OXFORD

Briefings in Bioinformatics, 2022, **23**(3), 1–16
https://doi.org/10.1093/bib/bbac102
Review

Protein design via deep learning

Wenze Ding, Kenta Nakai and Haipeng Gong

Corresponding authors: Wenze Ding, School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China. E-mail: dwz@nuist.edu.cn; Haipeng Gong, MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China. E-mail: hgong@tsinghua.edu.cn

Abstract

Proteins with desired functions and properties are important in fields like nanotechnology and biomedicine. *De novo* protein design enables the production of previously unseen proteins from the ground up and is believed as a key point for handling real social challenges. Recent introduction of deep learning into design methods exhibits a transformative influence and is expected to represent a promising and exciting future direction. In this review, we retrospect the major aspects of current advances in deep-learning-based design procedures and illustrate their novelty in comparison with conventional knowledge-based approaches through noticeable cases. We not only describe deep learning developments in structure-based protein design and direct sequence design, but also highlight recent applications of deep reinforcement learning in protein design. The future perspectives on design goals, challenges and opportunities are also comprehensively discussed.

Keywords: protein design, deep learning, deep reinforcement learning, protein structure, protein sequence

Introduction

Among all molecules in our sophisticated and wonderful world, proteins that participate in most biochemical reactions have been under the spotlight of fundamental scientific researches as well as medical and industrial applications for decades. According to the 'central dogma,' the basic biological principle articulated by Francis Crick in 1958, proteins are the executive ends of information flow systems in living organisms, each performing one or a few specifically encoded functions that jointly define the corresponding organism in turn. A wide variety of native proteins such as nuclear proteins, membrane proteins, hemoproteins, lipoproteins, heat-shock proteins, contractile proteins, etc. manifest strikingly excellent properties compared with man-made machines, including extremely high efficiency, economy and precision in operation, self-assembly upon synthesis and so on. Considering their enormous quantity, fantastic quality and consequent pluripotency, protein materials have attracted extensive attentions since they could provide possible solutions for many serious social challenges.

Due to the strictly limited working environment and relatively short operation life, native proteins, however, cannot meet the surging demands of human beings satisfactorily. Furthermore, since native proteins are optimized gradually through millions of years of evolution under the selective pressure of nature, they in principle are unlikely to handle challenges arising from human society within hundreds of years. Therefore, artificial protein modification, and even one step further, the design of brand-new proteins from scratch emerges as the times require. Fortunately, protein design becomes technically possible with the long-drawn accumulation of knowledge from past biochemical and biophysical studies of proteins [1].

Many impressive achievements have been made through protein design over the past decade, which intensively impacted and promoted synthetic biology in both academia and industry. Advances in immune signaling [2, 3], targeted therapeutics [4, 5], sense-response systems [6], protein switches [7, 8], self-assembly materials [9, 10] and other fields not mentioned here have shown the exciting potential of utilizing proteins as functional and reproducible materials. In addition, these breakthroughs in protein design also expand our exploration and understanding of protein sequence, structure and function spaces. Taking sequence space as an example, since all native protein sequences originated from a few ancient accidental events and gradually

Wenze Ding is currently a research scientist in School of Artificial Intelligence and School of Future Technology, NUIST, interested in structural bioinformatics and protein design.
Kenta Nakai is a professor in the Institute of Medical Science, the University of Tokyo. His research interests include sequence analysis in molecular biology and bioinformatics.
Haipeng Gong is an associate professor in School of Life Science, Tsinghua University. He focuses on protein bioinformatics and molecular dynamics simulations.
Received: December 16, 2021. **Revised:** February 26, 2022. **Accepted:** March 1, 2022

evolved with haphazard mutation and oriented selective pressure, they exist in the sequence space in the form of sprinkling clusters called protein families instead of even dispersion. The properties and functions of protein sequences located in the vast remaining space would never be sampled by natural evolution within a limited time scale, which thus endows the great significance of protein design.

The earlier protein design approaches such as directed evolution [11, 12] and the following rational engineering [13, 14] mainly focus on the imitation and/or acceleration of natural evolutionary processes. Through rounds of mutation library construction and high-throughput screening, these methods could successfully obtain proteins with improved performance or even new functions by chance [15–18]. Nevertheless, these approaches always confront the tradeoff between assay fidelity and throughput, and more importantly, their explorations are still restricted around the corresponding initial native proteins. With the development of computational devices and algorithms, shortages mentioned above are gradually overcome by computer-assisted protein engineering, which avoids the relatively random mutation strategy and provides some definite design blueprints based on biophysical and biochemical principles of proteins. Among many computer-assisted protein engineering methodologies, *de novo* protein design aiming to generate new proteins not existing in nature has drawn the most attention [1]. With copious valuable achievements, *de novo* protein design was nominated as one of the top 10 annual breakthroughs by *Science* in 2016 [19].

Basically, the task of *de novo* protein design is to find new sequences targeted for desired functions. In practice, however, there are some impediments in the construction of direct mapping between the protein sequence and function spaces. For example, information encoded in a protein sequence is hard to extract from the target sequence alone, since it is simply a permutation or combination of 20 kinds of amino acid residues. Besides, different protein functions could barely be quantitatively articulated. Since proteins need to form particular tertiary structures to perform their specific functions and structures usually contain richer information, e.g. the Cartesian coordinates of atoms stored in PDB files, protein structures are perfect media for the bidirectional mapping between sequences and functions. In addition, massive protein structural data accumulated from previous researches, such as protein fold classification, consequent clustering and reaction mechanism information described by binding interfaces, catalytic centers and allosteric regulations, would also be extremely helpful. Thus, *de novo* protein design proceeds mainly in the structure-based manner.

Structure-based *de novo* protein design usually has three domains or stages, i.e. backbone generation, sequence fitness and candidate scoring, exemplified by Top 7 [20], the first globular protein that was designed without natural homologs, as well as other famous related works. Generally, a specific folding topology with predefined secondary structural elements and/or geometric constraints (e.g. inter-residue distances and orientations) is designed at the first step. Then, compatible peptide fragments are picked under the evaluation by sequence-independent energy functions and several sequence-structure optimization iterations are executed. During these iterations, rotamers are substituted randomly based on the energy functions, following the Metropolis-Hastings algorithm. After that, candidates are scored, rated and selected to generate the final design outputs [21].

Despite the significant achievements [22–24], these conventional approaches are mainly knowledge based, relying on physical principles and statistical rules [25]. With the plenty of data accumulated for protein sequence, structure and function as well as their relationships [26–28], research interests of protein design gradually converted towards data-driven methods in recent years [29]. Among them, deep learning techniques, which have revolutionized many other fields like natural language processing and computer vision [30], made the most significant impacts.

Deep learning offers the simplest and also the most general approximation and parameterization methodology for high-order statistics and potentials by enlarging the receptive field with the support of big data, and thus could be integrated into all domains of structure-based protein design for further improvements and even breakthroughs. Besides, deep learning also sheds a light on the direct protein sequence design for specific functions or properties without the medium of structures. In this review, we orient our discussion to advanced protein design approaches based on deep learning techniques, the benefits offered by them and the predictable trends. It is noteworthy that many other advances that hugely promoted protein design, exemplified by DNA synthesis, protein structure prediction and protein manufacture, would not be detailed here.

Briefings of deep learning techniques related to this review

In a nutshell, deep learning trains an artificial neural network or a combination of related networks to approximate complicated unknown functions in a high-dimensional abstract space. Artificial neurons or nodes with non-linear activations are connected by specific affine transformations with parameterized weights and biases, which are modified in each training step through the back propagation of gradients computed from losses, i.e. the differences between current network outputs and corresponding ground truths.

Discriminative models

Convolutional neural network (CNN) is one of the most successful deep network architectures working with data

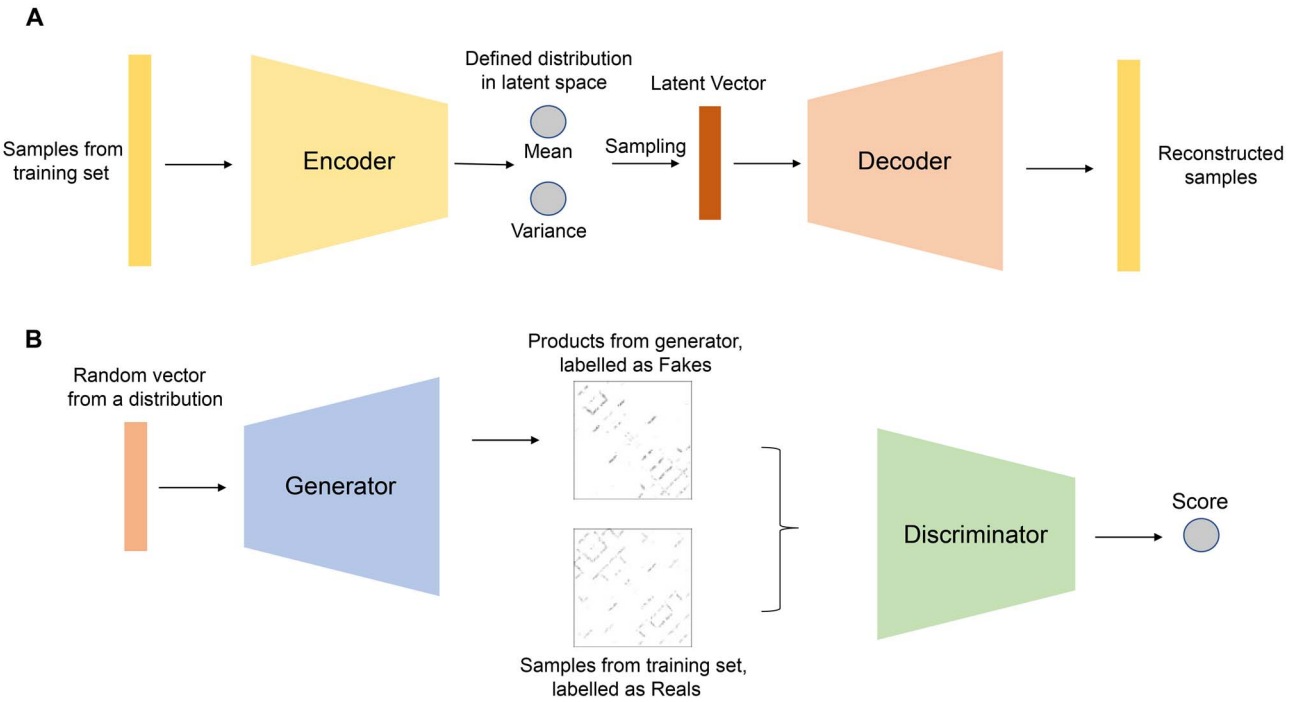

Figure 1. An illustration of the internal architectures of (**A**) VAEs and (**B**) GANs. Arrows represent corresponding dataflow.

