Applications of Deep Learning Algorithms in Protein Structure Prediction and Design

Yijie He *

College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen, China

* Corresponding Author Email: HeYijie123@outlook.com

Abstract. The three-dimensional configuration of a protein is important for its biological function, which is crucial in biological study and drug development. The progress of deep learning has greatly improved the prediction of protein structures. Especially, the AlphaFold series has increased the prediction precision and extended its use in the simulation of biomolecular complexes and drug discovery. This paper describes the latest progress in protein structure prediction and antibody design based on deep learning. The emphasis is on the enhancement and efficiency of the template-based method and the AlphaFold series. Furthermore, the practical applications of these methods in antibody engineering and drug discovery are illustrated by considering specific examples such as AbEgDiffuser for simultaneous sequence and structure design, and AlphaFold3 for the choice of covalent ligands. Experimental data indicate that deep learning can precisely forecast the protein structures and effectively improve the design and screening of potential substances, thereby offering new possibilities for rational drug design. In the future, more research should concentrate on predicting the conformational changes of proteins, designing novel antibody frameworks from scratch, and obtaining multiple drug-like properties at the same time.

Keywords: Deep Learning; Protein Structure Prediction; Antibody Design.

1. Introduction

Proteins play a main role in biological functions, and their actions are closely related to their three-dimensional structures. Therefore, it is important to study the protein structures in order to understand the biological mechanisms and to develop drugs rationally. Traditional experimental methods usually have long research periods, high expenses and low success rates. The appearance of deep learning has given a completely new way to solve these problems by significantly decreasing both the temporal and economic costs, and for the first time, enabling the rational design of antibodies which are suitable for particular targets. Thus, deep learning has greatly changed the strategy of antibody discovery and development.

Methods for predicting protein structures have changed from the template-based methods to deep generative models. The early techniques such as homology modelling and fold recognition depended mainly on the known structural databases and constructed models by sequence–structure alignment and energy evaluation. But these techniques usually failed in predicting the structures of proteins with low sequence similarity or new folding patterns. The hybrid methods, including the AlphaFold series, combine the information of templates, structural fragments, multiple sequence alignments and deep learning technologies to accomplish the complete prediction of protein structures with near-atomic accuracy.

AlphaFold2 uses the Evoformer architecture and Structure Module to integrate evolutionary information and spatial relationships and achieves excellent results in the CASP14 protein structure prediction contest. AlphaFold-Multimer also enhances these abilities for multi-chain protein complexes by incorporating special methods for multimer training and inter-chain optimization, which significantly improves the prediction of heteromeric assemblies. Recently, AlphaFold3 adopts a diffusion-based structure to produce atomic coordinates directly, thereby decreasing the reliance on multiple sequence alignments. Its integrated system can model proteins, nucleic acids, small molecules and post-translational modifications at the same time, showing outstanding performance

in forecasting complicated biomolecular assemblies. However, it still has some difficulties in simulating highly dynamic antibody-antigen interfaces.

Based on the progress in structure prediction, the deep generative models for antibody design have developed very quickly. AbEgDiffuser combines diffusion models with equivariant graph neural networks to achieve the simultaneous design of antigen-specific sequences and structures. Nevertheless, its present use is mainly limited to the optimization of complementarity-determining regions (CDRs) using existing antibody frameworks. Furthermore, the covalent drug virtual screening guided by AlphaFold3 has shown the capability to find potent Bruton's tyrosine kinase (BTK) inhibitors from over 900,000 compounds, indicating the new development from structure prediction to the discovery of functional molecules.

Although there have been these improvements, the present techniques still mainly concentrate on the prediction of static structures and localized molecular optimization. The problems of simulating conformational changes, allosteric influences and protein folding processes have not been solved yet. Additionally, the creation of completely new antibody frameworks, together with the thorough modification of drug-like properties, has not been accomplished fully.

This review gives an overall description of the recent advances in protein structure prediction using deep learning and antibody engineering. It describes the development of methods from template-based modelling to the AlphaFold series and current generative design systems, and discusses the principles, efficiencies and restrictions of the major models. Some examples of antibody sequence-structure joint design and covalent ligand screening are provided to illustrate the whole procedure from structure prediction to rational molecular design. In addition, the present problems are indicated and possible research directions like conformational dynamics prediction, de novo scaffold construction and multi-objective drug improvement are emphasized to provide a comprehensive guide for the improvement of next-generation protein therapeutic agents.

