Memory (LSTM) architecture on protein strands to pair contacts between residues [47]. [36], [39] showed that the LSTM language model was able to capture biological properties.

The first study to study protein structure using language model Transformer showed that information about the contact between residues could be recovered from the learned representation by performing a supervised linear projection of the protein structure [35]. Another study conducted an in-depth analysis of the self-attention mechanism in Transformers. It identified its relationship to relevant biological features and found that the various layers in the model contribute to the study of various features [48]. In particular, this study found a correlation between self-attention maps and contact patterns in proteins. Alley et al. and Heinzinger et al. performed a comparison of various protein language models using a deep residual network [35], [49]. Rao et al. [37] comparing the Potts model (trained against individual MSA) and Transformers (trained against a large string database). The result shows that just as the Potts model represents contacts directly through their pairwise components (weights), transformers also represent contacts directly via its pairwise component (self-attention). This study also shows the relationship between model performance, MSA depth and language model perplexity. Bhattacharya et al. [45] also showed that a single layer of self-attention could perform essentially the same computations as the Potts model. Based on the results of these two studies, the transport architecture is used to improve performance and perform sampling while maintaining contact [20].

Several other studies have explored alternatives to masked language modelling, such as the use of conditional generation [38], contrastive loss function

[50], and set of strands for supervision [51], [52]. Sturmfels et al. expanded the use of unsupervised language modelling to predict position-specific scoring matrix (PSSM) profiles [51]. Sercu et al. used amortized optimization to predict profile and pairwise coupling simultaneously [52].

Recently, a deep learning approach has been applied to perform MSA fittings. Heinzinger et al. showed that the factors studied by the Variational Autoencoder (VAE) model could be correlated with protein structure [39]. Smith et al. used the features of the Potts model with pseudolikelihood maximization to predict the pairwise distance with the deep residual network and optimize the final structure using Rosetta [14] [53].

3 Challenges

The deep learning approaches enable fast and accurate protein structure prediction. It can produce an individual protein structure in a couple of days compared to the experimental approach that usually takes months or years. However, several things that can be predicted using an experimental approach has not been successfully achieved using a computational approach.

3.1 Shallow MSA

Deep learning models are trained using the publicly available data consisting of hundreds of thousand protein structures that have been experimentally generated. The evolutionary-based approach uses MSA as an input. Structure prediction from very shallow MSA and even a single amino acid sequence remains a fundamental challenge.

3.2 Model Interpretation

A neural network is a flexible and powerful regression model. Furthermore, because of their highly recursive structure, neural networks are frequently referred to as "black boxes," meaning that the resulting parameters and functions are too complex for practitioners to comprehend. Although Vig et al. give reliable interpretations of Transformer architecture, particularly BERT, current deep learning models offer a limited understanding of the complex patterns they learn.

Reverse the process and works backwards to understand what information it was using was important as that might give some information into how the folding mechanism works within a cell. Because understanding the folding mechanism is one of the aspects that goes along with trying to crack the protein folding problem. If this could be understood, it may also then be possible to create a new protein structure and then to use the reverse mechanism to see what the original protein sequence would be.

3.3 Side Chain Location

Generally, the approach taken is to predict the protein backbone alone and then embeds the side-chain to the backbone structure. The side-chain conformation is predicted using conformation or energy-minimizing search [54]. Another approach uses Rosetta [53], optimizing the structure using a backbone-dependent rotamer library [55]. In both cases, the placement of the side-chain depends on the predicted backbone structure.

Research that only predicts the side-chain structure generally uses a physics-based approach, especially the energy function. However, Liu et al. tried to make predictions using deep neural network architecture without physics-based assumptions [56]. Several studies have attempted to combine backbone and side-chain [28], [57]. Yang et al. include a representation of the inter-residue orientation of the beta carbon to predict the structural features that help locate the position of side-chain atoms as well but does not provide a complete representation of the side-chain structure [28]. SidechainNet [57] generated a new dataset that extends the Protein-Net dataset [58]. This dataset includes atomic angle and coordinate information capable of describing all heavy atoms of each protein structure.