that have typical grid-structured topology like ordinary pictures and protein inter-residue distance maps. There are two major operators in an ordinary CNN: convolution and pooling, where convolution is a special linear operation with pair-wised multiplication while pooling is a proportional down-sampling manner. The unique mechanism of CNN empowers it to overcome shortages of common deep feedforward networks with approaches like parameter sharing, sparse interaction and equivariant representation, etc. Besides, convolution also makes it possible to handle input data of variable sizes. Modern practical implementations of CNN often involve huge networks containing architectural variants and millions of units. Among them, ResNet was one of the most famous propellants that promoted the development of protein bioinformatics in the last decade [31, 32].

Recurrent neural network (RNN) is another classic network architecture suitable for processing sequential data like natural languages and protein sequences. The underlying idea to unfold recursive computation into a computational graph with repetitive structure naturally results in large-scale parameter sharing. Typically, RNN produces a single output according to the information of entire sequence extracted and stored in hidden units with recurrent connections at the current time step. Variants of RNN like long short-term memory (LSTM) and gated recurrent unit (GRU) also played an important role in natural language processing and bioinformatics during the past decade [33–35]. Recently, a novel network called transformer that contains encoder–decoder architecture and attention mechanism exhibited its superior capability for sequence processing [36]. Within the transformer network, multi-head attention module consisting of multiple self-attentions could capture correlations of amino acid residues among different dimensions, which makes it appropriate for representation learning of protein sequences [37].

Deep graph neural network (GNN) operates on graph, a non-Euclidean data structure, and focuses on problems like clustering, link prediction and node classification. GNN has been widely applied to knowledge graphs, social networks, drug discovery and protein bioinformatics [38–41]. There are many kinds of GNNs, such as convolutional GNN, recurrent GNN, graph autoencoder and so on, which mainly generalize the corresponding operations from Euclidean data with grid or sequential structure to graph data. For example, similar to CNN, convolutional GNN generates the representation of a node through aggregating the features of its neighbors within the graph to expand the receptive field of corresponding neuron.

Generative models

Unlike discriminative models widely used in protein researches that construct mappings from the space of the input data to that of the output label by maximizing the respective likelihood of samples, generative models such as generative adversarial networks (GANs) [42] and variational auto-encoders (VAEs) [43] try to capture the underlying data distribution of training set and sample brand new instances according to the learned distribution. It is noteworthy that the relationship between GANs and VAEs is complicated. Although these two frameworks have a large intersection, the VAE architecture could be trained for some models that GANs could not and vice versa.

As shown in Figure 1, GAN generally contains two main parts: a generator and a discriminator. The generator takes samples from a learned distribution while the

discriminator distinguishes the generator's outputs from real samples in training dataset. Essentially, the joint training procedure of GAN is a two-player game. Therefore, if both parts have sufficient model capacity and enough network training is implemented, the Nash equilibrium of this specific game would appear and the distribution learned by the generator would be identical with the one of training data. Meanwhile, the architecture of ordinary VAE is similar to classical encoder–decoder except the encoder estimates the mean and variance of a normal distribution instead of producing a latent variable directly. Combing the advantages of Bayesian method, VAE with its elegant mathematical foundation, simple structure as well as satisfactory training cost and model performance, gradually becomes one of the common options for generative models and influences bioinformatics a lot.

Deep reinforcement learning

Combining the great fitting power of deep learning for high-dimensional function and the ability of reinforcement learning to interact with surroundings in various situations, deep reinforcement learning techniques contributed to many areas including protein design [44–52]. Basically, deep reinforcement learning divides the world into two parts, an environment and an agent. Within every training step, the agent chooses an available action according to its own policy, which slightly changes the environment, and then receives feedbacks called rewards from the environment. The positive rewards encourage the agent to strengthen its policy, i.e. making the same choice in a similar situation for the next time, while the negative ones spur the agent to change its policy.

Deep learning in structure-based protein design

Structure-based protein design could be treated as the reverse process of protein structure prediction. For the latter, some potential structures should be modeled for a given sequence, while for the former, some feasible sequences should be optimized for a backbone with the designed topology (Figure 2). Protein homology plays an important role in protein structure prediction, providing massive evolutionary information for precise inferences. Recently, deep learning has revolutionized protein structure prediction in many ways, from early efforts in the protein inter-residue contact prediction and contact-assisted structure modeling [31, 53–57] to the subsequent accurate prediction of inter-residue geometric properties and geometric-constraint-based protein folding [32, 58–62]. Furthermore, attention networks with the most advanced end-to-end training procedure developed by Google DeepMind shocked the public in the 14th Critical Assessment of protein Structure Prediction (CASP) experiments by providing a wonderful solution for the structure prediction of single-domain proteins [63–65]. Deep learning techniques utilized in protein structure prediction like the convolutional neural networks could efficiently capture fold-level structural features from co-evolutionary information harbored in the multiple sequence alignment [66]. These successes deepened our understanding of the sequence–structure relationship for proteins, which is also the foundation of structure-based design, and provided a bunch of practical tools that could be directly used in design problems.

In addition to circumstantial improvement of protein design through advances in structure prediction, customized deep learning approaches also made considerable contributions to protein design directly nowadays. Novel network architectures, training procedures and data manipulations aiming to serve various design objectives in diverse design stages sprang up continuously, vigorously promoting the exploration of proteins. We will detail these novelties, illustrate the differences between these approaches and conventional knowledge-based ones, and articulate corresponding significance in the following sections.

Backbone sampling and generation

Functions and structures of proteins are closely correlated. A protein will perform its unique function only when its specific 3D structure is correctly folded. Hence, generating a backbone conformation under some particular design purposes becomes the first step in general protein design routines. Just like the immense space of protein sequence, the space of backbone structure is also extremely vast, with thousands of degrees of freedom even for small peptides. Nevertheless, designable backbones usually cluster into minute regions that disperse sparsely in the space [67], because protein domains stabilized by complicated atom-level forces like hydrogen bonds and hydrophobic interactions have to adopt exquisite shapes with well packed cores and properly exposed interfaces.

The earliest routines redesigned existing native protein structures to get possible backbones with improved structural stability and perhaps new functions [68, 69] or systematically sampled helical bundles [23, 70, 71] under the constraints of Crick's parameterization. Ensuing *de novo* design methods generated protein backbones mainly through the combination of fragment-assembly-based simulations and human intuition [20, 72–81], exemplified by the famous Top7 mentioned previously [20]. As shown in Table 1, modern deep-learning-based approaches trained generative models to either generate 2D inter-residue geometric feature maps of sampled backbones or directly output their atom coordinates.

GANs were used to generate protein inter-residue distance maps for the completion of corrupted structures [82–84]. This task aimed to infer plausible backbones of missing residues for the target protein, analogous to an image-inpainting problem, i.e. inpainting a large distance map with small size-fixed patches (Figure 3). For example, deep convolutional GAN (dcGAN) [85] was chosen to learn a mapping from a low-dimensional standard

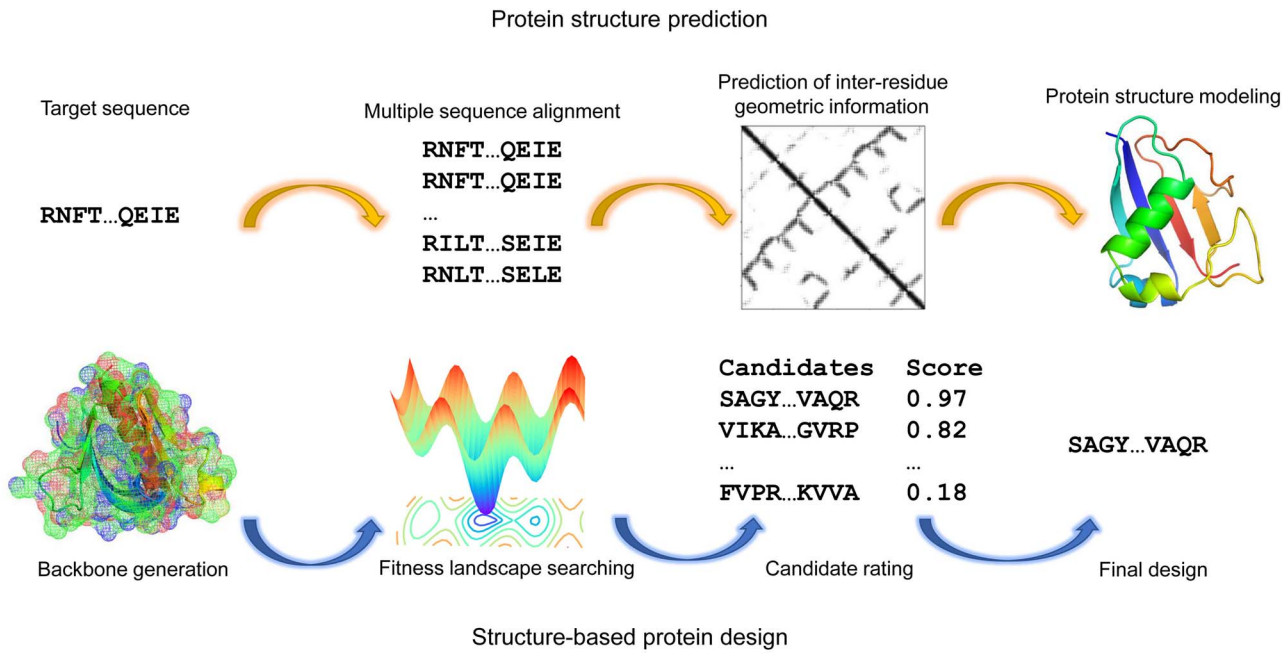

Figure 2. An illustration of two inverse processes, i.e. protein structure prediction (upper) and structure-based protein design (lower).