2. Deep Learning Paradigm for Protein Structure Prediction and Antibody Design

2.1. Fundamental Principles Underlying the Relationship Between Protein Structure and Function

Proteins play a crucial role in biological activities and are involved in various essential biological processes such as metabolism, energy conversion, signal transduction, gene regulation, immune responses and cell cycle control. The wide range of functions of proteins is due to their unique and stable three-dimensional configurations. By means of a complicated folding procedure, the amino acid sequence develops into specific secondary, tertiary and quaternary structures, thus forming functional sites for ligand recognition, substrate catalysis, signal transmission and ion transport.

From the viewpoint of drug discovery, structure-based drug design has become an important approach in current pharmaceutical research. It is necessary to have a clear knowledge of the spatial relationships between target proteins and the candidate compounds for the efficient selection, improvement and alteration of lead molecules. Therefore, both the experimental determination of protein structures by means of methods like X-ray crystallography and cryogenic electron microscopy (cryo-EM) and the computational prediction of protein conformations based on the amino acid sequences are of great importance in modern biological research and drug development. The investigation of the connection between protein structure and function is still one of the most basic and difficult goals in present life sciences [1].

2.2. Technical Bottlenecks of Traditional Experimental Approaches and the Paradigm Shift Driven by Deep Learning

Traditional methods used in antibody discovery show more and more disadvantages, mainly because of the long research period, large expenses and low success rates. In structural biology, the

conventional techniques, including X-ray crystallography and cryo-electron microscopy, can obtain detailed structural information; nevertheless, it usually takes several months or even years to collect high-quality antibody-antigen complex structures through repeated crystallization, data acquisition and model improvement. These procedures have strict conditions on sample purity, uniformity and stability, and are not capable of effectively presenting the conformational changes of antibodies during antigen-binding processes, thus preventing a thorough comprehension of the molecular recognition mechanisms.

In the process of antibody screening, the conventional methods like hybridoma technique and phage display need to build large antibody libraries, usually including 10^9 to 10^{11} clones, and require several rounds of selection based on affinity. These procedures are very time-consuming and laborious and are restricted by the diversity limit of natural or synthetic antibody repertoires [2]. It is known that the development of antibody therapies by using conventional techniques usually takes more than ten years and costs over one billion US dollars, and the success rate in obtaining regulatory approval is less than 10%. The above limitations originate from an "experiment first, optimize later" approach which is not efficient enough to explore the extensive sequence-structure-function space of proteins.

Recent developments in deep learning have brought about significant improvements for these problems. Large-scale protein language models, three-dimensional structure prediction systems like AlphaFold and RoseTTAFold [3], and generative models can acquire implicit physicochemical principles and structural representations from extensive sequence-structure databases directly. In comparison with the traditional screening techniques, the deep learning methods can evaluate billions of possible sequences simultaneously, thus shortening the screening period from several years to weeks or even days and greatly decreasing the experimental expenses.

Furthermore, deep learning has permitted the rational and controllable construction of antibodies. Instead of depending entirely on the naturally existing immune repertoire, researchers can now produce antibody molecules from scratch according to specified epitope characteristics and required binding properties [4]. This change in approach significantly improves the independence from experience and diversity of the conventional methods and is modifying the whole antibody discovery process rapidly.

2.3. Development of Protein Structure Prediction Methodologies

According to the use of known structural templates, protein structure prediction techniques can be divided into two categories [1]. One category is the template-based modelling (TBM) which builds three-dimensional models by detecting homologous proteins or structurally similar templates with known structures. When appropriate templates are accessible, this method usually gives high prediction accuracy. The other type does not depend on known structural templates; it forecasts the protein conformation from the amino acid sequence by means of physical energy functions or deep learning algorithms. Each category possesses its own advantages and disadvantages, and they serve as the primary methodological foundation for current protein structure prediction.

2.3.1. Template-Based Modeling

Template-based modeling includes two methods: homology modeling and threading (fold recognition). Homology modeling is based on the theory that protein tertiary structures change less rapidly than their amino acid sequences. Moreover, the types of protein folds are limited, thus enabling us to forecast the structure of an unknown protein from those with known structures that have similar folds.

The usual step is to examine protein structure databases to find appropriate template structures. The possible choices are judged by scoring functions which take into account the pair potentials of residues, the similarity of secondary structures, the solvent accessibility and other structural characteristics. Then, an alignment between the target sequence and the chosen template is done and the atomic coordinates of the target protein are rebuilt according to the template backbone. Finally, energy-based methods are used to check the structural practicability and enhance the model quality.

A representative tool is SWISS-MODEL [5], an initial web-based system for predicting protein structures. It contains a complete automatic process which consists of template selection, sequence alignment, model construction and structure evaluation, thereby simplifying the large-scale homology modelling.

