International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 12, Issue 9 [September. 2023] PP: 149-155

Research on Protein Tertiary Structure Generation

Zhichong Ma, Jinghua Chen, Changsheng Wang
College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai 200093, China
Corresponding Author: Jinghua Chen
Funding: The financial support for this work was supported by ZBKT202305 Key Lab of Intelligent and Green Flexographic Printing

ABSTRACT: *Proteins are a class of vital biomacromolecules in biological organisms, exhibiting four levels of structural organization: primary, secondary, tertiary, and quaternary structures. The tertiary structure of proteins is typically employed to depict the spatial characteristics of proteins, facilitating the analysis of their functions and actions. This article provides a comprehensive review and investigation of protein tertiary structure generation utilizing both physical experimental methods and deep learning approaches. Simultaneously, it conducts an analysis of their respective strengths and weaknesses, offering valuable insights and scientific significance for the future development and research of protein tertiary structure generation.*
Key words: *proteins, protein structures, protein tertiary structure generation*

Date of Submission: 10-09-2023 Date of acceptance: 24-09-2023

I. INTRODUCTION

Proteins represent intricate, crucial, and indispensable natural polymers [1], pivotal in sustaining the growth, differentiation, and repair of biological organisms, thus playing a paramount role in the realm of biology [2]. Based on structural hierarchy, proteins can be categorized into primary, secondary, tertiary, and quaternary structures [1]. The properties of lower-level structures can determine the quality of higher-level structures, thereby influencing their functionality and impact. The primary structure of a protein consists of multiple amino acids arranged in a specific sequence [3]. The secondary structure is formed when certain amino acids interact via hydrogen bonds, resulting in local spatial conformations such as α-helices, β-sheets, and β-turns, among others [4]. The tertiary structure, built upon the secondary structure, involves further coiling and folding of the polypeptide chain to create more complex, globular molecular structures, which can be categorized into α, β, α+β, and α/β types [5]. The quaternary structure is constituted by the assembly of two or more subunits, each possessing its own tertiary structure, through non-covalent interactions, resulting in a complex with a distinct tertiary structure [6]. The tertiary structure of proteins is the most commonly employed structure for representing the spatial characteristics of proteins.

The prediction of various protein properties and the inference of their functions and roles can be achieved through the analysis of protein tertiary structure. Currently, physical experimental methods employed for protein tertiary structure resolution include X-ray crystalline diffraction, nuclear magnetic resonance, and cryogenic electron microscopy.

(1) X-ray crystalline diffraction. The principle behind it is that when X-rays strike particles in a molecular crystal, they interact with the electrons in the crystal, leading to a diffraction effect. By collecting these diffraction signals with a detector, it is possible to determine the distribution of electron density within the crystal. This information allows us to obtain the spatial coordinates of the particles and ultimately derive the tertiary structure of the protein. X-ray crystalline diffraction is the most commonly used and highly accurate technique for determining protein tertiary structures [7], and the resulting protein tertiary structure is depicted in Figure 1. As shown in Table 1, this method offers a wide range of protein molecular weights, high resolution, and suitability for soluble proteins, pancreatic enzymes, and protein complexes. However, it requires protein crystallization. Additionally, X-rays themselves carry a significant amount of energy, making it easy to cause radiation damage to crystals. One limitation is that it is not suitable for analyzing proteins with larger molecular weights.

(2) Nuclear magnetic resonance. The underlying principle involves atomic nuclei with lone pair electrons undergo spin in response to an external magnetic field, resulting in energy level transitions through Zeeman splitting. This process leads to the absorption and emission of electromagnetic radiation, generating distinct resonance spectra. By recording changes in the spectrum, one can determine the location and relative abundance of the atoms within the molecule, enabling quantitative analysis, molecular weight determination, and molecular structure analysis. This technique is extensively utilized in structural biology

[8], and the resulting protein tertiary structure is depicted in Figure 2. As indicated in Table 1, this method provides a more accurate depiction of the natural structures of biomacromolecules and is suitable for investigating transient and unstable complexes. It is well-suited for studying protein dynamics and can also be applied to research the three-dimensional structures of membrane proteins. However, the presence of structural instability in proteins within a solution can make it challenging to obtain stable nuclear magnetic resonance signals, and there are limitations in its capacity when dealing with large biomolecules.