Table 1. Brief summary of recent researches focused on structure-based protein design

Reference order	Research objective	Data resource	Network architecture
[82]	Complete corrupted structures	Protein structures from PDB database	dcGAN
[91]	Hallucinate novel proteins through protein structure prediction networks	Completely arbitrary protein sequences with fixed length of 100 amino acids	trRosetta network within residue substitution step of a simulated annealing trajectory
[93]	Generate coordinates of immunoglobulin backbones	Antibody structures from AbDb database	VAE
[39]	Generate protein sequence with given geometric and amino acid constraints	Proteins extracted from UniProt database, sequence repository Gene3D	GNN
[110]	Optimize over protein sequences and structures simultaneously by backpropagating gradients through protein structure prediction networks	Proteins collected from a structure-refinement research (redundancy with trRosetta training set were reduced)	trRosetta network
[61]	Rate candidate predicted structures without explicit standards and answers	Known correct rankings	RankNet and LambdaRank

To maximize the usage of limited exhibition space in this paper, we only choose one research as representative from a bunch of researches with similar objectives or procedures.

normal distribution z to an unknown high-dimensional probability distribution in the space of protein inter-residue distance map with a fixed size [82]. After inpainting, backbone structures were obtained using either the alternating direction method of multipliers (ADMM) algorithm [86] to trace Cα positions with concrete coordinates or Rosetta [21] to sample fragments according to the generated distance constraints. Although satisfactory outcomes have been achieved by these works, some limitations still exist. For example, the distance maps generated via the dcGAN method mentioned above [82] were restricted to 16-, 64- and 128-residue fragments instead of arbitrary length for the intrinsic properties of dcGAN. This shortage, especially its incompetence to larger protein fragments, slashed its practicality. Meanwhile, VAEs that performed conditional generation through the introduction of a representative latent space were also shown to be very useful for protein backbone design [87–89]. With all these successful trials, the ability of generative models to produce protein backbones with multiple structural elements (e.g. secondary structures) has been validated and further related researches would surely acquire a greater depth in the coming future.

Deep neural networks originally trained for image recognition could be used to generate 'hallucinations' with a transformed style [90]. Similarly, information of protein sequence–structure relationships stored in billions of parameters in the powerful protein structure prediction networks could also be utilized inversely to generate new sequences and structures [91]. Completely random sequences of 100 residues were fed into trRosetta network [32], a well-performed predictor of protein inter-residue geometric properties based on sequence alignments, to derive the background

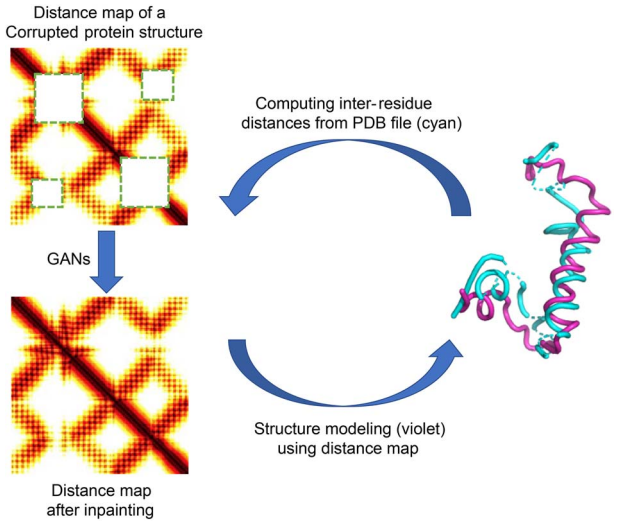

Figure 3. GANs are used as an inpainting tool to repair the inter-residue distance map for a corrupted protein structure. The missing part of the original corrupted distance map (upper) is highlighted with green dashed squares and the corresponding structure is represented in cyan (dotted line for corruptions). The distance map is repaired (lower) and the structure translated from it is represented in violet.

inter-residue distance distributions. Then, a Monte Carlo simulated annealing trajectory was produced for each initial random sequence to iteratively optimize this sequence and get compatible structures. Within the trajectory, a random single residue substitution was initiated at an arbitrary position and the distance distribution map of this mutated sequence would be immediately predicted by trRosetta for every time step. This substitution would be accepted only if the Kullback–Leibler divergence between distance distributions of the new sequence and corresponding background satisfied the Metropolis criterion. Through this procedure, diverse sequences and designable structures not observed in nature were generated. Subsequent *in vitro* synthesis showed that these 'protein hallucinations' were monomeric and stable, possessing designed structural elements. Furthermore, although constructed through trRosetta [32], this hallucination approach could be easily extended to more advanced protein structure prediction networks like AlphaFold2 [63] and RoseTTAFold [65] to improve its 'hallucinating power.' The significance of this work is not limited in showing a feasible exploration for structure or sequence generation. More importantly, it also exhibits a more straightforward avenue to construct supporting scaffolds around predetermined activation sites for protein design, where structures are not required to be mapped out beforehand.

The translation from protein inter-residue geometric matrices to backbone coordinates could also be undertaken by approaches related to deep learning [32, 61, 92]. Some of them incorporated energy based optimization [32, 93] while others employed self-adaptive data screening [61]. It is notable that some researches skipped two-dimensional structural representations and generated backbones with 3D atom coordinates directly. For example, a VAE-based architecture modeled backbone flexibility of immunoglobulin proteins via catching the related structure distribution, compressing it into a low-dimensional latent space and interpolating that space to sample structures with predefined complementarity determining regions (CDRs) [93].

Sequence design in protein fitness landscape

Almost all information of a protein is encoded in its sequence. However, inferring possible sequences for a predefined structure with desired function from the vast multidimensional sequence space termed as the protein fitness landscape [94] is extremely struggling and impossible to be handled with brute force, considering the countless permutations formed by the 20 usual proteinogenic amino acids [95].

Generally, protein fitness landscape searching methods cluster residue side-chain conformations as different rotamers [96], abstract the sequence optimization of a given backbone to a discrete energy minimization problem and then search combinations of rotamers around the global minimum [97]. The energy optimization process is analogous to mountain hiking (minimizing energy equivalent to maximizing its opposite), during which a hiker tries to arrive at the global optimal point through a meandering route consisting of multiple tiny trail steps. Despite the previous achievements, traditional approaches confronted restrictions like the powerlessness in multi-body interaction design and the excessive homology of outputs. Although similar in general, the learning process of deep neural network differs from conventional energy minimization in several ways. Thus, deep learning with its intrinsic advantages and training techniques accumulated in earlier researches could substantially mitigate the limitations of regular procedures either by replacing the entire optimization routine or by eliciting local amelioration within their frameworks.

Deterministic approaches could solve the fitness problem accurately for small backbones [98] but become powerless for large ones due to the exponential increase of computational complexity. Statistical sampling methods, exemplified by Monte Carlo simulations, have been used to solve this dilemma and could achieve acceptable approximations in practice [99]. Because the backbone energy evaluated by existing force fields is highly sensitive to conformational changes, backbone flexibility is usually considered in these methods by simultaneously optimizing rotamers and the corresponding local structures [100–103]. Besides, the hydrogen-bond network is also an important point that should be carefully attended to in sequence optimization procedures [104].

Deep learning approaches excel at optimizing the joint probability of residues under the given backbone constraints. Thus, applying them to sequence fitness problem could effectively alleviate or even address the challenges in conventional methods. In analogy to Sudoku puzzles, a deep GNN called ProteinSolver

was proposed by converting the sequence fitness for a predetermined backbone into a constraint satisfaction problem, where amino acids were assigned such that the atom-level inter-residue forces could be compatible with the given fold [39]. Through training on more than 70 million sequences corresponding to over 80 thousand structures, GNN elucidated the rules governing these constraints by inferring hitherto hidden patterns. Unlike other works in this topic that mainly used computational metrics to validate the accuracy and quality of their designs, *in vitro* validations of ProteinSolver by circular dichroism experiments testified its capability to fit protein sequences. It is noteworthy that ProteinSolver was only trained and tested with the constraints derived from existing proteins, and thus, its ability to sample reasonable sequences of novel proteins still needs further validation.

