Application Study of Three Mainstream Protein Structure Prediction Tools

Yifan Bai *

School of Biotechnology, East China University of Science and Technology, Shanghai, China

* Corresponding Author Email: 23012317@mail.ecust.edu.cn

Abstract. AlphaFold2, RoseTTAFold and ProteinGAN/ProtGPT2 are currently three widely used artificial intelligence tools for protein structure prediction and design. In this article, we make a comparative analysis of these three tools based on literature research and exemplary simulation. AlphaFold2 and RoseTTAFold use deep neural networks to predict three-dimensional structure of proteins from amino acid sequence. AlphaFold2 almost reaches the accuracy of experiment in single protein prediction, while RoseTTAFold uses three track architecture to model protein-protein complexes, and its accuracy is slightly lower. ProteinGAN and ProtGPT2 use Generative Adversarial Networks (GANs) and language models to generate new amino acid sequences. ProteinGAN and ProtGPT2 expand protein sequence and sometimes can fold into stable sequence. Their refers to the research results of protein generation tools, which have opened up new directions for the application of artificial intelligence in the field of protein engineering. In this article, we summarize the principle, performance characteristic and typical application of each method. The research results in this paper can provide a reference for selecting the appropriate tool in different situations. AlphaFold2 or RoseTTAFold can be used for accurate structural prediction in bioinformatics and structural biology, and ProtGPT2/ProteinGAN can be used for new design of protein in enzyme engineering and synthetic biology. By mastering the advantages of each tool, researchers can better use computational protein modeling to accelerate scientific innovation.

Keywords: Protein Structure Prediction; AlphaFold2; RoseTTAFold; ProteinGAN/ProtGPT2.

1. Introduction

All living systems contain proteins, and their three-dimensional structure determines their function at the level of the cell. Understanding a protein's structure can be essential to understanding its mechanism—from how an enzyme catalyzes a reaction to how a signaling protein mediates a cellular response. Yet experimental structure determination (e.g. X-ray crystallography, NMR, cryo-EM) is tedious and time-consuming: after decades of progress that has resulted in over 200,000 entries to the Protein Data Bank, there are still many more proteins without structures than with. This gap – the classic "protein folding problem" of relating amino acid sequence to 3D structure – has been a limiting step in molecular biology and drug discovery. There is a strong scientific and practical need for computational methods to accurately predict protein structures that would accelerate research by providing models of protein structure when experimental data are unavailable.

In the past few years, deep learning approaches have dramatically improved protein structure prediction. The protein structure prediction problem was recently solved by DeepMind's AlphaFold2, which in 2020 released an algorithm that demonstrated that AI could predict protein structures with atomic-level accuracy. In the CASP14 blind assessment, AlphaFold2's models were often essentially indistinguishable from experimental structures, leading to a consensus that the protein structure prediction problem for single chains has been solved. AlphaFold2's architecture, which includes a novel neural network design that integrates evolutionary sequence information, physics-inspired features, and attention-based learning, achieves a typical prediction error of the order of only ~1 Å. This accuracy allowed AlphaFold2 to predict structures for essentially all known protein sequences, resulting in the release of over 200 million protein models (covering entire proteomes and vastly extending structural coverage). AlphaFold's success represents a sea-change in computational biology with huge implications for understanding protein function and accelerating biotechnology research.

After solving this problem, AlphaFold's authors released source code and models, and other complementary tools have since emerged. Of note, the Baker lab released RoseTTAFold, which uses a three-track neural network to iteratively learn protein sequence, distance, and coordinate representations and predict protein structures. RoseTTAFold achieved accuracy approaching that of AlphaFold2 and was also able to model protein–protein interactions, enabling the prediction of multimeric complexes in addition to single chains. AlphaFold generally achieves higher accuracy for individual proteins, but RoseTTAFold provides a robust open-source alternative and achieves good performance in certain use cases, such as multi-chain assemblies. It is interesting and scientifically useful to study the differences in these models' architectures and prediction outputs, as this gives us a window into the principles of protein folding learned by AI and helps researchers choose appropriate tools for their specific tasks.