(3) Cryogenic electron microscopy. The principle involves rapidly freezing biomacromolecules in milliseconds, embedding them in a glassy state of ice. Highly coherent electrons are employed as a light source, and an acceleration voltage of 80 to 300 kV is applied to accelerate the electron beam. The electron beam passes through the sample and the surrounding ice layers, causing the sample to scatter. Subsequently, the scattered signals are recorded using a detector and lens system, collecting two-dimensional projections of the biomacromolecule from various orientations. Afterward, image processing and three-dimensional reconstruction techniques are employed to compute the fine tertiary structure of the biomacromolecule. This method combines the advantages of X-ray crystalline diffraction and nuclear magnetic resonance, making it the most suitable alternative for structural research [9]. The resulting protein tertiary structure is depicted in Figure 3. As shown in Table 1, this method allows for the direct acquisition of the structure and conformational changes of biomacromolecules under near-physiological conditions. It does not require protein crystallization, demands relatively small sample quantities, and is applicable to a wide range of researches, and resolution of the three-dimensional fine structure of macromolecular complexes. However, it imposes high demands on sample preparation, particularly for samples with lower molecular weights and poor homogeneity, and the resolution achieved may be comparatively lower.

Table 1. Comparation of physical experimental methods for protein tertiary structure resolution

	X-ray crystalline diffraction	**Nuclear magnetic resonance**	**Cryogenic electron microscopy**
Advantages	1. A wide range of protein molecular weights 2. High resolution 3. Suitability for soluble proteins, pancreatic enzymes, and protein complexes	1. A more accurate depiction of the natural structures of biomacromolecules 2. It is suitable for investigating transient and unstable complexes 3. It is well-suited for studying protein dynamics 4. It can also be applied to research the three-dimensional structures of membrane proteins	1. Direct acquisition of the structure and conformational changes of biomacromolecules under near-physiological conditions. 2. It does not require protein crystallization 3. It demands relatively small sample quantities 4. It is applicable to a wide range of researches 5. Resolution of the three-dimensional fine structure of macromolecular complexes.
Limitations	1. Protein crystallization requirement 2. X-rays themselves carry a significant amount of energy, making it easy to cause radiation damage to crystals 3. It is not suitable for analyzing proteins with larger molecular weights	1. The presence of structural instability in proteins within a solution can make it challenging to obtain stable nuclear magnetic resonance signals 2. There are limitations in its capacity when dealing with large biomolecules.	1. It imposes high demands on sample preparation, particularly for samples with lower molecular weights and poor homogeneity 2. Resolution achieved may be comparatively lower

These experimental methods often necessitate prolonged timeframes, involve intricate measurement techniques with numerous steps, carry a risk of experimental failure, incur high costs, and require protein crystallization. With the rise of deep learning technology, utilizing deep learning for protein tertiary structure generation has become mainstream. This method offers high efficiency, rapid processing, reduced reliance on extensive manual labor and experiments, and lower costs. The process of protein tertiary structure generation based on deep learning involves four stages: generating a multiple sequence alignment file, generating residue distance and orientation geometric constraint files, generating tertiary structure files, and visualizing the tertiary structure.

Depending on the different approaches to deep learning model generation, protein tertiary structure generation models can be categorized into three types: based on autoregressive models, based on variational autoencoders (VAE), and based on generative adversarial networks (GAN). The following is a detailed exploration of these tertiary structure generation models.

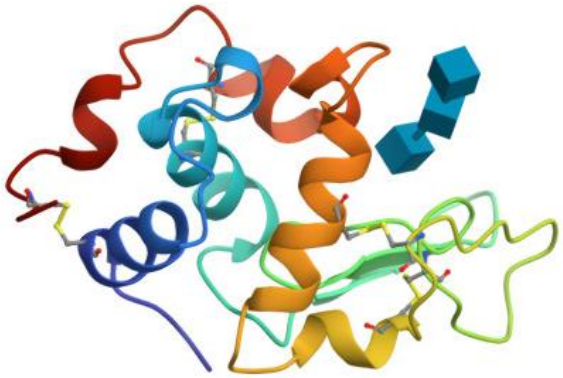

Figure 1. Protein tertiary structure determination using x-ray crystalline diffraction

