Article published online: 2024-09-03

Deep Learning-Based Self-Adaptive Evolution of Enzymes

Shuiqin Jiang[1]* Dong Yi[1]*

[1] Research Center for Systems Biosynthesis, China State Institute of Pharmaceutical Industry, Shanghai, People's Republic of China

Pharmaceut Fronts 2024;6:e252–e264.

Address for correspondence Shuiqin Jiang, PhD, Research Center for Systems Biosynthesis, China State Institute of Pharmaceutical Industry, 285 Gebaini Road, Shanghai 201203, People's Republic of China (e-mail: jiang.shuiqin@sysbiosyn.com).

Dong Yi, PhD, Research Center for Systems Biosynthesis, China State Institute of Pharmaceutical Industry, 285 Gebaini Road, Shanghai 201203, People's Republic of China (e-mail: yid@sysbiosyn.com).

Abstract

Biocatalysis has been widely used to prepare drug leads and intermediates. Enzymatic synthesis has advantages, mainly in terms of strict chirality and regional selectivity compared with chemical methods. However, the enzymatic properties of wild-type enzymes may or may not meet the requirements for biopharmaceutical applications. Therefore, protein engineering is required to improve their catalytic activities. Thanks to advances in algorithmic models and the accumulation of immense biological data, artificial intelligence can provide novel approaches for the functional evolution of enzymes. Deep learning has the advantage of learning functions that can predict the properties of previously unknown protein sequences. Deep learning-based computational algorithms can intelligently navigate the sequence space and reduce the screening burden during evolution. Thus, intelligent computational design combined with laboratory evolution is a powerful and potentially versatile strategy for developing enzymes with novel functions. Herein, we introduce and summarize deep-learning-assisted enzyme functional adaptive evolution strategies based on recent studies on the application of deep learning in enzyme design and evolution. Altogether, with the developments of technology and the accumulation of data for the characterization of enzyme functions, artificial intelligence may become a powerful tool for the design and evolution of intelligent enzymes in the future.

Keywords

- artificial intelligence
- deep learning
- protein engineering
- directed evolution
- biopharmaceuticals

Introduction

Biocatalysis has attracted much attention in the field of biopharmaceuticals. It has the advantages of mild reaction conditions, environmental friendliness, strict regioselectivity, and stereoselectivity, and can be used to prepare precursors, intermediates, and final chiral products.[1–3] For example, iron- and α-ketoglutarate-dependent oxygenases were engineered to improve the hydroxylation activity of *N*-succinyl-*threo*-3,4-dimethyoxyphenylalanine to produce *N*-succinyl-L-*threo*-3,4-dimethoxyphenylserine, a precursor to a psychoactive drug, Droxidopa.[4] (*S*)-Pregabalin, a drug for the treatment of epilepsy, neuropathic pain, fibromyalgia, and generalized anxiety disorders, can be synthesized using regioselectivity of nitrilase and the chiral selectivity of lipase.[5] Throughout natural evolution, wild-type enzymes are well-adapted to natural substrates but are less active against unnatural substrates for industrial applications. Most enzymes are not stable enough for industrial production.

received
September 19, 2023
accepted
June 25, 2024
article published online
September 3, 2024

DOI https://doi.org/10.1055/s-0044-1788317.
ISSN 2628-5088.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Therefore, protein engineering has become an important method to solve this problem.[6,7]

From the perspective of the catalytic mechanism, enzyme-, substrate-, and cofactor-binding conformations that are spatially and chemically conducive to the target reaction are crucial to improving the enzyme's catalytic activity.[8–10] Enzyme function is affected by global sites, and there are synergistic effects of multisite association.[11] Through site-specific mutations, efficient mutants can alter the network of residue interactions, further facilitating the formation of favorable conformations in the reaction environment. These mutation sites can be located near the catalytic center of the enzyme or on the surface far away from the catalytic center.[12–14] For a protein sequence of length N, the number of potential sequence combinations is 20^N, as each site can be occupied by any of the 20 amino acid residues. For the target properties, locating the important sites of the enzyme to obtain efficient mutants is the key problem of protein engineering.