In parallel, a second generation of AI models address protein structure problems from a design perspective. Generative models such as ProteinGAN and ProtGPT2 seek to expand the protein sequence universe by generating new proteins in silico. ProteinGAN is a generative adversarial network trained on natural protein sequences that can generate new sequence variants. ProtGPT2 is a Transformer-based language model trained on millions of sequences; it generates de novo protein sequences that appear statistically consistent with natural proteins. The great majority of sequences generated by ProtGPT2 are predicted to fold into stable globular structures, and when exposed to structure prediction (e.g. AlphaFold) their resulting models are often well-folded proteins with topology not seen in existing protein databases. These AI design tools provide complementary tools to structure prediction methods and can help advance enzyme engineering and synthetic biology by generating proteins with potentially novel functions and folds.

This review compares AlphaFold2, RoseTTAFold, and ProteinGAN/ProtGPT2 through literature analysis and hands-on simulation. We study how each method works – from evolutionary couplings used in AlphaFold/RoseTTAFold to ProteinGAN and ProtGPT2 sequence-generation strategies – and discuss their typical performance and use cases. The core objective of this comparative review is to provide recommendations for researchers in the fields of bioinformatics, enzyme engineering, and synthetic biology on how to use these cutting-edge tools to promote the development of protein science and biotechnology.

2. Fundamental Logic of Protein Structure Prediction

The fundamental logic of protein structure prediction originates from the Anfinsen hypothesis, which states that the amino acid sequence of a protein determines its three-dimensional structure and that the native conformation corresponds to the state of lowest free energy [1]. Regardless of whether the method is traditional homology modeling, physics-based molecular dynamics, or modern deep learning approaches such as AlphaFold2 and RoseTTAFold, all follow this core principle.

The prediction process can be summarized in three hierarchical levels. Firstly, the amino acid sequence determines the structure. The chemical properties of amino acid residues (such as hydrophobicity, charge, and side-chain volume) define their potential spatial positions. Secondly, the structure tends toward energy minimization. Proteins fold into stable conformations through non-covalent interactions such as hydrogen bonds, hydrophobic interactions, van der Waals forces, and electrostatic attractions. Thirdly, evolutionary information provides constraints. Multiple sequence alignments (MSA) identify co-evolving residue pairs, allowing inference of their spatial contact relationships [2]. Modern deep learning methods build upon these principles by introducing neural networks that integrate sequence data, evolutionary coupling information, and geometric constraints, enabling a direct mapping from sequence to structure with significantly improved predictive accuracy.

3. AlphaFold2: Single-Protein Structure Prediction

3.1. Introduction and Features of AlphaFold2

AlphaFold2, an advanced model created by DeepMind, uses deep learning in conjunction with evolutionary coupling information extracted from multiple sequence alignments (MSA) to predict spatial distances and angles between amino acid residuesn [3]. Amino acids that tend to mutate in tandem in evolution are likely to be close to each other in three-dimensional space. AlphaFold2 also uses a Transformer-like attention mechanism [4], that models' long-range interactions between residues and iteratively refines inter-residue relationship matrices to output three-dimensional models with near-experimental accuracy [5].

AlphaFold2 is very accurate at predicting protein folding patterns and long-range interactions, so it is very suitable for single-protein prediction and stable-structure prediction. However, it is not very suitable for multi-protein complexes, membrane proteins and dynamic conformational changes prediction. It also relies on homologous sequences availability. Furthermore, AlphaFold2 is a "black-box" model, so it lacks biological interpretability, making it hard to understand its physical and chemical mechanisms [6].

3.2. Application Examples of AlphaFold2 in SARS-CoV-2 and GPCR Structure Modeling

AlphaFold2 was applied to predict the structure of several SARS-CoV-2 proteins, which is helpful for vaccine and drug development in the COVID-19 pandemic [7]. In the SARS-CoV-2 study, Yang et al. applied AlphaFold2 to model spike proteins of ten major variants. The structural superposition based on cryo-EM results verified the accuracy of AlphaFold2 in modeling the global conformational changes of RBD and NTD. For example, in the Delta variant, several significant loops in RBD moved outwards/insides and the antigenic loops in NTD were repositioned, which might be the mechanisms for immune escape. On the basis of the model, the authors also applied virtual screening and experimental screening of inhibitors. AlphaFold2 was successful in virtual screening of inhibitors.

In engineering of enzymes, due to the universality of folding pattern, the prediction of folding pattern of novel enzymes has laid the foundation for subsequent functional modification and synthetic verification. This is the routine application of AlphaFold2. In the prediction of G protein-coupled receptors (GPCRs), which are typical targets with high conformational flexibility and experimental difficulty, AlphaFold2 has potential. He et al. applied AlphaFold2 to model GPCRs with different receptor subtypes. The overall location of TMDs was predicted with high accuracy (RMSD ~2–3Å), while the extracellular domains (e.g., GLP1R's ECD) were less accurately located. The model showed limited ability to capture active/inactive state transitions for movements like TM6 outward shifts [8]. Moreover, side-chain positioning in the ligand binding pocket needs further improvement. All in all, AlphaFold2 gets expected results in static structure modeling, but still needs improvement in predicting functionally relevant conformations.