Another method based on conditional generative model and graph representation also improved the reliability and computational speed of sequence fitness compared with traditional methods like Rosetta [38]. More specifically, in this work, a spatial k-nearest neighbor graph was used in a multi-head self-attention encoder to develop the backbone representation independent of sequence. Then, conditioned on previously generated s amino acids and the given structure, the $(s+1)$ th residue was predicted autoregressively by a decoder, similar to common procedures in language modeling. Other deep-learning-based methods constructed their networks with various architectures including auto-encoders [105], 3D convolutional neural networks [106], DenseNets [107] and GANs [84] to predict sequence probability profiles from a given backbone structure. Since these data-driven approaches are capable of assimilating co-evolutionary information from protein sequence databases, integrating high-dimensional hints, catching the inconspicuous internal patterns and deducing the most possible solutions, protein sequence profiles generated by them usually exhibit better agreement with the natural molecular evolution than those profiles sampled by conventional knowledge-based methods lacking the help of deep learning.

Deep learning also contributes to the energy evaluation process of protein fitness landscape searching. In comparison to traditional knowledge-based energy functions that are typically combinations of statistical and empirical potential terms [21, 97, 108], deep learning models could provide a more general and more accurate description of the multidimensional potential functions in the real world. A 3D convolutional neural network was trained in an autoregressive manner to learn the distribution of sequences conditioned on a predetermined backbone directly from the protein structure data [109]. In absence of any human-specified priors, potentials learned by this network could precisely predict side chain conformations without using any conventional forcefields. *In vitro* experimental data, especially the high-resolution crystal structures of two designed TM-barrel proteins, validated the design capability of this network and corresponding structural agreements. Compared with the classical molecular mechanics force fields with great complexity and cost, this data-driven method only needed a few hours for training, which exhibited its practical applicability and huge potentiality.

In addition, networks originally constructed for protein structure prediction could also be repurposed for sequence design by energy landscape optimization. With gradients backpropagated from the predefined structures to input protein sequences through the trRosetta network [32], sequences and structures could be optimized simultaneously [110]. This research hints that future combination of the low-resolution trRosetta model that considers the full conformational landscape and the high-resolution Rosetta model that is good at single point energy estimation would further improve protein design methodologies.

Scoring function and candidate rating

Usually, iterations of sequence–structure optimization would produce a set of candidate sequences. To lighten the burden of downstream laboratorial synthesis, it is necessary to select a small subset of candidates that have the largest probabilities for the intended protein properties and functions. A typical approach is to rank all candidates by scoring functions and only retain the top k. One of the most frequently used scoring functions is the potential energy mentioned above, since the chosen sequence should be able to fold into the correct topology with acceptable stability. Candidate rating is thus often simplified as identifying sequence–structure pairs with the lowest energies. Some summaries have been articulated in the last two sections since this step has a close relationship with previous steps and many researches integrate them all together.

Scoring functions in the Rosetta program range from statistical potentials established using Bayesian methods [111] to complicated modern force fields [112]. Thus, rating systems of many protein design routines are derived from Rosetta. Meanwhile, a distinct approach introduces deep ranking networks called RankNet and LambdaRank [113] in recommendation systems for candidate rating [61]. Instead of directly optimizing potential items for precise energy estimation, these networks update themselves according to the discrete ranking fitness, i.e. difference between current order ranked by the network and the supposed one. Although this work is originally proposed to address the protein structure prediction problem, the underlying fundamental concept could be easily generalized to protein design.

Deep learning in direct sequence design

As described above, the major task of protein design is to find sequences capable of stably exhibiting desired properties and conducting expected functions. Besides,

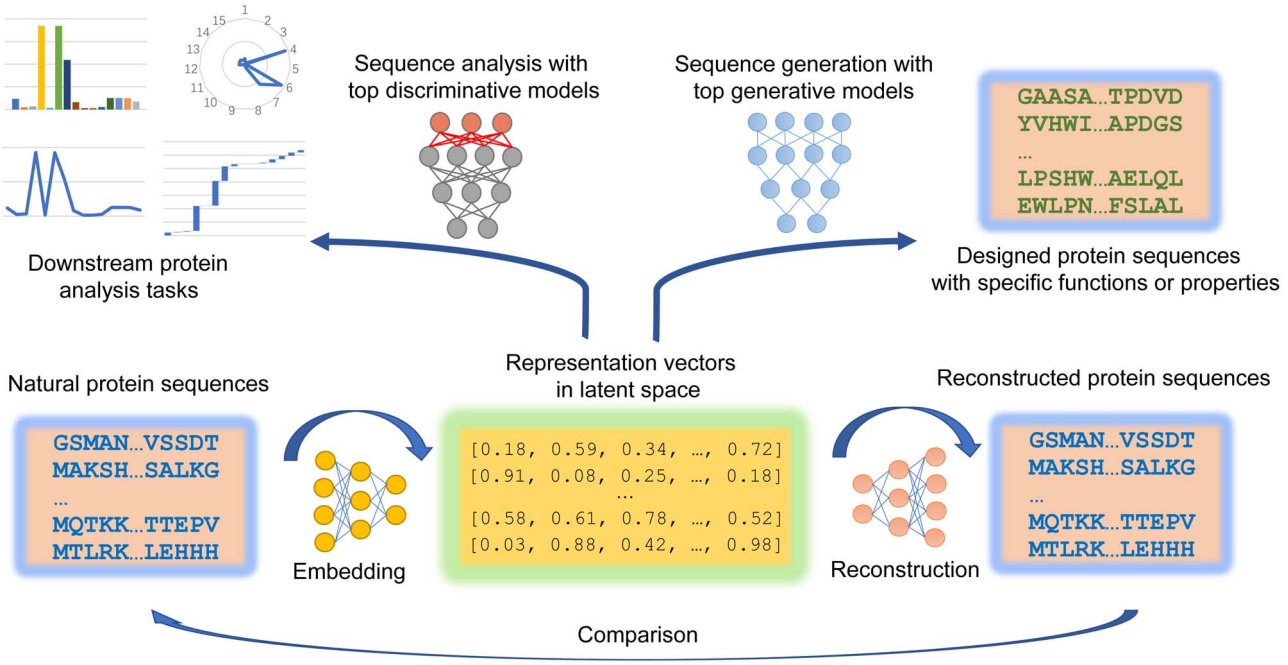

Figure 4. An illustration of protein representation learning, direct protein sequence design and related downstream protein analysis applications. Protein representations with fundamental features are obtained through protein language models (bottom). In combination with different kinds of top models, these representation vectors could be used for either protein sequence design or other analysis tasks (top).

longer information pathway with more transit points would generally introduce unnecessary transformation and transmission of data, which might cause larger signal deviations. Thus, in principle, directly mapping the spaces of protein sequence and function seems to be advantageous over design procedures that need predetermined structural topologies as intermedia. More importantly, due to advances in sequencing technology, the accumulation speed of protein sequence data is much faster than its structural counterpart, especially after the introduction of metagenomics [114]. Tremendous number of unlabeled sequences in combination with the powerful capability of deep learning for feature extraction, pattern recognition and objective generation make it possible and valuable to directly explore the sequence space and improve the protein design paradigm.

Different from protein fitness landscape searching for a given backbone, direct sequence design learns a meaningful distribution of sequence representation in a latent space and generates sequences in real space according to speculative representations derived from the learned distribution (Figure 4). Therefore, generative models are more widely used in this area compared with discriminative ones (as exhibited in Table 2). In this section, we will focus on two major aspects of direct protein sequence design with concrete cases to look through the past achievements and anticipate the future trends.

Representation learning

Although deep learning has shown huge success in many sub-fields of protein bioinformatics, there are still two major obstacles impeding its further development. The first one is the expensive cost of protein characterization, which leads to the data scarcity of sequence-label pairs for the training of deep neural networks. The second one is the lack of method generalization, since most domain-specific deep learning methods have not sufficiently exploited the fundamental features of protein sequences and thus are hard to be transferred from one problem to another through simple fine-tuning. One possible solution to overcome these obstacles is the representation learning using protein language models. Protein sequence and natural language both have internal long-range dependencies of distant contexts. Thus, inspired by natural language processing [115], protein language models treat a complete sequence as a paragraph or a sentence and the amino acids within it as single words [116, 117]. Through supervised or unsupervised training, a dictionary of word vectors, i.e. amino acid embeddings, could be optimized and the representation of a protein sequence with its fundamental features could be inferred in a latent space.

A method called unified representation (UniRep) trained a multiplicative long–short-term-memory RNN (mLSTM RNN) [35] with 1900 hidden units to learn the fundamental representation of protein sequences and encode arbitrary sequences into length-fixed vectors [33]. UniRep was trained with approximately 24 million sequences from the UniRef50 database [118] and its self-supervised training procedure [119] utilized input sequences themselves as the corresponding labels. Specifically, it iterated through amino acids of a sequence sequentially and compared the true next residue with the one predicted by the model based on its dynamic summary of all previously visited residues. With this training procedure, the model of UniRep gradually

Table 2. Brief summary of recent researches focused on direct protein sequence design

Reference order	Research objective	Data resource	Network architecture
[33]	Extract fundamental features of unlabeled protein sequences into a statistical representation	Protein sequences from UniRef50 database	mLSTM RNN
[37]	Train a deep contextual protein language model to produce generalized features	Protein sequences from UniParc database	Transformer
[34]	Build precise virtual protein fitness landscape based on protein sequence representation	A few mutants of natural target protein and their functional characterizations	Single-layer linear regression model on the top of UniRep
[127]	Generate synthetic genes coding proteins with desirable functions or biophysical properties	Peptides with 5–50 residues from UniProt dataset	WGAN with an external feedback loop
[121]	Generate functional protein sequences by learning natural sequence diversity	Bacterial MDH sequences from UniProt dataset	Tailored GAN with temporal convolution and self-attention

To maximize the usage of limited exhibition space in this paper, we only choose one research as representative from a bunch of researches with similar objectives or procedures.

maximized the conditional probability of correct amino acid type for next residue and learned a progressively better protein sequence representation by adjusting its parameters and revising its hidden state construction manner. In the absent of any evolutionary, structural, physicochemical and other kinds of related data explicitly, representation vectors of protein sequences encoded by UniRep intrinsically contained the required information and thus could be easily clustered by these properties. When evaluated on a comprehensive set of critical protein engineering problems, UniRep with simple linear or non-linear models trained on the top of it showed generalizable and superior performance. Although the data mining ability of RNN architecture used by UniRep might be inferior compared with current popular ones in the field of nature language processing like transformer, the basic conceptions it came up with and the impressive extensions it showed still influenced following researches a lot.