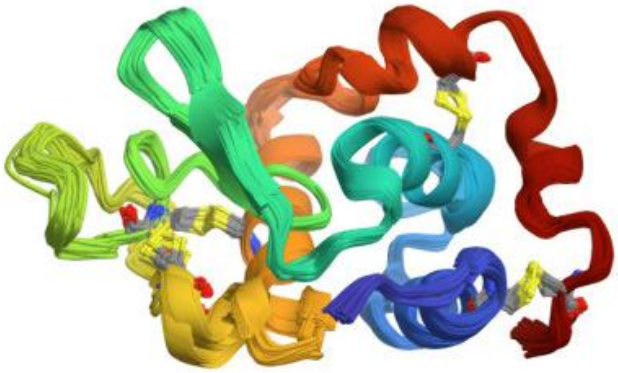

Figure 2. Protein tertiary structure determination using nuclear magnetic resonance

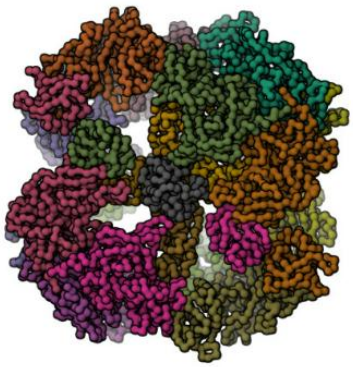

Figure 3. Protein tertiary structure determination using cryogenic electron microscopy

II. BASED ON AUTOREGRESSIVE MODELS

The autoregressive model refers to a process in which a variable is regressed on its own past values. It utilizes the historical time data of the variable to predict its current value [10]. As shown in Table 2, classical models like ProteinSolver [11] employ deep graph neural networks to rapidly and highly accurately design sequences that can fold into predetermined shapes. They can generate desired geometric sequences from existing proteins and combine them with backbone detection methods to design new structures more quickly, enabling more extensive sampling of large conformational spaces. However, the learning curve is steep, and a high degree of domain-specific expertise is required. The ESM-1b Transformer [12] utilizes unsupervised language models based on the Transformer technology to generate new diverse sequences. It then automatically learns features to accurately represent contact maps, secondary structures, and tertiary structures. The network can generalize, with a scale comparable to models in the text domain, and it outperforms ISTM. However, due to limited model capacity, even the highest-capacity trained model cannot fit sequence datasets. The MSA Transformer [13], which takes multiple sequence alignment as input, is a language model based on the Transformer technology. It is used to reconstruct a disrupted MSA and predict contact maps and secondary structures with high accuracy. It offers higher parameter efficiency compared to previous state-of-the-art protein language models and outperforms current unsupervised contact prediction methods, suitable for all depths of

multiple sequence alignments. However, bundled attention and parameter counting can affect contact accuracy, and the diversity of input sequences has a significant impact on structural inference. RefineGNN [14] utilizes sequence autoregression decomposition and iteratively improves its predicted global structure for antibody sequence and structure co-design. It can modify generated subgraphs to accommodate the addition of new residues. It significantly outperforms sequence-based and graph-based methods in three antibody generation tasks. However, it cannot be experimentally tested in wet labs, and the PPL given by the structural-conditioned model is lower than expected.

Table 2. Research on protein tertiary structure generation based on autoregressive models

Author	Model	Targets	Limitations
Strokach, A	ProteinSolver[11]	Rapidly and highly accurately design sequences that can fold into predetermined shapes, and combine with backbone detection methods to design new structures more quickly and sample a larger conformational space at a faster pace.	Its learning curve is steep, and a high degree of domain-specific expertise is required
Rives, A	ESM-1b Transformer[12]	Generate new diverse sequences and automatically learn features to precisely represent contact maps, secondary structures, and tertiary structures.	Due to limited model capacity, even the highest-capacity trained model cannot fit sequence datasets
Rao, R	MSA Transformer [13]	Reconstruct disrupted multiple sequence alignments (MSA) and predict contact maps and secondary structures with high accuracy.	Bundled attention and parameter counting can affect contact accuracy, and the diversity of input sequences has a significant impact on structural inference
Jin, W	RefineGNN[14]	Perform antibody sequence and structure co-design by iteratively improving predicted global structures through sequence autoregression decomposition.	It cannot be experimentally tested in wet labs, and the PPL given by the structural-conditioned model is lower than expected

