PERSPECTIVE

Current progress, challenges, and future perspectives of language models for protein representation and protein design

Tao Huang[1,*] **and Yixue Li**[1,2,3,4,5,*]

[1]Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

[2]Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China

[3]Guangzhou Laboratory, Guangzhou 510005, China

[4]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

[5]Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai 200433, China

*Correspondence: huangtao@sibs.ac.cn (T.H.); yxli@sibs.ac.cn (Y.L.)

Received: February 15, 2023; Accepted: May 18, 2023; Published Online: May 21, 2023; https://doi.org/10.1016/j.xinn.2023.100446

Citation: Huang T. and Li Y. (2023). Current progress, challenges, and future perspectives of language models for protein representation and protein design. The Innovation **4(4)**, 100446.

The sequence-structure-function paradigm of protein is the basis of molecular biology. What is the underlying mechanism of such sequence and structure/function corresponding relationship? We reviewed the methods for protein representation and protein design. With these protein representation models, we can accurately predict many properties of proteins, such as stability and binding affinity. Progen, Chroma, RF Diffusion, SCUBA, and other protein design models have demonstrated how human-designed artificial proteins can have desired biological functions. The protein design will revolutionize drug development. And more efficient artificial enzymes that break down industrial waste or plastics will contribute to carbon neutrality. We also discussed the three greatest challenges of protein design in future and possible solutions.

With the rapid growth of ChatGPT (https://chat.openai.com/), which has over 100 million active users in just two months, people hear language model over and over. However, many people are unaware that the language models have been used extensively in protein research.[1] We do know that the famous software AlphaFold2 can accurately predict the protein structure based only on the protein sequence with attention and transformer architecture.[2] The attention mechanism proposed by Google in 2017 is a revolutionary innovation for natural language processing. The basis of ChatGPT and GPT-4 are large language models, which essentially belong to natural language processing. Protein sequences can be treated as sentences and the 20 amino acids are the words. Since the sequence-structure-function paradigm of protein is fundamental to molecular biology, we can use language models borrowed from computer science to investigate the underlying mechanisms of this relationship. Language models have two primary applications in protein sciences: protein representation and protein design (Figure 1).

The protein representation is a problem of how to represent a protein sequence with numerical vectors. Due to the varying lengths of different proteins, it is not possible to use the original amino acids directly. The most intuitive method is the amino acid composition which counts the frequencies of the 20 amino acids. Then, it is found that some amino acids share similar properties, and pseudo-amino acid composition is developed. In addition to composition, the transition and distribution of amino acids have been found to be important and are also used for protein representation. Physical energy-based approaches are widely used to calculate the most possible structure since proteins tend to fold toward state of lower free energies.[3] The optimization goal of physical energy-based approaches is to minimize free energy. The biological meaningful sequence patterns shared by a protein family are summarized as motifs or domains. The structural features, such as intrinsically disordered regions, evolutionary conserved regions, protein-protein interaction sites, enzyme active sites, and PTM (post-translational modification) sites, are also critical for protein characterization. Subsequently, network-based functional features were proposed, including KEGG and GO enrichment scores of a protein's network microenvironment, which refers to the interaction neighbors on the protein-protein network. These features have demonstrated significant improvements in predicting protein stability.[4]

With the advent of deep learning, people find that a protein sequence can be considered as a sentence made up of 20 words, i.e., amino acids. It is like human language. Therefore, an increasing number of deep neural network-based language models have been proposed for protein representation. There are over 20 protein representation learning models.[5] ProtVec used uncontextualized word2vec (feedforward neural network, FNN-based). SeqVec used contextualized ELMo (long short-term memory, LSTM-based). ProtBERT replaced LSTMs with a Transformer (Bidirectional Encoder Representation from Transformers, BERT-based). T5-XL-BFD used a larger/different transformer and more data. T5-XL-U50 fine-tuned T5-XL-BFD on non-redundant data. With these protein representation models, we can accurately predict many properties of proteins. That leads to another application, protein design.

Could we design a non-existing protein with desired properties, such as a protein with a circular 3D structure, a protein that can bind SARS-CoV-2, a protein with antifreeze, or an antibacterial protein? Could we speak the language of protein and create any desired proteins as we wish?

Protein structure prediction is a crucial part of protein design. It has been widely used for the evaluation of designed candidate proteins with topology or symmetry constraints. There are several protein structure prediction rivals of well-known AlphaFold2, including ESMFold (https://github.com/facebookresearch/esm) developed by the Meta AI team, which predicts the structures of 600 million uncharacterized proteins from bacteria, viruses, and other microorganisms; RoseTTAFold (https://github.com/RosettaCommons/RoseTTAFold) developed by David Baker, which has a long development history since 1998 and great academic reputation; Uni-Fold (https://github.com/dptech-corp/Uni-Fold) developed by DP Technology, which can work on more hardware platforms and has much higher efficiency; and MEGA-Protein (MindSpore for Evolutionary Generation & Assessment Protein, https://gitee.com/mindspore/mindscience) developed by Huawei, which has better performance on the orphan sequences. Each program has its own advantages. When designing proteins for different organisms using different hardware platforms, they can be replaced with each other.