4. RoseTTAFold: Complex System Prediction

4.1. Introduction and Features of RoseTTAFold

RoseTTAFold by the Rosetta team at the University of Washington is very similar to AlphaFold2 except that it has a third three-track network that processes one-dimensional amino acid sequences, two-dimensional residue distance/contact maps, and three-dimensional coordinates [9]. These three levels are iteratively exchanged and integrated in information to enable the model to not only predict single-protein structures but also protein complexes and protein-ligand interactions.

RoseTTAFold shows improvements over AlphaFold2 on multi-protein complex and protein-ligand system predictions. It can be easily combined with the Rosetta modeling platform, which has more physical energy functions, to further refine the structure, making it suitable for system-level modeling. The single-protein prediction accuracy of RoseTTAFold is in general lower than that of

AlphaFold2, especially for highly conserved folds. RoseTTAFold is also more computationally expensive and thus requires more GPU resources as well as more user expertise. AlphaFold2 has better community support and ecosystem, which RoseTTAFold is still lacking. Therefore, its usability and scalability are still being addressed.

4.2. Application Examples of RoseTTAFold in GPCR and Multi-Enzyme System Modeling

RoseTTAFold shows promising results on modeling structurally heterogeneous and flexible protein systems, including transmembrane proteins and macromolecular complexes. Baek et al. used their three-track neural network to predict the structure of various human GPCRs, including several key activation-related conformations such as TM6 movement outward. The authors aligned the predicted structures and the cryo-EM structures and achieved RMSD around 2–4 Å. Thus, RoseTTAFold can also be used as a reliable tool for static structure modeling. Furthermore, they built a comprehensive structural database of human GPCRome using RoseTTAFold. This structural resource can be used as a reference for virtual drug screening and target discovery [10].

In terms of protein complex prediction, Humphreys et al. used RoseTTAFold to predict the structures of antibody–antigen and enzyme–inhibitor complexes from their families. Modeling results showed that the model can recapitulate interface regions and can identify key binding residues for complexes with good co-evolutionary signal. However, the predicted models for immune-related complexes with poor evolutionary coupling deviated more [11].

In addition, in the field of synthetic biology, Chen et al. used RoseTTAFold in synthetic biology to predict the spatial organization of multi-enzyme assemblies such as polyketide synthases. Modeling results can be used to guide the arrangement of catalytic centers and channels for improved performance of multi-enzyme assemblies in metabolic pathway design. Moreover, the models can also be used to assemble artificial protein devices such as "Rosettazymes" with nanoscale organization and resolution [12].

Flexible architecture and efficient modeling enable RoseTTAFold to be used in GPCR and multi-enzyme assembly prediction. RoseTTAFold application is not limited to these three scenarios, but covers various structural biology and synthetic biology use cases.

5. Case Study: Sweet Taste Receptor (9OPW)

The human sweet taste receptor (Sweet taste receptor) is one of the GPCRs resolved by cryo-electron microscopy, with the PDB code 9OPW. It can be seen in figure 1. This receptor belongs to the taste receptor family and is a heterodimer composed of the T1R2 and T1R3 subunits. It plays a crucial role in sensing sugars and artificial sweeteners, as well as regulating energy intake and metabolic balance, making it an important target in the study of metabolic diseases such as diabetes and obesity. The 9OPW structure consists of two chains with a resolution of 2.86 Å, forming a heterodimer (A1B1) that includes multiple α-helices and β-sheets.