III. BASED ON VARIATIONAL AUTOENCODERS

The Variational Autoencoder (VAE) consists of an encoder and a decoder. The encoder maps high-dimensional input data into a lower-dimensional latent space, obtaining features, which are subsequently decoded by the decoder into an output of the same form as the input [15]. As shown in Table 3, CO-VAE and DCO-VAE [16] employ variational autoencoders to reconstruct three-dimensional protein structures from generated contact maps. The generated contact maps and structure quality are high, and they can effectively capture the sample distribution of observed contact maps, resulting in diverse generated maps. However, it is not a direct end-to-end protein tertiary structure generation model, and there exists a gap between contact map generation and the formation of three-dimensional structures. CogMol [17], combining molecular SMILES and variational autoencoders, is used for designing highly affinity and off-target selective novel virus proteins. It can handle the constrained design of synthesizable, low-toxicity, drug-like molecules and exhibits high target specificity and selectivity without target-dependent fine-tuning of frameworks or target structure information. However, due to biases in training data or inaccuracies in predictors used to control generation, it may not always generate molecules with the desired attributes. Mathematical-neural network [18] combines the SRVF function, ResNets network, and G-VAE model network to compare, deform, and generate similar yet distinct novel three-dimensional protein structures. It can accurately infer missing portions of protein structures and restore their shapes. However, there are instances of unreasonable phenomena in the generated α-helices in the middle portion of proteins, and it is still unable to generate complete structures that are chemically valid and realistically complex in tertiary structure. SPP-VAE [19], which combines a convolutional variational autoencoder network with spatial pyramid pooling, can learn directly from the tertiary structures of proteins of different lengths in a protein database, enabling the precise generation of corresponding diversity distance matrices. The output results are unaffected by input size. However, it requires a relatively large training dataset, and training the model using the PISCES-derived database may significantly reduce the accuracy of the distance matrix. Ig-VAE [20] is used to directly generate the three-dimensional coordinates of immunoglobulins using a variational autoencoder. It does not require recovering coordinates from bidirectional distance constraints, thus avoiding issues with the validity of distance matrices. It can generate novel, high-quality backbones that satisfy specific design constraints while also providing a compatible distribution of supporting elements. However, refining with Rosetta did not improve the accuracy of backbone reconstruction, and the quality decreases when the structure deviates too far from the training data.

Table 3. Research on protein tertiary structure generation based on variational autoencoder

Author	Model	Target	Limitations
Guo, X	CO-VAE DCO-VAE[16]	Generate diverse, high-quality contact maps and protein tertiary structures	It is not a direct end-to-end protein tertiary structure generation model, and there exists a gap between contact map generation and the formation of three-dimensional structures
Chenthamarakshan, V	CogMol[17]	Be used for designing high-affinity and off-target selective targeted novel virus proteins and can handle the constrained design of synthesizable, low-toxicity, drug-like molecules	Due to biases in training data or inaccuracies in predictors used to control generation, it may not always generate molecules with the desired attributes
Huang, H	mathematical-neural network[18]	Compare, deform, and generate similar yet distinct novel three-dimensional protein structures, while accurately inferring the missing portions of protein structures to restore their shapes.	There are instances of unreasonable phenomena in the generated α-helices in the middle portion of proteins, and it is still unable to generate complete structures that are chemically valid and realistically complex in tertiary structure
Alam, F	SPP-VAE[19]	Learn directly from the tertiary structures of proteins of different lengths in a protein database, allowing for the precise generation of corresponding diversity distance matrices, with the output results unaffected by input size.	It requires a relatively large training dataset, and training the model using the PISCES-derived database may significantly reduce the accuracy of the distance matrix
Eguchi, R.R	Ig-VAE[20]	Avoid the need to recover coordinates from bidirectional distance constraints, thus mitigating issues with the validity of distance matrices. This approach can generate novel, high-quality backbones that satisfy specific design constraints while also providing a compatible distribution of supporting elements.	Refining with Rosetta did not improve the accuracy of backbone reconstruction, and the quality decreases when the structure deviates too far from the training data