The second method is the threading method, also called fold recognition. When the sequence similarity between the target protein and the available templates is too small, homology modeling cannot find trustworthy structural analogues. In this situation, the threading methods can be used.

Not only depending on the sequence similarity, but the threading algorithms also consider the match between a target sequence and the known structural patterns. This allows the detection of proteins having similar three-dimensional structures although there is little detectable sequence homology. By evaluating the sequence-structure fitness in a library of known patterns, the threading methods can predict possible structural models for difficult prediction objects. An example is GenTHREADER [6], which integrates the sequence profile analysis and the assessment of structural compatibility to enhance the fold recognition ability, especially for proteins not having close homologs in the present databases.

2.3.2. Hybrid Modeling Approaches and the AlphaFold Series

Hybrid modeling methods combine template information, structural fragments, energy functions and machine learning techniques to enhance the prediction accuracy and applicability. Among these, the models in the AlphaFold framework are the most representative achievements. Particularly, AlphaFold-Multimer is developed by the DeepMind team for predicting the structures of multimeric protein complexes, as an extension of AlphaFold [7]. Compared with AlphaFold2, this model has the ability to handle multi-chain training and inference, possesses better inter-chain information exchange mechanisms, uses redesigned loss functions and optimized cropping strategies, thus showing significant improvement in the prediction of protein complex structures. Benchmark tests have confirmed the effectiveness of AlphaFold-Multimer. On one hand, it achieves a successful interface prediction rate of 70% for heterodimers with no available structural templates in a dataset of 17 complexes; 26% of these predictions are considered as high-accuracy ones. For homodimeric complexes, the success rate reaches 72%, with 36% being high accurate predictions. Moreover, AlphaFold-Multimer introduces the interface predicted Template Modeling score (ipTM) as a confidence measure to evaluate the predicted protein-protein interfaces, giving structural biologists a powerful tool for high-accuracy modeling of protein complexes. Although it performs well, AlphaFold-Multimer still has some limitations. Specifically, its prediction ability for antibody-antigen complexes is relatively poor, due to the flexibility and complexity of the binding interfaces and the recognition mechanisms of the immune system [8].

Introduced in 2021, AlphaFold2 is an important progress in protein structure prediction. The main part of this system is the Evoformer network, which transforms the protein structure prediction into a three-dimensional geometric problem by considering both the evolutionary information and the spatial connections among amino acid residues [9]. This model is a two-stage integrated system. The protein sequence and multiple sequence alignment (MSA) are the input data, and the final result includes the three-dimensional coordinates of all heavy atoms. The whole structure is divided into two major parts: the Evoformer Module dealing with the evolutionary features from the input sequences, and the Structure Module constructing the three-dimensional protein structure through iterative procedures based on the acquired representations. Firstly, AlphaFold2 extracts the co-evolutionary tendencies of residues from a large database of protein sequences and then uses these evolutionary constraints to predict the spatial location of the residues. By incorporating the sequence information, evolutionary hints and geometric reasoning into one model, the performance reaches a high degree of structural prediction accuracy. The excellent results of AlphaFold2 can be seen in the 14th Critical Evaluation of Protein Structure Prediction (CASP14) contest. In 87 target protein domains, the median value of backbone Cα Root Mean Square Deviation (RMSD95) obtained by AlphaFold2 is 0.96 Å and the median all-atom RMSD is 1.5 Å. Compared with the second best method whose values are 2.8 Å and 3.5 Å respectively, these outcomes indicate that AlphaFold2 has

achieved nearly experimental and atomic level prediction precision and significantly outperforms the conventional structure prediction methods. However, there are some limitations of AlphaFold2. Its prediction ability mainly depends on the availability of homologous sequence data. When the number of effective sequences in the MSA is less than about thirty, the prediction accuracy declines obviously [10]. Besides, although the model works well for monomeric proteins and homomeric assemblies, it is inefficient in simulating heterogeneous protein complexes.

In 2024, DeepMind released AlphaFold3, an upgraded version of the AlphaFold system. Compared with its previous versions, AlphaFold3 applies a diffusion-based generative structure and decreases the reliance on multiple sequence alignments. Therefore, the entire network architecture is simplified while keeping the same high prediction accuracy. Instead of determining the rigid-body positions of the residues and then combining the atomic coordinates, AlphaFold3 forms the three-dimensional atomic structures directly through a diffusion process. Furthermore, the model makes use of the precise predictions made by AlphaFold-Multimer for knowledge distillation, which assists in controlling the diffusion model and reduces the possibility of unreasonable conformations, especially in disordered regions. One of the main advantages of AlphaFold3 is its complete structure which can deal with various types of molecular substances such as proteins, nucleic acids, small molecules and post-translational modified residues at the same time. This function allows the model to predict the molecular interactions in one computational environment.

