Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080120
ISSN 2371-8838 Vol. 8 Num. 1

Deep Learning-Driven Protein Design

Ao You[a,*], Xiaobing Xu[b]

School of Information, Yunnan Normal University, Kunming, China
[a]2123100055@ynnu.edu.cnl, [b]3188944206@qq.com
[*]Corresponding author

Keywords: Deep learning; Protein design; Language model; Generative model; Protein sequence; Protein structure

Abstract: Protein, the material basis of all living organisms and the primary carrier of life activities, plays a pivotal role in regulating various physiological functions. Composed of specific amino acid sequences, proteins can fold into distinct structures, enabling them to perform diverse functions such as biocatalysis, metabolic regulation, immune defense, transport, and storage. The design of novel proteins with targeted biological functions is a central task in protein engineering, with broad applications in synthetic biology and drug research. This paper provides a comprehensive overview of the recent advancements in deep learning-assisted protein design research. It primarily introduces related language models and generative models, and discusses the associated research and existing challenges from both sequence and structural perspectives. Finally, it offers a forward-looking perspective on the future development of deep learning-assisted protein design research.

1. Introduction

In the early 1990s, Chen and his colleagues [1] pioneered the approach of Directed Evolution for the creation of new and improved enzymes. As the focus on High-throughput screening has increased, the equipment and detection instruments used for high-throughput drug screening have seen significant advancements [2-3]. However, due to the vast sequence space, functional mutations need to be screened from thousands of proteins. This process still requires a lengthy screening cycle, substantial manpower, and the unavoidable and non-standardizable operational errors between personnel. The continuous progress in high-throughput sequencing technologies has provided unprecedented data on natural sequence diversity. Therefore, how to circumvent the long research and development cycle and more effectively explore the secrets of protein sequence evolution has become a research direction of interest for many researchers.

Advancements in high-performance computing devices have enabled deep learning models to process vast amounts of data. In recent years, a plethora of deep learning models have been developed in the fields of natural language processing and computer vision, leveraging massive datasets [4-6]. The evolution of these technologies has significantly propelled the progress of artificial intelligence. Similarly, interdisciplinary researchers have started utilizing deep learning methods to model extensive biological datasets, thereby advancing biology. Deep learning, a crucial research branch in machine learning, can be implemented using various architectures. Each layer in

deep learning gradually extracts features and passes them to the subsequent layer. By processing the input of each layer, higher-order features in the data are extracted. The backpropagation algorithm is employed to modify the internal parameters, thereby discovering complex structures in large datasets. Deep learning can be categorized into supervised and unsupervised learning, depending on whether the input data is labeled. Deep learning can transform complex and voluminous unstructured data into abstract and high-level representations through neural networks. The advantage is that features can be reused, and as the number of layers increases, more abstract features can be obtained. Thus, deep learning exhibits greater power and flexibility.

2. Models based on data learning methods

The objective of protein design is to create amino acid sequences that can fold into a specific structure and perform a particular function. With the exponential increase in protein sequence data, traditional methods have become inadequate for protein sequence generation. Although machine learning methods have made significant strides in the field of protein design, early work primarily focused on training discriminative models to guide protein design, which still presents many challenges. Generative models can address some of the issues inherent in existing machine learning models and hold substantial research and application potential in the field of protein sequence design. Currently, the abundance of protein sequences in databases provides ample data for training a protein language model. Our experiments demonstrate that language models can be applied to a variety of protein understanding and design tasks, and significant progress has been made in this area. Optimization in the protein sequence space is extremely challenging due to the large, discrete, and unstructured nature of the search space. Generative modeling of protein design aims to model the data distribution, with the key being to understand and control the biophysical properties learned by the model to generate new samples with properties similar to those on which the model was trained.

2.1. Generative Language Models

Over the course of several decades, Natural Language Processing (NLP) technology has evolved to the point where it can autonomously learn from a vast amount of unlabeled texts. This capability allows it to effectively capture textual information, and it has been extensively applied in various fields such as question answering, machine translation, sentiment analysis, and speech recognition. It has also been demonstrated that it is feasible to transfer the models and techniques associated with NLP to study the function of protein sequences with a large volume of data. In recent years, pre-trained language models have been increasingly utilized in protein engineering to enhance our understanding and interpretation of the functional information conveyed by protein sequences.