Trained on 250 million sequences with breadth and diversity from the UniParc database [120], a deep transformer called ESM-1b also learned protein sequence representations with fundamental features [37]. The model consisted of 33 layers, having around 650 M parameters. It utilized another self-supervised strategy, masking language modeling objective, for its training. ESM-1b Transformer integrated residue contexts across the entire input protein sequence through many stacked self-attention modules. It constructed a complicated representation space for protein sequences. Representation vectors derived from this space carried distinguishable protein features of the corresponding sequences. For example, secondary and tertiary structural properties could be identified from the generated sequence representations. Superiority over other state-of-the-art input features across a wide range of applications like mutational effect prediction further testified its generalizability and advantages. Furthermore, with the rapid accumulation of protein sequence data and the usage of network architectures with higher complexity and capability, the future versions of ESM-1b were expected to have additional improvements in protein sequence representation. However, the training cost of such a huge protein language model would not be something that ordinary small research groups could afford and it would be meaningless to repeat the construction of these infrastructures for the whole academic community. Thus, the sharing spirit existed in this work and many other famous researches should be advocated and kept for a long time.

Other works focusing on representation learning adopted deep generative model architectures like GANs, VAEs and autoregressive ones. They compressed discrete protein sequences into a continuous latent space by capturing contextual information within these sequences [121–123]. For example, trained by sequences from the Swiss-Prot database, a VAE model called BioSeqVAE learned good sequence representations, which could be used as input features for multiple downstream applications [124]. Since different researches of representation learning generally use self-built datasets and have no unified evaluation process or standard, it is difficult for people to compare them and consider the accuracy and efficiency, advantages and disadvantages of each [125]. Hence, a work introduced a set of protein bioinformatics tasks with clear definitions, data and assessing metrics to construct a standard evaluation system for protein transfer learning [126]. This task set called tasks assessing protein embeddings (TAPE) contained five concrete problems within three major aspects: protein structure prediction, remote protein homolog detection and protein design. The authors also benchmarked several representation learning methods, of which the methodology could be easily generalized to recent works mentioned above.

Sequence generation

Representation learning has laid a solid foundation for sequence generation. By condensing, integrating and extracting fundamental features within sequence statistics, learned representations embody protein properties like function, structure, stability, dynamics, half-life, binding, etc. Therefore, in combination with downstream generative models or methods, proteins of desired functions but with unseen sequences could be

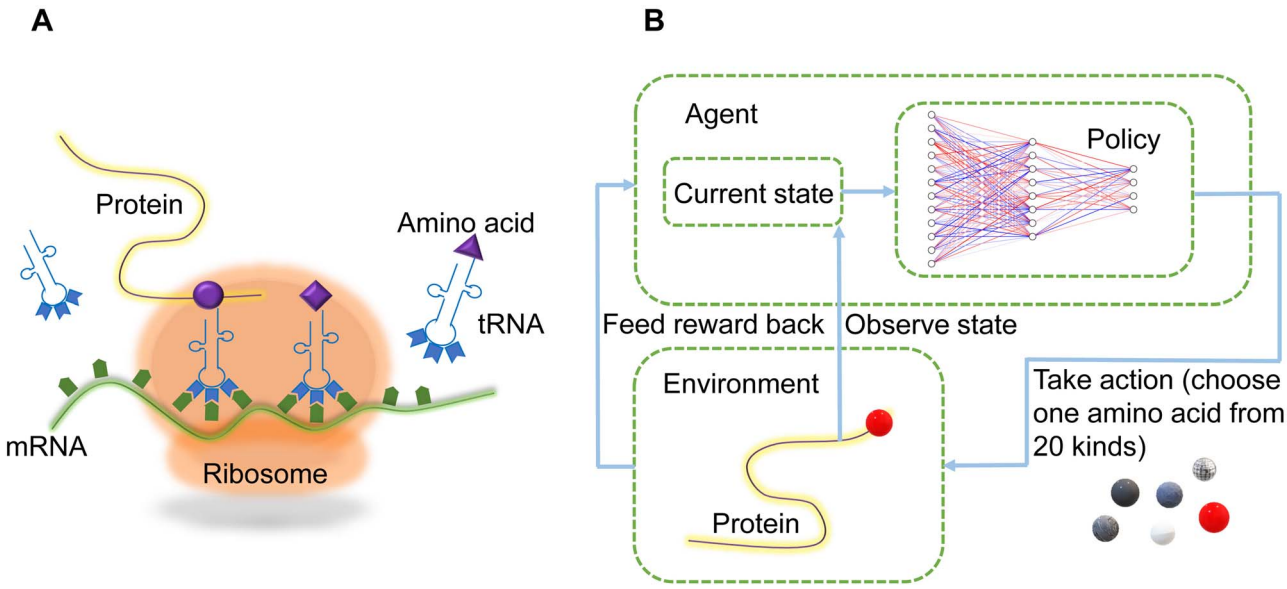

Figure 5. Deep-reinforcement-learning-based protein design is analogous to natural protein synthesis process. (**A**) An illustration of the natural protein synthesis process. (**B**) Protein sequence generation from left to right by deep reinforcement learning. The agent takes an available action (what kind of amino acid to pick in the next step) according to its policy conditioned on the current state.

generated in a high throughput manner. For example, a low-N protein engineering method [34] was reported based on representation learning of UniRep [33]. A simple supervised top model taking the sequence representations as input was trained on a limited number (as few as 24 sequences, and this is the source of its name 'low-N') of functionally assayed random mutants of the target protein to rate arbitrary sequences. Then, *in silico* directed evolution was executed through a Markov Chain Monte Carlo procedure on the surrogate fitness landscape provided by sequence representations and the rating model.

GANs also played an important role in direct protein sequence generation. A Wasserstein GAN (WGAN) [84] combined with a novel external feedback-loop mechanism (denoted as a function analyzer) was trained to generate DNA sequences encoding proteins [127]. The function analyzer could be in any form, differentiable or non-differentiable, as long as it took a sequence as input and output a score. The training procedure of this so-called FBGAN system contained two parts. Firstly, the WGAN was pretrained with general DNA sequences converted from protein sequences reversely to generate valid genes. Within every training step, sequences produced by the generator of WGAN were fed into the function analyzer to evaluate their related properties and those with scores exceeding a predetermined threshold were chosen to replace the oldest samples in the original training set of the discriminator. This feedback-loop mechanism finetuned the distribution mapping between the latent space and the real DNA sequence space for specific downstream optimization objectives. Successful applications on the generation of antimicrobial peptides [128] and helical proteins supported the good generalizability of this model. In addition, this unique network architecture and training procedure could be easily extended to other domains beyond genomics and protein sequences. However, there were also some compromises in this work beyond its success. For example, FBGAN focused on gene sequence generation though the research objective was protein sequence design, because gene sequences had clear codon structures to instruct start/stop positions and much simpler vocabulary (only four nucleotides) compared with proteins. Thus, direct generation of longer and more complex protein sequences would still be an important task for follow-up researches of FBGAN.

ProteinGAN was another GAN architecture constructed to expand functional protein sequence space [121]. Implemented with customized temporal convolutional network [129] and self-attention mechanism [130], ProteinGAN could not only learn useful sequence motifs and critical long-range inter-residue interactions simultaneously, but also concentrate on functional areas like catalytic centers. To validate its contribution to real protein engineering, ProteinGAN was trained on a family of bacterial malate dehydrogenase (MDH) enzymes. By uniformly interpolating the latent space, the model successfully generated 20 thousand protein sequences exhibiting sequence properties highly correlated with the latent dimensions, which supported its ability to capture the intrinsic features of native sequences and their inter-relationships. Among 55 generated sequences tested experimentally, 24% of them stably existed in physiological solutions with blatant catalytic activity, which further demonstrated its potential to generate new, diverse functional protein sequences.

Other direct sequence generative models adopted different architectures suitable for specific generation demands [87, 88, 122, 131]. For example, an attention-based transformer model was trained on the Swiss-Prot database to generate functional signal

peptide sequences and experimental tests proved its practicality [131]. As all roads lead to Rome, distinct networks with various customized training procedures all serve one similar goal: learning to sample diverse protein sequences that are previously unseen in nature and to enhance the likelihood of those satisfying desired criteria.

Design with deep reinforcement learning

Protein design approaches based on deep reinforcement learning are just like *in silico* simulations of natural protein synthesis processes (Figure 5). With the application of more advanced technologies, these methods can help us excavate more intrinsic principles of proteins and get more high-quality functional protein materials. For example, DyNA PPO [132] was such a deep reinforcement learning model based on proximal-policy optimization [133] for sequence design. The model generated sequences from left to right one amino acid after another, with the overall procedure regarded as a Markov decision process. Before the completion of sequence generation, the reward to the agent remained 0. At the end of each round, sequence fitness measurement given by a panel of machine learning models that tried to approximate surrogate fitness functions was taken as the final reward. DyNA PPO balanced the tradeoff in reward estimation by using a bunch of models to learn different aspects of the sequence fitness landscape but only using the most suitable one with sufficient accuracy to update its policy. Although its superiority has been shown in the large-scale benchmarking across several methods, the report of DyNA PPO did not exhibit any verification through wet lab experiments. Thus, its practicability still needs to be testified in future researches. Alternatively, reinforcement learning could be used to finetune some pre-trained generative models for protein design. For example, a RNN was tuned by a policy-based reinforcement learning approach to generate desirable compounds [134]. The most important inspiration from this research would be the attempt and success of decreasing the catastrophic forgetting risk [135], a common problem for protein generative models.