AlphaFold3 shows excellent results in predicting biomolecular complexes [11]. For instance, in the complex consisting of a bacterial CRP/FNR-family transcriptional regulator, DNA and cyclic guanosine monophosphate (cGMP) (PDB ID: 7PZB), the predicted structure obtains an overall Local Distance Difference Test (LDDT) score of 82.8 and a Global Distance Test (GDT) score of 90.1. Likewise, for the human coronavirus OC43 spike protein–neutralizing antibody complex (PDB ID: 7PNM) which consists of 4,665 amino acid residues, AlphaFold3 achieves an LDDT score of 83.0 and a GDT score of 83.1.

Moreover, AlphaFold3 achieved the best scores in various benchmark datasets such as PoseBusters and the newly published Protein Data Bank (PDB) structures as well as RNA-related targets from CASP15. The performance was judged based on some criteria including the precision of ligand and covalent modification prediction, the percentage of binding-pocket ligand predictions with an RMSD less than 2 Å, LDDT values for protein-nucleic acid interfaces, LDDT values for single protein and nucleic acid structures, and the percentage of protein-protein and antibody-antigen interfaces with DockQ scores over 0.23. Statistical differences were calculated by two-sided Fisher's exact tests and Wilcoxon signed-rank tests, indicating that AlphaFold3 performed significantly better than other methods in all the tested criteria. In conclusion, these findings indicate that the model has a good ability to predict the structures of different and complex biomolecular complexes.

3. Applications of Deep Learning in Antibody Design and Drug Discovery

In recent years, deep learning has led to significant progress in antibody therapeutic research, especially in the fields of antibody sequence-structure co-design and covalent ligand virtual screening. Some typical studies have shown the capability of artificial intelligence to speed up the molecular design and drug discovery processes. In this part, the basic technique and present limitations of the AbEgDiffuser system will be introduced, then the application of AlphaFold3 to covalent drug virtual screening will be discussed and the experimental results confirming its efficiency will be examined.

3.1. Antibody Sequence–Structure Co-design Based on AbEgDiffuser

AbEgDiffuser, a deep generative model proposed in 2025, is used for the simultaneous design of antigen-specific antibody sequences and their three-dimensional structures [12]. It combines diffusion models with equivariant graph neural networks to optimize both the antibody sequences and the three-dimensional structures at the same time. Compared with the conventional methods which mainly concentrate on either the sequence optimization or the structural refinement separately,

AbEgDiffuser takes into account the mutual relationship between the amino acid sequences and the spatial conformations. Thus, it can produce antibody variations which keep the structural stability while showing enhanced antigen-binding properties.

The framework consists of two main components. The first component applies equivariant graph neural networks to acquire structural information and preserve its invariance under rotation and translation. The second component carries out a diffusion process to gradually enhance the representation of antibody sequences and structures. Through repeated denoising, the model generates antibody candidates which satisfy the specified structural and functional requirements. Experimental results showed that AbEgDiffuser could improve the design of complementarity-determining regions (CDRs), mainly involved in antigen recognition and binding specificity. In comparison with the existing methods of antibody design, the generated antibody candidates had higher structural feasibility and better predicted binding abilities. These findings indicate that combining geometric deep learning with diffusion-based generation is a promising approach for the next generation of antibody engineering.

The present method mainly concentrates on improving CDR regions together with the preservation of the original antibody frameworks. Therefore, the model is incapable of producing entirely new antibody structures from scratch efficiently. Moreover, experimental verification is required to confirm if the computed antibodies show the expected biological effects under normal conditions. AbEgDiffuser indicates the possibility of integrating deep learning and generative modeling in antibody sequence-structure co-design, which is an important progress towards the fully automatic discovery and improvement of antibodies.

3.2. AlphaFold3-Guided Covalent Drug Discovery

Although AlphaFold3 can predict the structures of covalent protein–ligand complexes correctly, its practical application in virtual screening has not been verified experimentally, and there was no previous benchmark dataset for testing the performance of covalent ligand screening. In order to fill this gap, researchers used the AlphaFold3 model developed by DeepMind for all-atom structural prediction and applied it for the first time to covalent ligand virtual screening. At the same time, they established a benchmark dataset called COValid, which is the first standardised dataset specially designed to evaluate the efficiency of covalent ligand virtual screening and to compare the performance of AlphaFold3 with that of conventional covalent docking methods [13]. The experimental results showed that when the mean Predicted Aligned Error (mPAE) is below 0.85 Å, about 96.6% of the predicted compounds are found to be effective ligands with measurable biological activity, indicating that AlphaFold3 performs better than the traditional covalent docking tools.