SCUBA (Side Chain-Unknown Backbone Arrangement) developed by Haiyan Liu[6] proposes a statistical model that uses neural network-form energy terms. It is a *de novo* protein design model that does not need a template. The *de novo* protein design workflow of SCUBA has two steps, neighbor counting followed by neural network training for learning. After the artificial proteins are generated, ABACUS2 selects the sequence for natural backbones. SCUBA and ABACUS2 are available at https://doi.org/10.5281/zenodo.4533424.

RoseTTAFold Diffusion (RF Diffusion, https://github.com/RosettaCommons/RFdiffusion) developed by David Baker proposes a guided diffusion model for generating new proteins by adding and removing noise. RF Diffusion performs well for a broad range of protein design problems, such as topology-constrained protein design and enzyme active site scaffolding for therapeutic protein design. To achieve atomically accurate design, David Baker and his colleagues applied reinforcement learning for top-down design of protein architectures.[7] The deviation between the designed and real protein structure is on average smaller than the size of a single atom.

Chroma (https://generatebiomedicines.com/chroma) developed by Generate Biomedicines creates new proteins with desired structural or functional properties using a generative model. It combines a structured diffusion model for protein backbones with scalable molecular neural networks for backbone synthesis and all-atom design. The designed proteins will have the required functional

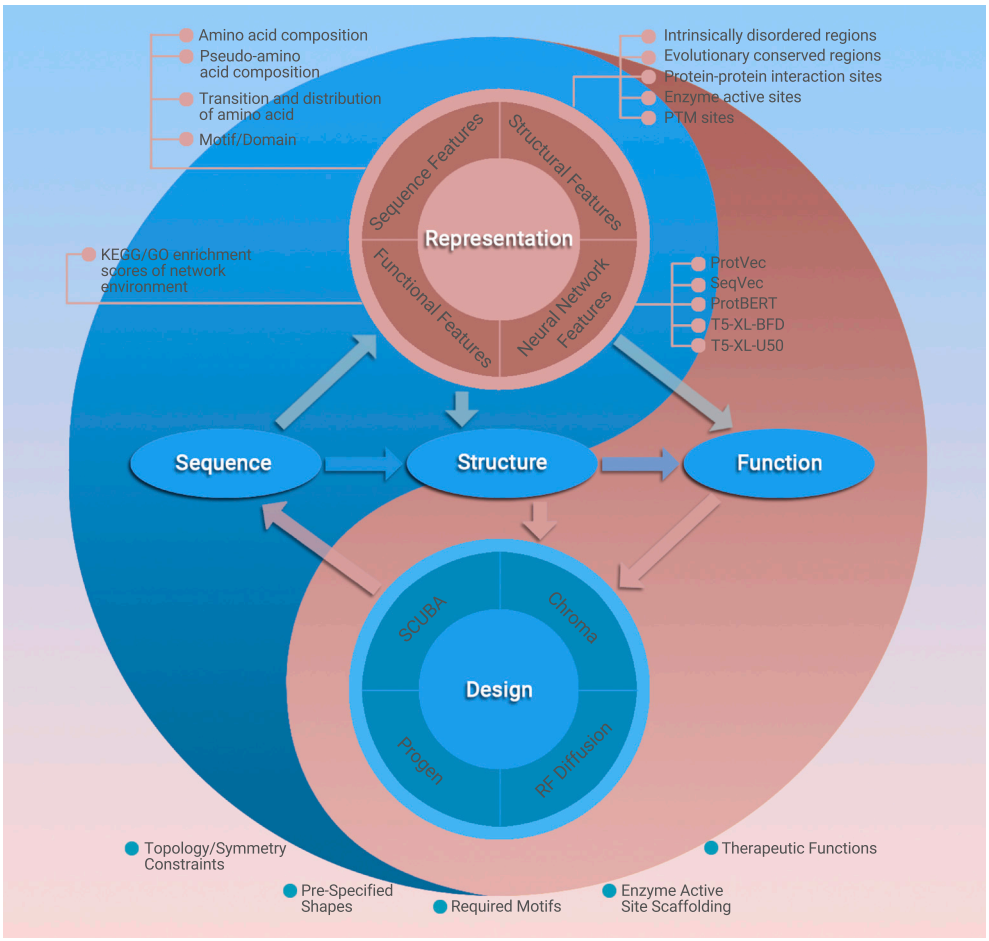

Figure 1. The scheme for protein representation and protein design

structural motifs, with symmetry constraints or in a pre-specified shape. It can even design proteins with 3D structures in arbitrary given shapes, such as alphabet and numbers (https://cdn.generatebiomedicines.com/video/alphabet_padded.mp4). The proteins with shapes like the alphabet have existed in nature for a long time,[8] but Chroma can design novel ones with the same shapes.