3.4 Protein Quaternary Structure

The current state of the art in protein structure prediction can predict the tertiary structure of individual proteins accurately. However, because many proteins act in a cell as a complex (a form of quaternary structure), how one protein is sandwiched amongst other proteins and knowing a full pitch of the complete complex provides far more information than just that protein individually. While experimental approaches can predict protein complex information, computational approaches have not shown an accurate prediction.

3.5 Protein Interaction

Protein does not work alone. Protein might include DNA in its structure. How the protein interacts with the DNA, RNA or other small molecules can only be predicted using an experimental approach. The existing deep learning models only predict protein structure based on one individual amino acid sequence.

4 Conclusion

We have discussed the current state-of-the-art deep learning techniques applied to the problem of protein structure prediction. Having protein structure gives us insight into the protein's function, and that just useful for understanding what cells doing and understanding what's going on within each of the cells. There are a variety of things we can get with a protein structure. Firstly, it gives an understanding of interactions, for example, with different drugs. So this is very useful for drug discovery. Secondly, a lot of diseases are caused by a genetic mutation within different genes, and those genetic mutations can often alter the amino acid

presence within a protein. Having this predictive system helps researchers see how the protein changes when that amino acid is changed and try to understand that disease mechanism. Third, diseases such as Alzheimer and Diabetes is often seen that protein aggregated. Understanding the aggregation process can be greatly aided by having the structures of these different proteins. However, the recent achievement in protein structure prediction can only predict individual protein backbone structure. AlphaFold2 is one of the most significant advancements to date, yet there are still many questions to be answered, as with all scientific research. There are still many challenges to explore, such as how to predict structure from shallow MSA, how numerous proteins form complexes, how proteins interact with DNA, RNA, or tiny molecules, how to precisely locate all amino acid side chains and how to reverse engineer the process.

This work was supported by Indonesia Endowment Fund for Education (LPDP), Ministry of Finance, Republic of Indonesia.