In the subsequent phase of the screening experiments, the researchers selected 13 new candidate molecules from a compound library which includes more than 900,000 substances and synthesized them for experimental verification. Finally, three efficient covalent ligands were found. Among these, a substance called YS1 has a novel chemical structure and is different in structure from the current Bruton's tyrosine kinase (BTK) inhibitors. The pharmacological properties of YS1 are very promising: it shows an IC50 value of 30 nM for BTK in biochemical tests and an IC50 value of 170 nM in cell-based tests. Additionally, the test on 362 kinases indicates only six off-target effects, indicating high selectivity.

AlphaFold3 has correctly predicted the structures of the YS1-BTK and YS2-BTK covalent complexes bound to Cys481 in the BTK kinase domain. The determined ligand conformations accord well with the experimental crystal structures, having heavy-atom root-mean-square deviation (RMSD) values of only 0.50 Å and 0.41 Å respectively. These results verify the accurate prediction capability of AlphaFold3 for covalent protein-ligand systems and indicate its possible application in artificial intelligence-assisted drug discovery.

4. Conclusion

This review gives an overall summary of the development of deep learning methods in protein structure prediction and antibody design. It mentions the improvement from the previous template-based model to the AlphaFold series of models and also examines the application of deep generative models in antibody engineering. By contrasting some representative approaches, the significant progress achieved by deep learning in enhancing the accuracy of structure prediction, modeling of biomolecular complexes and the investigation of biomolecular interactions is highlighted. Furthermore, the use of AbEgDiffuser for antibody co-design and the screening of covalent ligands assisted by AlphaFold3 indicates the present transition from structure prediction to rational molecular design.

Deep learning allows for the quick design and assessment of candidate molecules in considerably less time, thus offering new technical methods for antibody drug discovery and laying an important basis for achieving efficient and accurate pharmaceutical development in the future. However, there are still some difficulties to overcome. Presently, most models mainly concentrate on predicting static structures and have little ability to describe conformational changes and allosteric effects. Moreover, the self-generated creation of completely new molecular frameworks is still in an immature stage, and many computational predictions lack sufficient systematic experimental verification. In the future, the combination of molecular dynamics simulations with various kinds of biological data may be beneficial, which will help to improve the automation of the whole process from structure prediction to molecular creation.

References

[1] Bertoline L M F, Lima A N, Krieger J E, Teixeira S K. Before and after AlphaFold2: an overview of protein structure prediction[J]. Frontiers in Bioinformatics, 2023, 3: 1120370.

[2] Almagro J C, Daniels-Wells T R, Perez-Tapia S M, Penichet M L. Progress and challenges in the design and clinical development of antibodies for cancer therapy[J]. Frontiers in Immunology, 2018, 8: 1751.

[3] Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network[J]. Science, 2021, 373: 871–876. https://doi.org/10.1126/science.abj8754.

[4] Watson J L, Juergens D, Bennett N R, et al. De novo design of protein structure and function with RFdiffusion[J]. Nature, 2023, 620: 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.

[5] Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes[J]. Nucleic Acids Research, 2018, 46(W1): W296–W303.

[6] McGuffin L J, Jones D T. Improvement of the GenTHREADER method for genomic fold recognition[J]. Bioinformatics, 2003, 19(7): 874–881. https://doi.org/10.1093/bioinformatics/btg097.

[7] Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-Multimer[J]. bioRxiv, 2021: 2021.10.04.463034.

[8] Yin R, Feng B Y, Varshney A, Pierce B G. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants[J]. Protein Science, 2022, 31(8): e4379. https://doi.org/10.1002/pro.4379.

[9] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583–589.

[10] Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome[J]. Nature, 2021, 596: 590–596. https://doi.org/10.1038/s41586-021-03828-1.

[11] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3[J]. Nature, 2024, 630(8016): 493–500.

[12] Zhu Y, Shi X, Zhang J, Sun W, Wang L. AbEgDiffuser: antibody sequence–structure co-design with equivariant graph neural networks and diffusion models[J]. Journal of Chemical Theory and Computation, 2025, 21(21): 11307–11317. https://doi.org/10.1021/acs.jctc.5c00990.

[13] Shamir Y, Gabizon R, Rogel A, Lin D Y, Andreotti A H, London N. Discovery of covalent ligands with AlphaFold3[J]. Journal of the American Chemical Society, 2026. https://doi.org/10.1021/jacs.5c22222