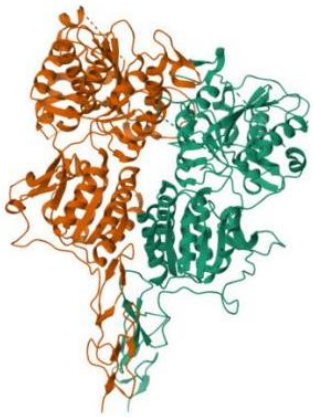

Fig. 1 PDB Database Model of 9OPW [13]

The experimental procedure was as follows: First, the amino acid sequence of the target protein was downloaded from the PDB database in FASTA format. Then, the AlphaFold Serve online platform was used by uploading the FASTA file on the Submit Job page, with parameters set as follows: Model preset = multimer, Copies = 1, MSA search = on, and Templates = use template. Next, in ColabFold, the RoseTTAFold2 Notebook (experimental version, WIP) was opened, and parameters were configured as: params: RF2_apr23, input the amino acid sequence, set symmetry settings to sym: X and order: 1, specify msa_concat_mode: diag, choose msa_method: mmseqs2, and set pair_mode to unpaired_paired. In the RoseTTAFold2 settings section, num_recycles was set to 0, max_msa to 128, and num_models to 1. The "Run all" command was executed sequentially to generate the predicted protein model. Finally, the predicted pdb file was opened in PyMOL, where structural alignment (global alignment) was performed with the experimental structure, and the RMSD value was calculated to evaluate the prediction accuracy. The actual operation demonstration is shown in Fig. 2, Fig. 3, and Fig. 4.

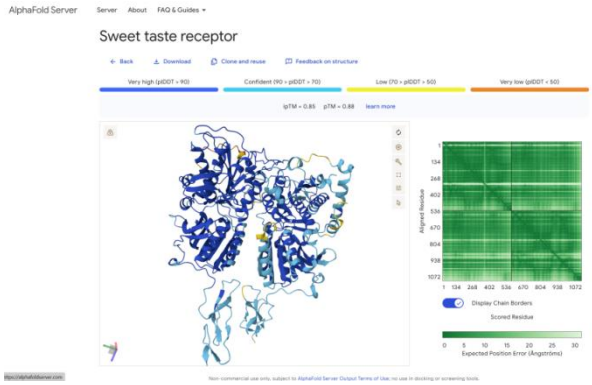

Fig. 2 AlphaFold Predicted Protein Model

Picture credit: Original

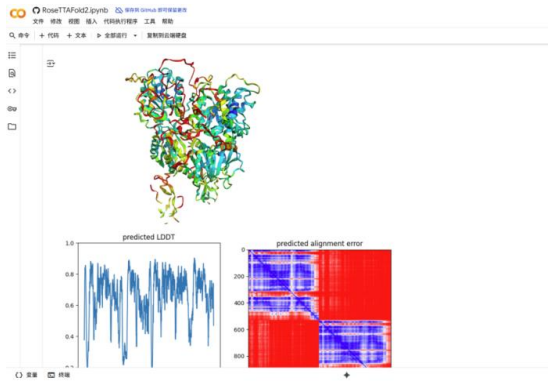

Fig. 3 RoseTTA2 Predicted Protein Model

Picture credit: Original

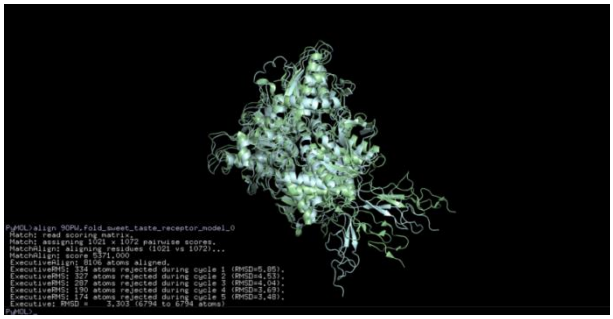

Fig. 4 RMSD analysis in PyMOL for AlphaFold's top-ranked predicted model

Picture credit: Original

Table 1. Comparison of RMSD between 9OPW, AlphaFold, and RoseTTAFold models

Tool	RMSD				
9OPW	2.86Å				
AlphaFold	3.303	3.866	3.443	3.262	3.342
RoseTTAFold	7.973				

Table 1 shows RMSD between 9OPW, AlphaFold, and RoseTTAFold models compared using Pymol. When comparing the experimentally determined structure (9OPW) with predicted models, we can see a large deviation in the lower part of the structure. This is due to the fact that in the extracellular domain of sweet taste receptor there are multiple subdomains (e.g., VFT, CRD) connected with flexible linkers, and therefore this part of the protein has a high flexibility. AlphaFold, which is based on neural network architecture, is accurate in predicting intra-domain folding, but it makes mistakes in the relative positioning of different domains (especially when there is no information about ligand, glycosylation, or inter-chain pairing information). Therefore, the overall RMSD is largely affected by the inter-domain misalignment, and when only the upper core domain is aligned, the RMSD decreases greatly, meaning that the deviations are located mostly in the linkers and stem.