References

1. "HUPO - What is Proteomics?" [Online]. Available: https://www.hupo.org/Whats-Proteomics. [Accessed: 24-Feb-2021].
2. A. Bateman, "UniProt: A worldwide hub of protein knowledge," Nucleic Acids Res., vol. 47, no. D1, pp. D506–D515, Jan. 2019.
3. "Nomenclature and Symbolism for Amino Acids and Peptides: Recommendations 1983," Eur. J. Biochem., vol. 138, no. 1, pp. 9–37, 1984.
4. D. Sakakibara et al., "Protein structure determination in living cells by in-cell NMR spectroscopy," Nature, vol. 458, no. 7234, pp. 102–105, Mar. 2009.
5. "Home - CASP14." [Online]. Available: https://predictioncenter.org/casp14/index.cgi. [Accessed: 15-Apr-2021].
6. The Critical Assessment of protein Structure Prediction, "Artificial intelligence solution to a 50-year-old science challenge could 'revolutionise' medical research," Press Release, 30-Nov-2020. [Online]. Available: https://predictioncenter.org/casp14/doc/CASP14_press_release.html. [Accessed: 15-Apr-2021].
7. M. AlQuraishi, "End-to-end differentiable learning of protein structure," bioRxiv. bioRxiv, p. 265231, 14-Feb-2018.
8. J. M. Jumper, N. F. Faruk, K. F. Freed, and T. R. Sosnick, "Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics," PLoS Comput. Biol., vol. 14, no. 12, p. e1006342, Dec. 2018.
9. T. Lazaridis and M. Karplus, "Effective energy functions for protein structure prediction," Current Opinion in Structural Biology, vol. 10, no. 2. Current Biology Ltd, pp. 139–145, 01-Apr-2000.
10. K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet - a deep learning architecture for molecules and materials," J. Chem. Phys., vol. 148, no. 24, Dec. 2017.
11. J. Chen, J. Chen, G. Pinamonti, and C. Clementi, "Learning Effective Molecular Models from Experimental Observables," J. Chem. Theory Comput., vol. 14, no. 7, pp. 3849–3858, Jul. 2018.
12. S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K. R. Müller, "Machine learning of accurate energy-conserving molecular force fields," Sci. Adv., vol. 3, no. 5, p. e1603015, May 2017.
13. J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost," Chem. Sci., vol. 8, no. 4, pp. 3192–3203, Mar. 2017.
14. J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," J. Chem. Phys., vol. 148, no. 24, p. 241733, Jun. 2018.
15. J. Hermann, R. A. DiStasio, and A. Tkatchenko, "First-Principles Models for van der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications," Chemical Reviews, vol. 117, no. 6. American Chemical Society, pp. 4714–4758, 22-Mar-2017.
16. B. Nebgen et al., "Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks," J. Chem. Theory Comput., vol. 14, no. 9, pp. 4687–4698, Sep. 2018.
17. S. T. John and G. Csányi, "Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials," J. Phys. Chem. B, vol. 121, no. 48, pp. 10934–10949, Dec. 2017.
18. M. K. Scherer, B. E. Husic, M. Hoffmann, F. Paul, H. Wu, and F. Noé, "Variational selection of features for molecular kinetics," J. Chem. Phys., vol. 150, no. 19, p. 194108, May 2019.
19. A. J. Riesselman, J. B. Ingraham, and D. S. Marks, "Deep generative models of genetic variation capture the effects of mutations," Nat. Methods, vol. 15, no. 10, pp. 816–822, Oct. 2018.
20. R. Rao et al., "MSA Transformer," bioRxiv, p. 2021.02.12.430858, Feb. 2021.
21. M. AlQuraishi, "Parallelized Natural Extension Reference Frame: Parallelized Conversion from Internal to Cartesian Coordinates," J. Comput. Chem., vol. 40, no. 7, pp. 885–892, Mar. 2019.
22. A. W. Senior et al., "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)," Proteins Struct. Funct. Bioinforma., vol. 87, no. 12, pp. 1141–1148, Dec. 2019.
23. J. Jumper et al., "High Accuracy Protein Structure Prediction Using Deep Learning," 2020.