IV. BASED ON GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs) consist of a generator and a discriminator. The generator attempts to produce realistic data to 'fool' the discriminator, while the discriminator tries to distinguish it from real data. In this continuous adversarial training, the generator's generation capability and the discriminator's discrimination ability gradually improve, ultimately producing high-quality outputs [21]. As shown in Table 4, GANs [22] employ Generative Adversarial Networks to generate fixed-length, all-atom protein backbones in a fast and fully differentiable manner. It can recover the generated folds after forward folding, eliminating the need for explicitly encoding structural invariance to arbitrary rotations and translations. What's more, it can model long-range contacts. However, generating structures can be challenging, and it consumes a relatively high amount of energy. The coordinate recovery network trained may not generalize to arbitrary generated mappings without local backbone net errors, requiring additional optimization for each structure. RamaNet [23] utilizes a combination of GAN architecture and Long Short-Term Memory units to design de novo protein backbones. It can generate a large number of random, logical, rigid, and compact helical protein backbone topologies without the need for user-defined topological values. However, it is only suitable for dense protein main chains containing 80 to 150 amino acids. It has not been successful in generating structures that include beta sheets, and it relies on other simulation tools to obtain proteins with specific sequences. ContactGAN [24] utilizes a generative adversarial network to refine and denoise contact maps, improving the accuracy of protein tertiary structure models. This approach significantly enhances the precision of contact maps predicted by CCMpred, DeepCov, and DeepContact. However, improving the accuracy of contact map predictions, particularly for crucial long-range contacts between residues 12-18 and 112-118, remains challenging when using trRosetta.

Table 4. Research on protein tertiary structure generation based on generative adversarial networks

Author	Model	Target	Limitations
Anand, N	GANs[22]	Generate fixed-length, all-atom protein backbones in a fast and fully differentiable manner, recover the generated folds after forward folding, and model long-range contacts.	Generating structures can be challenging, and it consumes a relatively high amount of energy. The coordinate recovery network trained may not generalize to arbitrary generated mappings without local backbone net errors, requiring additional optimization for each structure.
Sabban, S	RamaNet[23]	Automatically generate a large number of random, logical, rigid, and compact helical protein backbone topologies without the need for user-defined topological values.	It is only suitable for dense protein main chains containing 80 to 150 amino acids. It has not been successful in generating structures that include beta sheets, and it relies on other simulation tools to obtain proteins with specific sequences
Maddhuri, V	ContactGAN[24]	Refine and denoise contact maps to improve their accuracy, leading to the generation of higher-precision protein tertiary structure models.	Improving the accuracy of contact map predictions, particularly for crucial long-range contacts between residues 12-18 and 112-118, remains challenging when using trRosetta

V. CONCLUSION

This article elaborates on the significance of proteins in biology and their structural classifications. Among these, the tertiary structure of proteins is commonly employed to depict their spatial characteristics, allowing for the analysis of protein functionality and roles based on their tertiary structures. The generation of protein tertiary structures can be achieved through physical experimental methods, including X-ray crystalline diffraction, nuclear magnetic resonance, and cryogenic electron microscopy, as well as through deep learning-based methods, such as autoregressive-based models, variational autoencoders (VAE), and generative adversarial networks (GAN). The article provides detailed explanations of these methods and summarizes their advantages and limitations. Clearly, compared to physical experimental methods, the methods based on deep learning for protein tertiary structure generation offer higher efficiency, faster processing, reduced reliance on extensive manual work and experiments, and lower costs.

REFRENCES

[1]. Bordoloi, Hemashree. [2021]. "Analytical Model to Predict Protein Structure using Soft-Computing Approach" Bioscience Biotechnology Research Communications, Vol. 14, No. 6: pp.324-327

[2]. Jiang, Q., Jin, X., Lee, S.J. and Yao, S. [2017]. " Protein secondary structure prediction: A survey of the state of the art" Journal of Molecular Graphics and Modelling, Vol. 76: pp.379-402.

[3]. De Brevern, AG., Etchebest, C., and Hazout, S. [2000]. "Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks" Proteins, Vol. 41: pp.271–287.

[4]. Chou, PY., Fasman, GD. [1974]. "Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins" Biochemistry, Vol. 13: pp.211–222.

[5]. Levitt, M., Chothia, C. [1976]. "Structural patterns in globular proteins" Nature, Vol. 261: pp. 552–558.