Large language models have the capacity to learn diverse information from sequences, demonstrating robustness and generalizability. As illustrated in Table 1, protein language models tailored for various tasks typically necessitate extensive data for training. The ESM-1b[7], a high-capacity Transformer language model, assimilates biological properties from 86 billion amino acids across 250 million protein sequences. It can discern the secondary and spatial structures of proteins, aligning with organizational principles from physicochemical properties to long-range homology. ProGen[8], a protein generation language model, harnesses roughly 280 million protein sequences to produce evolutionarily diverse sequences through unsupervised generation, categorized by classification and keyword labels. Elnaggar[9] et al. developed two autoregressive language models and two autoencoder models trained on 80 billion amino acids from 200 million protein sequences, alongside ProtTrans, a language model trained on 393 billion amino acids from 2.1 billion protein sequences. These unsupervised language models have been shown to capture fundamental

biophysical characteristics of proteins, underscoring the benefits of scaling up language models with more extensive data support. The UniRep[10] model accurately predicts the stability of both native and newly designed proteins by modeling unlabeled amino acid sequences and distilling essential protein features into statistically sound representations grounded in semantics, structure, evolution, and biophysics. Transfer learning leverages a vast corpus of unlabeled protein sequences for pre-training, extracting general protein features and representations, which are then fine-tuned using a limited set of labeled data to tailor the model for specific downstream tasks. The TAPE[11] model assesses the embedding efficacy of pre-trained language models across five tasks, including structure prediction, remote homology detection, and protein engineering, revealing that no single model excels in all areas. Pre-trained language models, when applied to extensive and varied protein sequence databases, can predict protein function from experimental measurements without additional supervision and are readily deployable for a spectrum of protein comprehension and design endeavors. Despite many protein language models capturing the general context of protein sequences, the vast protein landscape remains incomplete, posing ongoing challenges for numerous specific proteins in development.

2.2. Deep Generative Models

Employing deep generative models to assimilate evolutionary traits from established functional protein sequences enables the generation of innovative protein sequences. This approach facilitates the exploration of previously untapped functional sequence diversity. Consequently, it diminishes the necessity to empirically test an extensive array of non-functional protein sequence variants, thereby optimizing the efficiency of protein engineering endeavors.

Deep generative models are adept at learning the joint probability distribution of sample data, capturing the essence of data distribution, addressing hidden variable samples, and synthesizing new samples that reflect the characteristics of the training data. Anand et al.[12] introduced an innovative method for the generation and reconstruction of 3D structures using deep generative models, employing Generative Adversarial Networks (GANs) to create novel protein structures. This trained model also has the capability to predict the missing segments of damaged protein structures. Greener et al.[13] utilized Conditional Variational Autoencoders (CVAE) to generate protein sequences with specific desired properties, such as introducing potential copper and calcium binding sites into proteins that typically do not bind metals. Shin et al. [14] developed an autoregressive generative model that leverages the information inherent in natural sequences to comprehend the constraints on amino acid positions without the need for sequence alignment. This model employs autoregressive likelihood to craft and design the complementarity-determining regions of antibodies. Repecka et al.[15] crafted a novel approach based on self-attention mechanisms.

Our variant of the generative adversarial network, ProteinGAN, directly assimilates the evolutionary relationships and the diversity of natural protein sequences from the intricate multidimensional space of amino acid sequences, generating a plethora of new sequence variants that retain natural physical properties. Xian et al.[16] proposed a conditional generative model to address the scarcity of labeled training data, merging the strengths of VAE and GAN to discern the edge feature distribution of unlabeled images via an unconditional discriminator. Furthermore, we demonstrate the interpretability of the learned features by reconverting them into pixel space.

3. Models based on data types

Deep learning utilizes statistical methods to construct models that mirror the complexities of the real world, harnessing the power of extensive datasets. In bioinformatics, particularly in protein

sequence generation research, deep learning has demonstrated its distinctive worth and potential. Technologies for designing protein sequences with deep learning are primarily categorized into two strategies: sequence-based design [17] and structure-based design[18].