Progen (https://github.com/salesforce/progen) developed by Salesforce Research is one of the latest language models for protein engineering.[9] The model is trained on 280 million protein sequences, and it generates one million artificial proteins. 100 sequences are selected based on generation quality and diversity. 72 out of 100 Progen-generated artificial proteins express equally well in cells. Two artificial proteins have the same function against the cell walls. Without language models, it is time-consuming and costly to design enzymes.[10] First, you need to do multiple sequence alignment of natural homologs. Second, you calculate the empirical statistics of amino acids. Third, you infer a statistical model that generates artificial sequences. Fourth, you test them with a high-throughput assay for desired functions. The statistical model is difficult to construct, and its performance is usually poor since you do not know which features you should use and which model and parameters you should choose. Therefore, it requires a lot of feature engineering and parameter tunning to get a workable statistical model. The language model is an end-to-end model. It is easy to build and performs well. These results demonstrate how well language model-designed artificial proteins can have desired biological functions and how efficient they are.

A more comprehensive list of papers about protein design using deep learning can be found at https://github.com/Peldom/papers_for_protein_design_using_DL. The sequence and structure benchmark datasets, papers on function to scaffold, scaffold to sequence, function to sequence, and function to structure, are available.

The protein design will revolutionize drug development. Most drugs work by binding to proteins and triggering changes in their function. If we can generate proteins with the desired structure, we can design drugs or repurpose existing drugs to effectively bind those target proteins. What's more, more efficient artificial enzymes that break down industrial waste or plastics will contribute to carbon neutrality.

There are still several challenges ahead of us. First, the structure of a protein is dynamic *in vivo*. The point mutations, PTMs, such as phosphorylation, glycosylation, ubiquitination, methylation, acetylation, lipidation, and proteolysis, and interaction with other molecules, such as protein-protein interaction, protein-RNA binding, and protein-DNA binding, can all affect the structures and functions of proteins. The biological functions are usually regulated by the disordered regions, but the structures of such disordered regions are still poorly predicted. The ligands and flexibility (ability to change structure) should be considered for deep learning modeling. Second, more specific benchmark databases and prediction models should be constructed. If we want to design a specific enzyme, we may need high-quality data on such enzymes. And the model must be tuned for this specific task. There is no general model that can work on all problems. Let's not forget the no-free-lunch theorem for machine learning. Third, what is the ultimate truth of protein? Now, we still need to train the model on many existing proteins. Although it works well, it is essentially data fitting. Can we formulate the protein design with principles or equations rather than based on big data? Is there an end game of protein design, like from data-based AlphaGO to principle-guided AlphaZero? We believe that the language model will help understand the folding processes of proteins and their biological functions.

REFERENCES

1. Vu, M.H., Akbar, R., Robert, P.A., et al. (2023). Linguistically inspired roadmap for building biologically reliable protein language models. Nat. Mach. Intell. ***10***, 1038.
2. Tunyasuvunakool, K., Adler, J., Wu, Z., et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature ***596***, 590–596.
3. Dill, K.A., and MacCallum, J.L. (2012). The protein-folding problem, 50 years on. Science ***338***, 1042–1046.
4. Huang, F., Fu, M., Li, J., et al. (2023). Analysis and prediction of protein stability based on interaction network, gene ontology, and KEGG pathway enrichment scores. Biochim. Biophys. Acta, Proteins Proteomics ***1871***, 140889.
5. Unsal, S., Atas, H., Albayrak, M., et al. (2022). Learning functional properties of proteins with language models. Nat. Mach. Intell. ***4***, 227–245.
6. Huang, B., Xu, Y., Hu, X., et al. (2022). A backbone-centred energy function of neural networks for protein design. Nature ***602***, 523–528.
7. Lutz, I.D., Wang, S., Norn, C., et al. (2023). Top-down design of protein architectures with reinforcement learning. Science (New York, N.Y.) ***380***, 266–273.
8. Howarth, M. (2015). Say it with proteins: an alphabet of crystal structures. Nat. Struct. Mol. Biol. ***22***, 349.
9. Madani, A., Krause, B., Greene, E.R., et al. (2023). Large language models generate functional protein sequences across diverse families. Nat. Biotechnol. ***10***, 1038.
10. Russ, W.P., Figliuzzi, M., Stocker, C., et al. (2020). An evolution-based model for designing chorismate mutase enzymes. Science (New York, N.Y.) ***369***, 440–445.

ACKNOWLEDGMENTS

This work was supported by Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200, XDA26040304), National Key R&D Program of China (2022YFF1203202, 2018YFC2000205), and Self-supporting Program of Guangzhou Laboratory (SRPG22-007).

DECLARATION OF INTERESTS

The authors declare no competing interests.