Conclusions and future perspectives

In the last decade, protein design has achieved great successes, helping mankind deal with social challenges in multiple facets. Examples could be found everywhere in our daily life, including designed small-molecule binding proteins that are used in *in vivo* biosensors [136, 137], designed biomedical inhibitors that aim to prevent viral infections [138], designed enzymes that have attractive catalytic efficiencies [139–141], designed highly symmetric self-assembly materials that endow vaccine applications with multivalent presentation of antigens [10, 142], etc. Recently, deep learning techniques have shown preliminary but impressive impacts to the field of protein design. Through their incredible power of extracting and integrating statistical patterns within existing protein data, artificial deep neural networks learn fundamental protein features, store them in billions of parameters and generalize them for inferences in different sub-fields.

However, roadblocks still stand in our path to routinely design arbitrary proteins using deep learning methods. For example, our understanding to protein folding mechanism, one of the most important and essential problems in bioinformatics and also the paramount theoretical principle of all kinds of protein design methods, is far from sufficiency. Many efforts combining deep learning, physical modeling and *in silico* simulations have been made in this area. Perhaps deep reinforcement learning trying to build policies and find possible trajectories from extended protein chains to well-folded structures would also be helpful.

Diverse and abundant well-annotated data are necessary for all fields adopting deep learning, just as the influence of ImageNet database [143] to the development of computer vision. However, for protein design with a specific objective, related data of protein functions and properties are usually not only scarce but also collected without unified and standard experimental conditions. The scarcity of training data would hinder the accurate design, consequently leading to the demand for additional experimental optimization. Although some databases exemplified by ProtaBank [144] have been constructed to alleviate this phenomenon, lots of efforts still need to be made. Another important direction to overcome this deficiency might be the few-shot learning [145, 146], and to our knowledge, related exploration has not been tried yet.

Scoring accuracy and computational speed of energy functions in protein design also need to be further improved, since they guide the optimization direction and are used repeatedly in every step. Compared with traditional potential terms, energy functions learned by deep neural networks evaluate designs more precisely but slowly. The adoption of more advanced and lightweight network architectures as well as knowledge distillation [147] and network pruning [148] may partially handle this dilemma. Another plight for both protein design and its reverse procedure, protein structure prediction, is that current approaches for optimization are usually adept in landscapes with only one minimum, while many proteins perform their functions and properties through structural transformation among different conformations. This requires deep learning methods to design proteins with multiple and distinct energy minima. Future researchers should attend to such complexity.

Another important and imminent assignment of deep-learning-based protein design is promoting its application scope. Many researches of this field focused on algorithm development and *in silico* evaluation with barely few experimental verifications and practical applications. Taking pharmacy and therapeutics as

an example, although conventional drug discovery methodologies concentrated on molecular dynamics simulations and molecular docking [40] have made great achievements, protein design approaches are gradually showing their impressive capability and promising future. There are many roadmaps involving protein design in this field, which aim at various diseases afflicting human beings. One possible procedure is designing a modular protein sensor-actuator switch where small ligands could directly change downstream transductions of corresponding cellular signal pathway by binding to the designed targets [6, 73, 149]. Another approach might be designing mimetics of natural immune proteins with augmented therapeutic affinity and activity but diminished immunogenicity and toxicity [2, 150, 151]. Besides, through treating short peptides (usually less than 50 residues) as small molecules and utilizing knowledge about protein–protein interactions (PPI) [41] instead of drug–target interactions (DTI), high-throughput protein design methods could be constructed for therapeutics with specific targets [4, 152]. In the context of current worldwide pandemic of COVID-19, protein design is especially important since designed mini-protein inhibitors of ACE2 receptor (coronavirus binder) have provided a good start for corresponding therapeutics [5, 153]. However, almost all above successes were achieved by traditional knowledge-based design methodologies. Getting out the *in-silico* limit and putting the advanced data-driven algorithms into effect should be another key point of future researches focusing on deep-learning-based protein design.

Many challenges confronting protein design could be ameliorated or even handled by combining deep learning efforts with complementary advances in conventional *de novo* methods, while others still await the development of new methodologies from the ground up. No matter which case it is, proteins are important gifts from nature to mankind, and with the blueprints glimpsed by deep learning, we could craft desired tools as we want to make our world a better place after iterations of trials and errors.

Key Points

- Recently, the introduction of deep learning has shown preliminary but transformative influence to the field of protein design.
- Deep learning could provide fast, high-throughput and precise *in silico* protein design methodologies.
- We retrospect current advances in deep-learning-based protein design procedures mainly within the past 2 years and illustrate their novelty, advantage and significance in comparison with traditional knowledge-based approaches through important milestones. We also comprehensively discuss the coming challenges and opportunities in the near future.
- This review could help people get familiar with this field and promote relevant researches.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This work has been supported by the National Natural Science Foundation of China (#32171243), the Beijing Advanced Innovation Center for Structural Biology, the Japan Society for the Promotion of Science (JSPS) KAKENHI (19H03213 and 18H0298) and the Startup Foundation for Introducing Talent of NUIST.

References

1. Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. *Nature* 2016;**537**:320–7.
2. Silva D-A, Yu S, Ulge UY, *et al.* De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* 2019;**565**:186–91.
3. Mohan K, Ueda G, Kim AR, *et al.* Topological control of cytokine receptor signaling induces differential effects in hematopoiesis. *Science* 2019;**364**:750.
4. Chevalier A, Silva D-A, Rocklin GJ, *et al.* Massively parallel de novo protein design for targeted therapeutics. *Nature* 2017;**550**:74–9.
5. Cao L, Goreshnik I, Coventry B, *et al.* De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* 2020;**370**:426–31.
6. Glasgow AA, Huang Y-M, Mandell DJ, *et al.* Computational design of a modular protein sense-response system. *Science* 2019;**366**:1024–8.
7. Langan RA, Boyken SE, Ng AH, *et al.* De novo design of bioactive protein switches. *Nature* 2019;**572**:205–10.
8. Dawson WM, Lang EJM, Rhys GG, *et al.* Structural resolution of switchable states of a de novo peptide assembly. *Nature Commun* 2021;**12**:1–10.
9. Shen H, Fallas JA, Lynch E, *et al.* De novo design of self-assembling helical protein filaments. *Science* 2018;**362**:705–9.
10. Hsia Y, Bale JB, Gonen S, *et al.* Design of a hyperstable 60-subunit protein icosahedron. *Nature* 2016;**535**:136–9.
11. Kan SBJ, Lewis RD, Chen K, *et al.* Directed evolution of cytochrome c for carbon-silicon bond formation: bringing silicon to life. *Science* 2016;**354**:1048–51.
12. Kan SBJ, Huang X, Gumulya Y, *et al.* Genetically programmed chiral organoborane synthesis. *Nature* 2017;**552**:132–6.
13. Savile CK, Janey JM, Mundorff EC, *et al.* Biocatalytic asymmetric synthesis of chiral amines from ketones applied to Sitagliptin manufacture. *Science* 2010;**329**:305–9.
14. Sun Z, Liu Q, Qu G, *et al.* Utility of B-factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering Thermostability. *Chem Rev* 2019;**119**:1626–65.
15. Hammer SC, Kubik G, Watkins E, *et al.* Anti-Markovnikov alkene oxidation by metal-oxo-mediated enzyme catalysis. *Science* 2017;**358**:215–8.
16. Zhang RK, Chen K, Huang X, *et al.* Enzymatic assembly of carbon-carbon bonds via iron-catalysed sp(3) C-H functionalization. *Nature* 2019;**565**:67–72.
17. Yu D, Wang J-B, Reetz MT. Exploiting designed oxidase-peroxygenase mutual benefit system for asymmetric cascade reactions. *J Am Chem Soc* 2019;**141**:5655–8.
18. Khoury GA, Smadbeck J, Kieslich CA, *et al.* Protein folding and de novo protein design for biotechnological applications. *Trends Biotechnol* 2014;**32**:99–109.