[6]. Dey, S., D. Levy, E. [2018]. "Inferring and Using Protein Quaternary Structure Information from Crystallographic Data" Protein Complex Assembly. Methods in Molecular Biology, Vol. 1764: pp. 357–375.

[7]. Sayers, Z., Avşar, B., CHOLAK, E, et al. [2017]. "Application of advanced X-ray methods in life sciences" Biochimica et Biophysica Acta (BBA)-GeneralSubjects, Vol. 1861, Issue 1: pp. 3671-3685.

[8]. Shi, Y. Y., Wu, J.H. [2007]. Progress in nuclear magnetic resonance spectroscopy used to structural biology" Acta Biophysica Sinica, Vol. 23, Issue 4: pp. 240-245.

[9]. Akbar, S., Mozumder, S., Sengupta, J. [2020]. "Retrospect and prospect of single particle cryo-electron microscopy: the class of integral membrane proteins as an example" Journal of Chemical Information and Modeling, Vol. 60, Issue 5: pp. 2448-2457.

[10]. Jin, Q., Cao, L., Yang, R. [2015]. "Auxiliary model based recursive and iterative least squares algorithm for autoregressive output error autoregressive systems" Applied Mathematical Modelling, Vol. 39, Issue 22: pp. 7008-7016.

[11]. Strokach, A., Becerra, D., Corbi-Verge, C., et al. [2020]. "Fast and Flexible Protein Design Using Deep Graph Neural Networks" Cell Syst, Vol. 11, Issue 4: pp. 402-411.e4.

[12]. Rives, A., Meier, J., Sercu, T., et al. [2021]. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences" Proceedings of the National Academy of Sciences of the United States of America, Vol. 118, Issue 15: pp. e2016239118.

[13]. Rao, R., Liu, J., Verkuil, R., et al. [2021]. "MSA Transformer" bioRxiv, 2021.02.12.430858.

[14]. Jin, W., Jeremy, W., Regina, B., T. Jaakkola. [2020]. "Iterative Refinement Graph Neural Network for Antibody Sequence-Structure Co-design" ArXiv abs/2110.04624: pp. n. pag.

[15]. Medbouhi, A.A., Polianskii, V., Varava, A., Kragic, D. [2023]. "InvMap and Witness Simplicial Variational Auto-Encoders" Machine Learning and Knowledge Extraction, Vol. 5, Issue 1: pp. 199-236.

[16]. Guo, X., Tadepalli, S., Zhao, L., Shehu, L. [2020]. "Generating tertiary protein structures via an interpretative variational autoencoder" arXiv:2004.07119.

[17]. Chenthamarakshan, V., Das, P., Samuel C. et al. [2020]. "CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models" arXiv: 2004. 01215.

[18]. Huang, H., Lin, L., Zhu, F., et al. [2021]. "G-VAE, a Geometric Convolutional VAE for Protein Structure Generation" arXiv:2106.11920.

[19]. Alam, F., Shehu, A. [2021]. "Generating Physically-Realistic Tertiary Protein Structures with Deep Latent Variable Models Learning Over Experimentally-available Structures" in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): pp. 2463-2470.

[20]. Eguchi, R.R., Choe, C.A., Huang, P.S. [2022]. "Ig-VAE: Generative modeling of protein structure by direct 3D coordinate

generation" PLoS Comput Biol, Vol. 18, Issue 6: pp. e1010271.
[21]. Luo, W., Wang, P., Wang, J., An, W. [2019]. " The research process of generative adversarial networks" Journal of Physics: Conference Series, Vol. 1176, Issue 3: pp. 032008.
[22]. Anand, N., Eguchi, R., and Huang, P. S. [2019]. "Fully differentiable full-atom protein backbone generation" in Deep generative models for highly structured data, ICLR 2019 Workshop, 2019 (New Orleans, LA: ICLR 2019).
[23]. Sabban, S., Markovsky, M. [2020]. "Ramanet: Computational De Novo Helical Protein Backbone Design Using a Long Short-Term Memory Generative Adversarial Neural Network" F1000Research, Vol. 9, Issue 298: pp. 298.
[24]. Maddhuri, V. Subramaniya，SR., Terashi, G., Jain, A., et al. [2021]. "Protein Contact Map Refinement for Improving Structure Prediction Using Generative Adversarial Networks" Bioinformatics, Vol. 37, Issue 19: pp. 3168–3174.