Three main approaches in protein engineering are directed evolution, semirational design, and rational design. Directed evolution is combined with high-throughput screening to obtain the target mutants through iterative rounds of mutagenesis and screening.[15] However, this approach is highly dependent on the screening strategy[16] and experimental costs. Semi-rational design is based on enzyme sequences, three-dimensional (3D) structures, catalytic mechanisms, selection of specific sites, and establishment of small-scale mutant libraries to improve enzyme function.[17] Therefore, there are high requirements for the structure resolution of enzymes and an understanding of the catalytic mechanisms. A semi-rational design usually has a strong advantage in the active pockets of enzymes, yet, is more difficult to implement at sites far from the active center. It can easily fall into local optimal solutions, which limits its wide application in industry. Therefore, it is crucial to develop novel methods in protein engineering that can efficiently guide the development of enzyme catalysts and reduce research and development costs.

In recent years, owing to the rapid developments of gene sequencing and high-throughput experimental technology, several large biological databases, such as GenBank, UniProt, and the Protein Data Bank (PDB), have been established, laying the foundation for the application of artificial intelligence in life sciences. The artificial intelligence methods are novel data-driven strategies independent of enzyme crystal structure, catalytic mechanism cognition, multi-round iterations, and screening strategies.[18] The predicted mutation sites cover various parts of the global protein, allowing for the exploration of a larger portion of the protein sequence structure. Moreover, deep learning is largely seen as a supervised problem when applied to directed evolution. Its main task is to learn a function, also named the protein fitness landscape, from a set of protein sequences with associated labels (e.g., catalytic activity, selectivity, and stability), which can further predict the labels of previously unseen sequences. In each evolution cycle, the function is used to computationally evaluate a large number of protein sequences, which are then updated with feedback from laboratory results. As a result, deep learning achieves better evolution efficiency than laboratory screening alone.[19] Artificial intelligence, especially deep-learning-assisted enzyme development, has become a new development trend.[20,21]

In this review, we summarize the progress of deep-learning studies on enzyme evolution in recent years and discuss the advantages and limitations of artificial intelligence in assisting enzyme functional evolution to promote the application of biocatalysis in the biopharmaceutical field.

Deep-Learning Methods

Deep learning is an important branch of artificial intelligence that aims to design algorithms to help machines learn from data and improve their performance of specific tasks. Deep learning forms a deep network through multiple hierarchical neurons and has strong high-dimensional abstract learning capability.[22] It can automatically extract features from data without feature engineering, which makes it more advantageous when dealing with massive data. Commonly used deep-learning methods include the multilayer perceptron (MLP), recurrent neural network (RNN), convolutional neural network (CNN), graph neural network (GNN), variational autoencoder (VAE), generative adversarial network (GAN), transfer learning, and embedding.

Multilayer Perceptron

MLP, also known as an artificial neural network (ANN),[23] is an information-processing paradigm inspired by the way biological nervous system components, such as the brain, process information. It consists of an input layer, a hidden layer, and an output layer (►**Fig. 1**). The input layer receives the input data x. The hidden layer transforms the input data linearly, followed by the activation function to obtain the hidden features. The hidden features are transformed linearly to obtain the output y. The activation function introduces a nonlinear representation capability into the model. A gradient-based optimization algorithm is used to determine the model parameters. For model convergence, the maximum likelihood criterion is typically used for classification tasks, and the mean square error is often used as the loss function

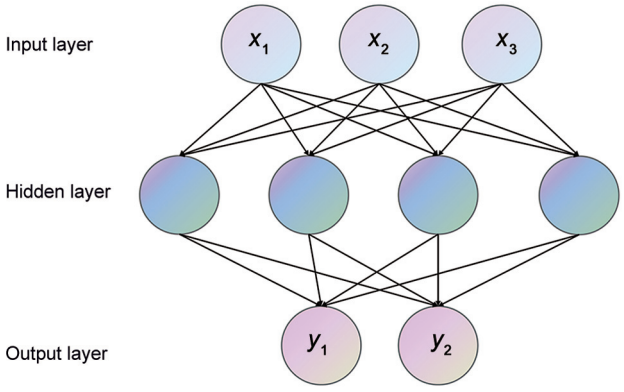

Fig. 1 The structure of MLP. MLP, multilayer perceptron.