3.1. Sequence-based design models

Sequence-based design methods are centered on the direct generation of proteins' primary structures—their amino acid sequences. These methods utilize deep neural networks to analyze and learn from vast datasets of known protein sequences, thereby discerning the intrinsic rules and patterns that govern sequence relationships. Through this sophisticated learning process, the models can navigate and interpolate within a meticulously crafted latent space, enabling the generation of novel protein sequences.

Biswas et al.[19] have mastered the art of learning natural latent representations by mining the vast landscape of protein sequences, utilizing a mere 24 functionally analyzed mutant sequences to craft an accurate virtual fitness landscape. Riesselman et al.[20], tapping into the latest strides in natural language processing and speech synthesis, have forged a generative deep neural network with an autoregressive model for biological sequences. This model, built on a residual causal dilation convolutional neural network architecture, adeptly captures functional constraints without depending on explicit alignment structures. Ding et al.[21] harnessed the distribution of family sequences within latent space to decode the protein fitness landscape, offering predictions on protein mutation stability and underscoring the pivotal role of stability in protein evolution. Their work reveals that points sharing similar adaptive landscapes cluster near the latent space sequence distribution, enabling the generation of new variant sequences via the VAE model decoder. Hawkins-Hooker et al.[22] engineered distinct VAE models for both unaligned and aligned sequences, demonstrating that those trained on multiple sequence alignment data more convincingly replicate the statistical nuances of structural and functional constraints that have been preserved throughout evolutionary history. Russ et al.[23] delineate a methodology to infer protein-specifying constraints solely from evolutionary sequence data, enabling the design and synthesis of gene libraries and their subsequent in vivo activity assessment through quantitative complementation analysis. These sequence-based statistical models prove to be robust tools, adequately specifying proteins and unlocking a vast expanse of functional sequences.

3.2. Structure-based design models

Structure-based design is predicated on the three-dimensional architecture of proteins. This method employs deep learning models to decipher the intricate interplay between amino acid sequences and their resultant spatial conformations. By modeling these relationships, scientists are able to engineer new protein structures that are theoretically robust and may also manifest unprecedented functionalities.

In 2013, De et al. posited that coevolution plays a crucial role in evolution, often leading to coordinated changes among regulatory proteins that help sustain the integrity of ecological and molecular networks. Computational methods grounded in coevolutionary principles have become instrumental in analyzing and predicting protein structure, function, and interactions. In 2015, Braun et al. combined evolutionary insights with iterative sampling strategies to enhance the accuracy of protein structure predictions. The prediction of protein residue contacts, as improved by Adhikari et al. in 2018 using two-level deep convolutional neural networks, provides invaluable insights for protein structure prediction, allowing for the simultaneous prediction of all contacts from a protein's comprehensive input information.

The CASP (Critical Assessment of Protein Structure Prediction) competition, convened by the

protein structure prediction scientific community, serves as a benchmark for the field, with the winners' level reflecting the pinnacle of global structure prediction capabilities. In 2019, Li et al. utilized a deep residual neural network to merge multiple original co-evolutionary features for contact map prediction in CASP13, demonstrating the robustness of the end-to-end training pipeline, attributed to sensitive MSA construction and sophisticated co-evolutionary feature integration strategies. AlphaFold emerged as a standout in CASP14 in 2020, with nearly two-thirds of its predictions achieving medium to low resolution experimental accuracy, nearly resolving the challenge of single-domain protein fold prediction. Subsequently, a research team led by David Baker introduced RoseTTAFold, a tool whose performance rivals that of AlphaFold. In 2021, Li et al. advanced the prediction of protein inter-residue contacts and distances by synergizing complementary coevolution features with deep residual networks in CASP14, indicating its potential to provide reliable distance metrics for ab initio protein folding.

4. Conclusions

The expansive compatibility of protein language models, coupled with the targeted data modeling by generative models, has significantly expedited the exploration of novel proteins. Leveraging the power of big data and deep learning, researchers are able to diminish reliance on domain-specific knowledge, bypass non-critical constraints, and delve into the discovery of potential new protein sequences that bear resemblance to actual proteins. While large and varied datasets of protein sequences are replete with information, there remains considerable potential for deep learning techniques to more precisely distill structural details into sequence data. Currently, the triumph of many models is heavily reliant on the availability of extensive homologous sequences. Consequently, datasets with limited samples present an ongoing challenge and an area ripe for future research and development

References

[1] CHEN K, ARNOLD F. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide [J]. PNAS, 1993(90):5618-5622.