When we compare the experimental model (9OPW) to the RoseTTAFold prediction, which RMSD is larger than that of AlphaFold? The reason for this is that our three-track network finds a simultaneous fit to both sequence, distance map, and 3D coordinate information. Therefore, in regions where there is high flexibility in protein conformation (e.g. the linker between VFT and CRD), the model will on average find multiple ways to "fold out" the correct inter-domain orientation. Thus, this suggests that we should think of RoseTTAFold more as modelling the inter-chain interaction in heterodimeric proteins; and that when there are no stabilizing factors like ligands or glycosylation missing in the model, proteins might be less well-defined in local domain positioning, giving rise to higher RMSD values. Importantly, this does not imply incorrect folding, but rather that RoseTTAFold adopts a more "relaxed" approach to inter-domain orientation. As long as only the core domains are aligned, the prediction looks very similar to the experiment.

On average, RoseTTAFold achieved higher RMSD values than AlphaFold for this system for two main reasons. First, the three-track network places more emphasis on predicting inter-chain and inter-domain interactions, thereby leaving the heterodimer or complex with a larger range of possible conformations. But this flexibility results in more variation in the inter-domain orientation and hence higher RMSD when there is no ligand or glycosylation present to provide a constraint. Second, AlphaFold has stronger monomeric fold performance and gets closer to a more stable estimate of the core relationship between inter-domain - hence there is closer agreement with experiment. Note that although RoseTTAFold has strengths for modeling multi-protein complexes and protein – ligand systems, the sweet taste receptor (9OPW) here is a heterodimeric membrane receptor where the interacting domains are highly conserved transmembrane protein domains, which are structurally stable. For this case, AlphaFold2, which has the deeper MSA database and optimal Transformer architecture, is able to more accurately model the subunit orientation and hence achieve higher prediction accuracy. The opposite would be true if RoseTTAFold had an architecture advantage for modeling flexible linkers; however, in this case there was no ligand/glycosylation information available, so the lack of this information led to inter-domain misalignment and a higher RMSD. Hence, the results of this case study do not indicate which tool is 'better' overall, but the differences in performance that we would expect to see under certain structural conditions.

6. ProteinGAN and ProtGPT2: De Novo Protein Sequence Design

6.1. Introduction and Features of ProteinGAN and ProtGPT2

Unlike AlphaFold2and RoseTTAFold, which are focused on protein structure prediction, ProteinGAN and ProtGPT2 are two new directions of generative artificial intelligence in the design of protein sequence, which are based on the hypothesis that protein sequences are also a kind of "language", with grammar and semantic rules. Generative models, including GANs and Transformers, can learn the statistical distribution of natural protein sequences and generate new sequences with structural plausibility and potential functionality.

ProteinGAN uses a Generative Adversarial Network (GAN) framework, where generator and discriminator play against each other. Through iterative adversarial training, the generator learns the distribution of natural sequences and can finally generate proteins with similar statistical and functional behaviors to natural ones [14]. The advantage of this method lies in the fact that its generative capability overcomes the limitation of natural evolution and can quickly explore new sequence space to speed up the discovery of new proteins.

However, challenges still exist: the generated sequences are only computationally plausible but not experimentally confirmed; the models are black boxes, it is hard to derive the physical mechanism of how to form structures in the model; the function of generated proteins is also limited, it is hard to design proteins with specific biological behaviors.

6.2. Application Examples of ProteinGAN and ProtGPT2 in De Novo Protein Sequence Generation

In practice, ProteinGAN and ProtGPT2 have shown remarkable ability in de novo protein design. Repečka et al. employed ProteinGAN (a self-attention GAN model) to synthesize thousands of new enzyme sequences (malate dehydrogenase serving as a template). Several of these A I-generated sequences were then synthesized and evaluated for their ability to fold stably and catalytic activity [13]. This was the first time that a protein engineered using a GAN has been shown to function, and thus it represents one of the first successful explorations of functional sequence space beyond natural evolution. The researchers demonstrated a new cycle of enzyme engineering in which generative adversarial networks can quickly generate highly diverse and previously unseen functional proteins.