24. "AlphaFold: a solution to a 50-year-old grand challenge in biology | DeepMind." [Online]. Available: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology. [Accessed: 07-Apr-2021].
25. A. W. Senior et al., "Improved protein structure prediction using potentials from deep learning," Nature, vol. 577, no. 7792, pp. 706–710, Jan. 2020.
26. 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 Comput. Biol., vol. 13, no. 1, p. e1005324, Jan. 2017.
27. Y. Liu, P. Palmedo, Q. Ye, B. Berger, and J. Peng, "Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks," Cell Syst., vol. 6, no. 1, pp. 65-74.e3, Jan. 2018.
28. J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker, "Improved protein structure prediction using predicted inter-residue orientations," bioRxiv. bioRxiv, 18-Nov-2019.
29. B. Adhikari, "DEEPCON: Protein contact prediction using dilated convolutional neural networks with dropout," Bioinformatics, vol. 36, no. 2, pp. 470–477, Jan. 2020.
30. C. Mirabello and B. Wallner, "RAWMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments," PLoS One, vol. 14, no. 8, Aug. 2019.
31. S. M. Kandathil, J. G. Greener, A. M. Lau, and D. T. Jones, "Deep learning-based prediction of protein structure using learned representations of multiple sequence alignments," bioRxiv. bioRxiv, 27-Nov-2020.
32. W. Russ et al., "Evolution-based design of chorismate mutase enzymes," bioRxiv, p. 2020.04.01.020487, Apr. 2020.
33. P. Tian, J. M. Louis, J. L. Baber, A. Aniana, and R. B. Best, "Co-Evolutionary Fitness Landscapes for Sequence Design," Angew. Chemie - Int. Ed., vol. 57, no. 20, pp. 5674–5678, May 2018.
34. T. Blazejewski, H. I. Ho, and H. H. Wang, "Synthetic sequence entanglement augments stability and containment of genetic information in cells," Science (80-.)., vol. 365, no. 6453, pp. 595–598, Aug. 2019.
35. A. Rives et al., "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," bioRxiv. bioRxiv, p. 622803, 29-Apr-2019.
36. E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church, "Unified rational protein engineering with sequence-based deep representation learning," Nat. Methods, vol. 16, no. 12, pp. 1315–1322, Dec. 2019.
37. R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives, "Transformer protein language models are unsupervised structure learners," bioRxiv. bioRxiv, p. 2020.12.15.422761, 15-Dec-2020.
38. A. Madani et al., "ProGen: Language modeling for protein generation," bioRxiv. bioRxiv, p. 2020.03.07.982272, 08-Mar-2020.
39. M. Heinzinger et al., "Modeling the language of life - Deep learning protein sequences," bioRxiv. bioRxiv, p. 614313, 19-Apr-2019.
40. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, vol. 2017-December, pp. 5999–6009.
41. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 4171–4186, Oct. 2018.
42. 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, Oct. 2018.
43. J. Ingraham, A. Riesselman, C. Sander, D. Marks, and H. M. School, "Learning Protein Structure With A Differentiable Simulator," Sep. 2018.
44. J. Xu, M. McPartlon, and J. Li, "Improved protein structure prediction by deep learning irrespective of co-evolution information," bioRxiv. bioRxiv, 12-Oct-2020.
45. N. Bhattacharya et al., "Single layers of attention suffice to predict protein contacts," bioRxiv. bioRxiv, p. 2020.12.21.423882, 22-Dec-2020.
46. A. Elnaggar et al., "ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing," bioRxiv, Jul. 2020.
47. T. Bepler and B. Berger, "Learning protein sequence embeddings using information from structure," arXiv, Feb. 2019.
48. J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani, "BERTology Meets Biology: Interpreting Attention in Protein Language Models," bioRxiv, Jun. 2020.
49. R. Rao et al., "Evaluating Protein Transfer Learning with TAPE," bioRxiv, Jun. 2019.
50. A. X. Lu, A. X. Lu, and A. Moses, "Evolution Is All You Need: Phylogenetic Augmentation for Contrastive Learning," arXiv, Dec. 2020.
51. P. Sturmfels, J. Vig, A. Madani, and N. F. Rajani, "Profile Prediction: An Alignment-Based Pre-Training Task for Protein Sequence Models," arXiv, Nov. 2020.
52. T. Sercu et al., "Neural Potts Model | OpenReview," 2020, pp. 1–13.
53. S. Raman et al., "Structure prediction for CASP8 with all-atom refinement using Rosetta," Proteins

Struct. Funct. Bioinforma., vol. 77, no. SUPPL. 9, pp. 89–99, 2009.

54. M. S. I. Bhuyan and X. Gao, “A protein-dependent side-chain rotamer library.,” BMC Bioinformatics, vol. 12 Suppl 14, 2011.
55. M. V. Shapovalov and R. L. Dunbrack, “A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions,” Structure, vol. 19, no. 6, pp. 844–858, Jun. 2011.
56. K. Liu et al., “Prediction of amino acid side chain conformation using a deep neural network.”
57. J. E. King and D. Ryan Koes, “SidechainNet: An All-Atom Protein Structure Dataset for Machine Learning,” 2020.
58. M. AlQuraishi, “ProteinNet: A standardized data set for machine learning of protein structure,” BMC Bioinformatics, vol. 20, no. 1, p. 311, Jun. 2019.
59. T. Wu, Z. Guo, J. Hou, and J. Cheng, “DeepDist: real-value inter-residue distance prediction with deep residual convolutional network,” BMC Bioinformatics, vol. 22, no. 1, p. 30, Dec. 2021.
60. N. Hiranuma, H. Park, M. Baek, I. Anishchenko, J. Dauparas, and D. Baker, “Improved protein structure refinement guided by deep learning based accuracy estimation,” Nat. Commun., vol. 12, no. 1, p. 1340, Dec. 2021.