[2] BLEICHER K H, BHM H J, MULLER K, et al. Hit and lead generation: beyond high-throughput screening [J].Nature reviews Drug discovery, 2003, 2(5): 369-378.

[3] MACARRON R, BANKS M N, BOJANIC D, et al. Impact of high-throughput screening in biomedical research [J].Nature reviews Drug discovery, 2011, 10(3):188-195.

[4] WU Z, JOHNSTON K E, ARNOLD F H, et al. Protein sequence design with deep generative models[J].Current opinion in chemical biology, 2021(65): 18-27.

[5] HIRANUMA N, PARK H, BAEK M, et al. Improved protein structure refinement guided by deep learning based accuracy estimation [J].Nature communications, 2021, 12(1):1340.

[6] DING W, NAKAI K, GONG H. Protein design via deep learning [J].Briefings in bioinformatics, 2022, 23(3): bbac102.

[7] RIVES A, GOYAL S, MEIER J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[J].bioRxiv,2019(10): 622803.

[8] MADANI A, MCCANN B, NAIK N, et al. Progen: Language modeling for protein generation [J].arXiv preprint arXiv, 2004(3497):2020.

[9] ELNAGGAR A, HEINZINGER M, DALLAGO C, et al. ProtTrans: Towards cracking the language of Life's code through self-supervised deep learning and high performance computing[J].arXiv preprint arXiv, 2007(06225).

[10] ALLEY E C, KHIMULYA G, BISWAS S, et al. Unified rational protein engineering with sequence-based deep representation learning[J].Nature methods, 2019, 16(12): 1315-1322.

[11] RAO R, BHATTACHARYA N, THOMAS N, et al. Evaluating protein transfer learning with TAPE[J].Advances in neural information processing systems, 2019:32.

[12] ANAND N, HUANG P. Generative modeling for protein structures [J].Advances in neural information processing systems, 2018:31.

[13] GREENER J G, MOFFAT L, JONES D T. Design of metallo proteins and novel protein folds using variational

autoencoders [J]. Scientific reports, 2018, 8(1): 16189.
[14] SHIN J E, RIESSELMAN A J, KOLLASCH A W, et al. Protein design and variant prediction using autoregressive generative models [J].Nature communications, 2021, 12(1): 2403.
[15] REPECKA, DONATAS. "Expanding functional protein sequence spaces using generative adversarial networks." [J]. Nature Machine Intelligence, 2021(4): 324-333.
[16] XIAN Y, SHARMA S, SCHIELE B, et al. f-vaegan-d2: A feature generating framework for any-shot learning[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019:10275-10284.
[17] Z. Wu, K. E. Johnston, F. H. Arnold, and K. K. Yang, "Protein sequence design with deep generative models," Curr. Opin. Chem. Biol., vol. 65, pp. 18–27, 2021.
[18] S. Ovchinnikov and P.-S. Huang, "Structure-based protein design with deep learning," Curr. Opin. Chem. Biol., vol. 65, pp. 136–144, 2021.
[19] BISWAS S, KHIMULYA G, ALLEY E C, et al. Low-N protein engineering with data-efficient deep learning[J]. Nature methods, 2021, 18(4): 389-396.
[20] RIESSELMAN A, SHIN J E, KOLLASCH A, et al. Accelerating protein design using autoregressive generative models [J]. BioRxiv, 2019: 757252.
[21] DING X, ZOU Z, BROOKS III C L. Deciphering protein evolution and fitness landscapes with latent space models [J].Nature communications, 2019, 10(1): 5644.
[22] HAWKINS-HOOKER A, DEPARDIEU F, BAUR S, et al. Generating functional protein variants with variational autoencod-ers[J].PLoS computational biology, 2021, 17(2): e1008736.
[23] RUSS W P, FIGLIUZZI M, STOCKER C, et al. An evolution based model for designing chorismate mutase enzymes [J]. Science, 2020, 369(6502): 440-445.