19. The runners-up. *Science (New York, NY)* 2016;**354**:1518–23.
20. Kuhlman B, Dantas G, Ireton GC, *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* 2003;**302**:1364–8.
21. Huang P-S, Ban Y-EA, Richter F, *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *Plos One* 2011;**6**:1–8.
22. Koga N, Tatsumi-Koga R, Liu G, *et al.* Principles for designing ideal protein structures. *Nature* 2012;**491**:222–7.
23. Joh NH, Wang T, Bhate MP, *et al.* De novo design of a transmembrane Zn2+−transporting four-helix bundle. *Science* 2014;**346**:1520–4.
24. King NP, Bale JB, Sheffler W, *et al.* Accurate design of co-assembling multi-component protein nanomaterials. *Nature* 2014;**510**:103–8.
25. Gainza P, Nisonoff HM, Donald BR. Algorithms for protein design. *Curr Opin Struct Biol* 2016;**39**:16–26.
26. Berman HM, Westbrook J, Feng Z, *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
27. Fox NK, Brenner SE, Chandonia J-M. SCOPe: structural classification of proteins-extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* 2014;**42**:D304–9.
28. Bateman A, Martin MJ, O'Donovan C, *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**:D158–69.
29. Madani A, Krause B, Greene ER, *et al.* Deep neural language modeling enables functional protein generation across families. In: *preprint: bioRxiv*, 2021. https://doi.org/10.1101/2021.07.18.452833.
30. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–44.
31. Wang S, Sun S, Li Z, *et al.* Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol* 2017;**13**:e1005324.
32. Yang J, Anishchenko I, Park H, *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A* 2020;**117**:1496–503.
33. Alley EC, Khimulya G, Biswas S, *et al.* Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 2019;**16**:1315–22.
34. Biswas S, Khimulya G, Alley EC, *et al.* Low-N protein engineering with data-efficient deep learning. *Nat Methods* 2021;**18**:389–96.
35. Radford A, Jozefowicz R, Sutskever I. Learning to generate reviews and discovering sentiment. In: *preprint: arXiv*, 2017. https://doi.org/10.48550/arXiv.1704.01444.
36. Ashish V, Noam S, Niki P, *et al.* Attention is all you need. In: *preprint: arXiv*, 2017. https://doi.org/10.48550/arXiv.1706.03762.
37. Rives A, Meier J, Sercu T, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* 2021;**118**:e2016239118.
38. Ingraham J, Garg VK, Barzilay R, *et al.* Generative models for graph-based protein design. In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, 2019.
39. Strokach A, Becerra D, Corbi-Verge C, *et al.* Fast and flexible protein design using deep graph neural networks. *Cell Systems* 2020;**11**:402–11.
40. Wang X, Flannery ST, Kihara D. Protein docking model evaluation by graph neural networks. *Front Mol Biosci* 2021;**8**:1–13.
41. Réau M, Renaud N, Xue LC, *et al.* DeepRank-GNN: a graph neural network framework to learn patterns in protein-protein interfaces. In: *preprint bioRxiv*, 2021. https://doi.org/10.1101/2021.12.08.471762.
42. Goodfellow I, Pouget-Abadie J, Mirza M, *et al.* Generative adversarial networks. *Commun ACM* 2020;**63**:139–44.
43. Doersch C. Tutorial on variational autoencoders. In: *preprint: arXiv*, 2021. https://doi.org/10.48550/arXiv.1606.05908.
44. Silver D, Huang A, Maddison CJ, *et al.* Mastering the game of go with deep neural networks and tree search. *Nature* 2016;**529**:484–9.
45. Staddon JER. Reinforcement learning: an introduction, 2nd edition. *J Exp Anal Behav* 2020;**113**:485–91.
46. Duan Y, Chen X, Houthooft R, *et al.* Benchmarking deep reinforcement learning for continuous control. In: *33rd International Conference on Machine Learning*, New York, USA, 2016.
47. Mnih V, Kavukcuoglu K, Silver D, *et al.* Human-level control through deep reinforcement learning. *Nature* 2015;**518**:529–33.
48. Mahmud M, Kaiser MS, Hussain A, *et al.* Applications of deep learning and reinforcement learning to biological data. *IEEE Trans Neural Netw Learn Syst* 2018;**29**:2063–79.
49. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. *Sci Adv* 2018;**4**:eaap7885.
50. Dong X, Shen J, Wang W, *et al.* Dynamical hyperparameter optimization via deep reinforcement learning in tracking. *IEEE Trans Pattern Anal Mach Intell* 2021;**43**:1515–29.
51. Deng Y, Bao F, Kong Y, *et al.* Deep direct reinforcement learning for financial signal representation and trading. *IEEE Trans Neural Netw Learn Syst* 2017;**28**:653–64.
52. Huang C, Mo R, Yuen C. Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning. *IEEE J Select Areas Commun* 2020;**38**:1839–50.
53. Hanson J, Peliwal K, Litfin T, *et al.* Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics* 2018;**34**:4039–45.
54. Shen T, Wu J, Lan H, *et al.* When homologous sequences meet structural decoys: accurate contact prediction by tFold in CASP14-(tFold for CASP14 contact prediction). *Prot-Struct Funct Bioinform* 2021;**89**(12):1901–10.
55. Ding W, Mao W, Shao D, *et al.* DeepConPred2: an improved method for the prediction of protein residue contacts. *Comput Struct Biotechnol J* 2018;**16**:503–10.
56. Mao W, Ding W, Xing Y, *et al.* AmoebaContact and GDFold as a pipeline for rapid de novo protein structure prediction. *Nat Mach Intell* 2020;**2**:25–33.
57. Li Y, Hu J, Zhang C, *et al.* ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* 2019;**35**:4647–55.
58. Xu J. Distance-based protein folding powered by deep learning. *Proc Natl Acad Sci U S A* 2019;**116**:16856–65.
59. Ding W, Gong H. Predicting the real-valued inter-residue distances for proteins. *Adv Sci* 2020;**7**:1–11.
60. Senior AW, Evans R, Jumper J, *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* 2020;**577**:706–10.
61. Ding W, Xu Q, Liu S, *et al.* SAMF: a self-adaptive protein modeling framework. *Bioinformatics* 2021;**37**:4075–82.
62. AlQuraishi M. End-to-end differentiable learning of protein structure. *Cell Syst* 2019;**8**:292–301.
63. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
64. Tunyasuvunakool K, Adler J, Wu Z, *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* 2021;**596**:590–6.

65. Baek M, DiMaio F, Anishchenko I, *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6.
66. Zhang C, Zheng W, Mortuza SM, *et al.* DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* 2020;**36**:2105–12.
67. Helling R, Li H, Melin R, *et al.* The designability of protein structures. *J Mol Graph Model* 2001;**19**:157–67.
68. Jiang L, Althoff EA, Clemente FR, *et al.* De novo computational design of retro-aldol enzymes. *Science* 2008;**319**:1387–91.
69. Tinberg CE, Khare SD, Dou J, *et al.* Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* 2013;**501**:212–6.
70. Huang P-S, Oberdorfer G, Xu C, *et al.* High thermodynamic stability of parametrically designed helical bundles. *Science* 2014;**346**:481–5.
71. Polizzi NF, DeGrado WF. A defined structural unit enables de novo design of small-molecule-binding proteins. *Science* 2020;**369**:1227–33.
72. Rocklin GJ, Chidyausiku TM, Goreshnik I, *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* 2017;**357**:168–74.
73. Dou J, Vorobieva AA, Sheffler W, *et al.* De novo design of a fluorescence-activating beta-barrel. *Nature* 2018;**561**:485–91.
74. Marcos E, Chidyausiku TM, McShan AC, *et al.* De novo design of a non-local beta-sheet protein with high stability and accuracy. *Nat Struct Mol Biol* 2018;**25**:1028–34.
75. Lin Y-R, Koga N, Tatsumi-Koga R, *et al.* Control over overall shape and size in de novo designed proteins. *Proc Natl Acad Sci U S A* 2015;**112**:E5478–85.
76. Marcos E, Basanta B, Chidyausiku TM, *et al.* Principles for designing proteins with cavities formed by curved beta sheets. *Science* 2017;**355**:201–6.
77. Park K, Shen BW, Parmeggiani F, *et al.* Control of repeat-protein curvature by computational protein design. *Nat Struct Mol Biol* 2015;**22**:167–74.
78. Huang P-S, Feldmeier K, Parmeggiani F, *et al.* De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat Chem Biol* 2016;**12**:29–34.
79. Cooper S, Khatib F, Treuille A, *et al.* Predicting protein structures with a multiplayer online game. *Nature* 2010;**466**:756–60.
80. Koepnick B, Flatten J, Husain T, *et al.* De novo protein design by citizen scientists. *Nature* 2019;**570**:390–4.
81. Yang C, Sesterhenn F, Bonet J, *et al.* Bottom-up de novo design of functional proteins with complex structural features. *Nat Chem Biol* 2021;**17**:492–U306.
82. Anand N, Huang P. Generative modeling for protein structures. In: *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada, 2018.
83. Li Z, Nguyen SP, Xu D, *et al.* Protein loop modeling using deep generative adversarial network. In: *IEEE 29th International Conference on Tools with Artificial Intelligence*, Boston, MA, USA. 2017, pp. 1085–1091, doi: 10.1109/ICTAI.2017.00166.
84. Karimi M, Zhu S, Cao Y, *et al.* De novo protein design for novel folds using guided conditional Wasserstein generative adversarial networks. *J Chem Inf Model* 2020;**60**:5667–81.
85. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. In: *preprint: arXiv*, 2016. https://doi.org/10.48550/arXiv.1511.06434.
86. Boyd S, Parikh N, Chu E, *et al.* Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends Mach Learn* 2011;**3**:1–122.
87. Greener JG, Moffat L, Jones DT. Design of metalloproteins and novel protein folds using variational autoencoders. *Sci Rep* 2018;**8**:1–12.
88. Hawkins-Hooker A, Depardieu F, Baur S, *et al.* Generating functional protein variants with variational autoencoders. *PLoS Comput Biol* 2021;**17**:1–23.
89. Guo X, Du Y, Tadepalli S, *et al.* Generating tertiary protein structures via an interpretative variational autoencoder. In: *preprint: arXiv*, 2020. https://doi.org/10.48550/arXiv.2004.07119.
90. Mordvintsev A, Olah C, Tyka M. *Inceptionism: going deeper into neural networks*. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html (18, November, 2021, date last accessed.).
91. Anishchenko I, Pellock SJ, Chidyausiku TM, *et al.* De novo protein design by deep network hallucination. *Nature* 2021;**600**(7889):547–52.
92. Anand N, Eguchi R, Huang PS. Fully differentiable full-atom protein backbone generation. In: *7th International Conference on Learning Representations (ICLR 2019)*. New Orleans, USA, 2019. https://openreview.net/pdf?id=SJxnVL8YOV.
93. Eguchi RR, Anand N, Choe CA, *et al.* IG-VAE: generative modeling of immunoglobulin proteins by direct 3D coordinate generation. In: *preprint: bioRxiv*, 2020. https://doi.org/10.1101/2020.08.07.242347.
94. Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* 2009;**10**:866–76.
95. Axe DD. Estimating the prevalence of protein sequences adopting functional enzyme folds. *J Mol Biol* 2004;**341**:1295–315.
96. Chandrasekaran R, Ramachandran GN. Studies on the conformation of amino acids. XI. Analysis of the observed side group conformation in proteins. *Int J Protein Res* 1970;**2**:223–33.
97. Shapovalov MV, Dunbrack RL, Jr. A smoothed backbone-dependent Rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* 2011;**19**:844–58.
98. Desmet J, De Maeyer M, Hazes B, *et al.* The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 1992;**356**:539–42.
99. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A* 2000;**97**:10383–8.
100. Ollikainen N, de Jong RM, Kortemme T. Coupling protein side-chain and backbone flexibility improves the re-design of protein-ligand specificity. *PLoS Comput Biol* 2015;**11**:1–22.
101. Georgiev I, Donald BR. Dead-end elimination with backbone flexibility. *Bioinformatics* 2007;**23**:I185–94.
102. Davey JA, Chica RA. Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Prot Struct Funct Bioinform* 2014;**82**:771–84.
103. Loshbaugh AL, Kortemme T. Comparison of Rosetta flexible-backbone computational protein design methods on binding interactions. *Prot Struct Funct Bioinform* 2020;**88**:206–26.
104. Boyken SE, Chen Z, Groves B, *et al.* De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* 2016;**352**:680–7.
105. O'Connell J, Li Z, Hanson J, *et al.* SPIN2: predicting sequence profiles from protein structures using deep neural networks. *Prot Struct Funct Bioinform* 2018;**86**:629–33.