Ferruz et al. 's ProtGPT2 is a Transformer-based protein language model that views amino acid sequences as a "biological language." ProtGPT2 was trained on approximately 50 million protein sequences and can generate de novo protein sequences with reasonable structural and physicochemical properties. The generated outputs have natural amino acid composition and are predicted to form proper structure. For example, of the generated sequences, 88% are predicted to be globular like natural proteins. Moreover, a handful of designs were validated using AlphaFold2 and shown to be likely to fold into stably ordered 3D structures. These designs are predicted to form well-formed hydrophobic cores and standard 2D secondary structure motifs, like natural proteins [15].

7. Conclusion

This study comprehensively compared three popular protein structure prediction software tools, AlphaFold2, RoseTTAFold and ProteinGAN/ProtGPT2 in both theoretical study and experimental application. Taking Sweet Taste Receptor (PDB: 9OPW) as an example, the results showed that AlphaFold2 was more accurate in modeling the stable transmembrane and conserved regions of the receptor, while larger RMSD values were obtained from RoseTTAFold due to its flexible treatment of inter-domain orientation. The differences in results reflected the two software tools' design philosophies: AlphaFold2 tends to have more conservative positioning of domains, while RoseTTAFold tends to place more emphasis on possible inter-chain and inter-domain orientations and thus provides more flexibility in modeling. In general, AlphaFold2 was more accurate in modeling single-proteins and structurally stable complexes, while RoseTTAFold was more flexible

in modeling multi-chain assemblies and protein-ligand systems. In contrast, ProteinGAN and ProtGPT2 are new players in the field of de novo design of proteins which can produce novel sequences beyond natural evolution, demonstrating the revolutionary potential of generative AI in structural biology.

These studies reveal the revolutionary influence of artificial intelligence on molecular and synthetic biology. With the assistance of deep learning, evolutionary information and generative information, the prediction of protein structure, function and enzyme design have become more efficient.

In the future, it is necessary to further enhance the interpretability and experimental verification of the structures predicted by AI. It is also worth exploring the combination of multi-omics information, dynamic simulation and energy-based refinement to improve the biological accuracy. With the gradual convergence of computational biology and experimental biology, the prediction and design of proteins and enzymes based on AI will still play a central role in revealing the molecular mechanism and designing functional biomolecules with precision.

References

[1] Anfinsen C. B. Principles that govern the folding of protein chains. Science, 1973, 181(4096): 223-230.

[2] Dill K. A., MacCallum J. L. The protein-folding problem, 50 years on. Science, 2012, 338(6110): 1042-1046.

[3] Karamanos T. K. Chasing long-range evolutionary couplings in the AlphaFold era. Biopolymers, 2023, 114(3): e23530.

[4] Yang Z., Zeng X., Zhao Y., Chen R. AlphaFold2 and its applications in the fields of biology and medicine. Signal Transduct Target Ther, 2023, 8(1): 115.

[5] Xu T., Xu Q., Li J. Toward the appropriate interpretation of AlphaFold2. Front Artif Intell, 2023, 6: 1149748.

[6] Fang X., Wang F., Liu L., et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. Nat Mach Intell, 2023, 5: 1087-1096.

[7] Yang Q., Jian X., Syed A. A. S., et al. Structural comparison and drug screening of spike proteins of ten SARS-CoV-2 variants. Research (Wash D C), 2022, (2): 9781758.

[8] He X. H., You C. Z., Jiang H. L., et al. AlphaFold2 versus experimental structures: evaluation on G protein-coupled receptors. Acta Pharmacol Sin, 2023, 44(1): 1-7.

[9] Chen L., Li Q., Nasif K. F. A., Xie Y., et al. AI-driven deep learning techniques in protein structure prediction. Int J Mol Sci, 2024, 25(15): 8426.

[10] Baek M., et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021, 373(6557): 871-876.

[11] Humphreys I. R., et al. Computational predictions of protein complexes with AlphaFold and RoseTTAFold. Science, 2021, 374(6573): eabm4805.

[12] Chen, L., Li, Q., Nasif, K.F.A., et al. AI-Driven Deep Learning Techniques in Protein Structure Prediction. Int. J. Mol. Sci, 2024, 25: 8426.

[13] Wang H, Chen X, Dai Y, Pidathala S, Niu Y, Zhao C, Li S, Wang L, Lee CH. Structure and activation mechanism of human sweet taste receptor. Cell Res. 2025, 35(10):775-778.

[14] Repecka D., Jauniskis V., Karpus L., et al. Expanding functional protein sequence spaces using generative adversarial networks. Nat Mach Intell, 2021, 3: 324-333.

[15] Ferruz N., Schmidt S., Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun, 2022, 13: 4348.