106. Zhang Y, Chen Y, Wang C, *et al.* ProDCoNN: protein design using a convolutional neural network. *Prot Struct Funct Bioinform* 2020;**88**:819–29.
107. Qi Y, Zhang JZH. DenseCPD: improving the accuracy of neural-network-based computational protein sequence design with DenseNet. *J Chem Inf Model* 2020;**60**:1245–52.
108. Yang J, Yan R, Roy A, *et al.* The I-TASSER suite: protein structure and function prediction. *Nat Methods* 2015;**12**:7–8.
109. Anand-Achim N, Eguchi RR, Mathews II, *et al.* Protein sequence design with a learned potential. In: *preprint: Biorxiv*, 2020. https://doi.org/10.1101/2020.01.06.895466.
110. Norn C, Wicky BIM, Juergens D, *et al.* Protein sequence design by explicit energy landscape optimization. In: *preprint: bioRxiv*, 2020. https://doi.org/10.1101/2020.07.23.218917.
111. Rohl CA, Strauss CEM, Misura KMS, *et al.* Protein structure prediction using Rosetta. *Methods Enzymol* 2004;**383**: 66–93.
112. Alford RF, Leaver-Fay A, Jeliazkov JR, *et al.* The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput* 2017;**13**:3031–48.
113. Burges CJC, Ragno R, Le QV, *et al.* Learning to rank with non-smooth cost functions. In: *19th Conference on Neural Information Processing Systems (NeurIPS 2006)*, Vancouver, Canada, 2007.
114. Bateman A, Martin M-J, Orchard S, *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9.
115. Devlin J, Chang M-W, Lee K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: *preprint: arXiv*, 2019. https://doi.org/10.48550/arXiv.1810.04805.
116. Asgari E, Mofrad MRK. Continuous distributed representation of biological sequences for deep proteomics and genomics. *Plos One* 2015;**10**:1–15.
117. 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. In: *preprint: arXiv*, 2020. https://doi.org/10.48550/arXiv.2007.06225.
118. Suzek BE, Wang Y, Huang H, *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;**31**:926–32.
119. Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space. In: *preprint: arXiv*, 2013. https://doi.org/10.48550/arXiv.1301.3781.
120. Bairoch A, Bougueleret L, Altairac S, *et al.* The universal protein resource (UniProt). *Nucleic Acids Res* 2008;**36**:D190–5.
121. Repecka D, Jauniskis V, Karpus L, *et al.* Expanding functional protein sequence spaces using generative adversarial networks. *Nat Mach Intell* 2021;**3**:324–33.
122. Shin J-E, Riesselman AJ, Kollasch AW, *et al.* Protein design and variant prediction using autoregressive generative models, nature. *IDAA Commun* 2021;**12**:1–11.
123. Sohn K, Yan X, Lee H. Learning structured output representation using deep conditional generative models. In: *28th Conference on Neural Information Processing Systems (NeurIPS 2015)* Montréal, Canada, 2015. https://proceedings.neurips.cc/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.
124. Costello Z, Martin HG. How to hallucinate functional proteins. In: *preprint: arXiv*, 2019. https://doi.org/10.48550/arXiv.1903.00458.
125. Unsal S, Atas H, Albayrak M, *et al.* Evaluation of methods for protein representation learning: a quantitative analysis. In: *preprint: bioRxiv*, 2020. https://doi.org/10.1101/2020.10.28.359828.
126. Rao R, Bhattacharya N, Thomas N, *et al.* Evaluating protein transfer learning with TAPE. *Adv Neural Inform Process Syst* 2019;**32**:9689–701.
127. Gupta A, Zou J. Feedback GAN for DNA optimizes protein functions. *Nat Mach Intell* 2019;**1**:105–11.
128. Izadpanah A, Gallo RL. Antimicrobial peptides. *J Am Acad Dermatol* 2005;**52**:381–92.
129. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. In: *preprint: arXiv*, 2018. https://doi.org/10.48550/arXiv.1803.01271.
130. Zhang H, Goodfellow I, Metaxas D, *et al.* Self-attention generative adversarial networks. In: *preprint: arXiv*, 2019. https://doi.org/10.48550/arXiv.1805.08318.
131. Wu Z, Yang KK, Liszka MJ, *et al.* Signal peptides generated by attention-based neural networks. *ACS Synth Biol* 2020;**9**: 2154–61.
132. Angermueller C, Dohan D, Belanger D, *et al.* Model-based reinforcement learning for biological sequence design. In: *8th International Conference on Learning Representations (ICLR 2020)*. Virtual Conference, 2019.
133. Schulman J, Wolski F, Dhariwal P, *et al.* Proximal policy optimization algorithms. In: *preprint: arXiv*, 2017. https://doi.org/10.48550/arXiv.1707.06347.
134. Olivecrona M, Blaschke T, Engkvist O, *et al.* Molecular de-novo design through deep reinforcement learning. *J Chem* 2017;**9**:48, 48.
135. Goodfellow IJ, Mirza M, Xiao D, *et al.* An empirical investigation of catastrophic forgetting in gradient-based neural networks. In: *preprint: arXiv*, 2015. https://doi.org/10.48550/arXiv.1312.6211.
136. Griss R, Schena A, Reymond L, *et al.* Bioluminescent sensor proteins for point-of-care therapeutic drug monitoring. *Nat Chem Biol* 2014;**10**:598–603.
137. Feng J, Jester BW, Tinberg CE, *et al.* A general strategy to construct small molecule biosensors in eukaryotes. *Elife* 2015;**4**: 1–23.
138. Koday MT, Nelson J, Chevalier A, *et al.* A computationally designed hemagglutinin stem-binding protein provides in vivo protection from influenza independent of a host immune response. *PLoS Pathog* 2016;**12**:1–23.
139. Kiss G, Celebi-Oelcuem N, Moretti R, *et al.* Computational enzyme design. *Angew Chem Int Ed* 2013;**52**:5700–25.
140. Garrabou X, Wicky BIM, Hilvert D. Fast Knoevenagel condensations catalyzed by an artificial Schiff Base-forming enzyme. *J Am Chem Soc* 2016;**138**:6972–4.
141. Kries H, Blomberg R, Hilvert D. De novo enzymes by computational design. *Curr Opin Chem Biol* 2013;**17**:221–8.
142. Correia BE, Bates JT, Loomis RJ, *et al.* Proof of principle for epitope-focused vaccine design. *Nature* 2014;**507**:201–6.
143. Deng J, Dong W, Socher R, *et al.* ImageNet: a large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Miami, USA. 2009, pp. 248–255.
144. Wang CY, Chang PM, Ary ML, *et al.* ProtaBank: a repository for protein design and engineering data. *Protein Sci* 2018;**27**: 1113–24.
145. Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: *31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, USA, 2017.
146. Sung F, Yang Y, Zhang L, *et al.* Learning to compare: relation network for few-shot learning. In: *2018 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, Salt Lake City, USA. 2018, pp. 1199–1208, doi: 10.1109/CVPR.2018.00131.

147. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. In: *preprint: arXiv*, 2015. https://doi.org/10.48550/arXiv.1503.02531.
148. Tang Y, Wang Y, Xu Y, *et al.* SCOP: scientific control for reliable neural network pruning. In: *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, 2020.
149. Polizzi NF, Wu YB, Lemmin T, *et al.* De novo design of a hyperstable non-natural protein-ligand complex with sub-angstrom accuracy. *Nat Chem* 2017;**9**:1157–64.
150. Larson RC, Maus MV. Recent advances and discoveries in the mechanisms and functions of CAR T cells. *Nat Rev Cancer* 2021;**21**:145–61.
151. Sesterhenn F, Yang C, Bonet J, *et al.* De novo protein design enables the precise induction of RSV-neutralizing antibodies. *Science* 2020;**368**:730.
152. Kintzing JR, Cochran JR. Engineered knottin peptides as diagnostics, therapeutics, and drug delivery vehicles. *Curr Opin Chem Biol* 2016;**34**:143–50.
153. Larue RC, Xing EM, Kenney AD, *et al.* Rationally designed ACE2-derived peptides inhibit SARS-CoV-2. *Bioconjug Chem* 2021;**32**:215–23.