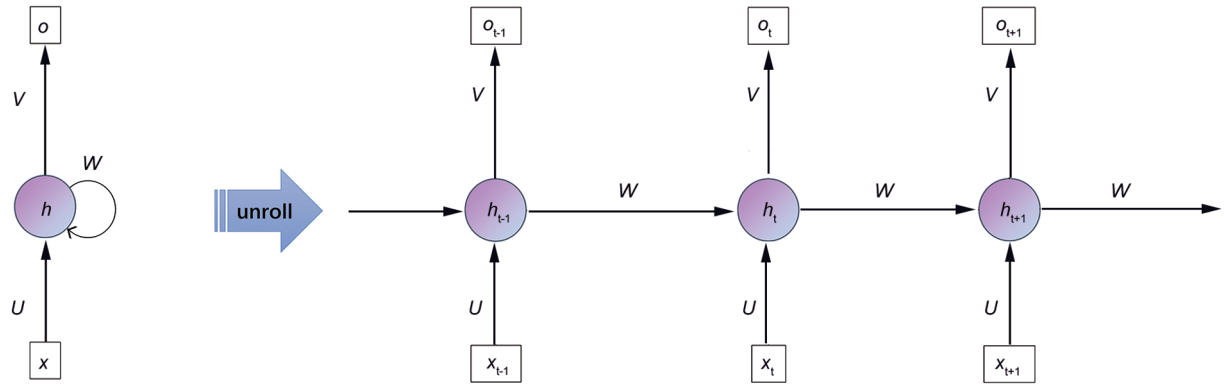

Fig. 2 The structure of RNN. RNN, recurrent neural network.

for regression tasks. MLP is often used as the base module for deep-learning models. MLPs have been used for enzyme catalysis tasks such as enzyme function prediction.[24]

Recurrent Neural Network

RNN has a temporal relationship with ANNs (►**Fig. 2**).[25] The input x_t of the current moment and the output h_{t-1} of the previous moment jointly determine the output o_t of the current moment. RNN adopts a parameter-sharing mechanism in a time series, which means that the input x shares the weight matrix U, the output h of the previous moment shares the weight matrix W, and the output layer shares the weight matrix V. RNN uses historical information, which gives it the ability to remember. RNNs are also used to learn protein sequence representations based on the approximately 24 million protein sequences in UniRef.[26]

Convolutional Neural Network

CNN is a deep-learning algorithm most often applied to analyze and learn visual features from large amounts of data.[27] CNN consists of a convolutional layer, a pooling layer, and a fully connected layer (►**Fig. 3**). Convolutional and pooling layers are added for comparison with ANN. The convolutional layer is the most important in a CNN and uses convolutional kernels to extract local features from the input data in a sliding window manner. CNN uses a parameter-sharing mechanism, in which the weights of each convolutional kernel in the convolutional layer are fixed. Thus, CNN effectively reduces the number of weights to be estimated and enables the network to learn in parallel. The pooling layer uses down-sampling to reduce the dimensionality of the features, which is usually performed by max and average pooling. The pooling layer is often located in the middle of continuous convolutional layers and is used to compress the feature data and various parameters to prevent the neural network from overfitting. The output feature data from the pooling layer are fed into the fully connected layer to calculate the output prediction. CNN has features such as parameter sharing and parallelizable learning that make it advantageous for processing high-dimensional data and excellent for tasks such as image recognition. CNNs have been used to encode protein sequences, and attention mechanisms have subsequently been used to learn the relationship between each amino acid residue and the conversion rate.[28]

Graph Neural Network

GNN is a method that enables deep learning of graph data.[29] One characteristic of graph data is that each node has unique features and structural information. GNN receives feature and structural information from the input graph, undergoes a multilayer computational transformation, and finally outputs the graph (►**Fig. 4**). The multilayer computational transformation is divided into three steps: node feature information extraction, node-local structural information fusion, and nonlinear transformation after information aggregation. A nonlinear transformation can increase the expressiveness of the model. GNN is capable of end-to-end learning of both feature and structural information of a node and is presently the best model for graph data-learning tasks. The results from GNN are superior to those of other methods for tasks such as node classification and edge prediction. GNNs have been used to learn the representation of 3D

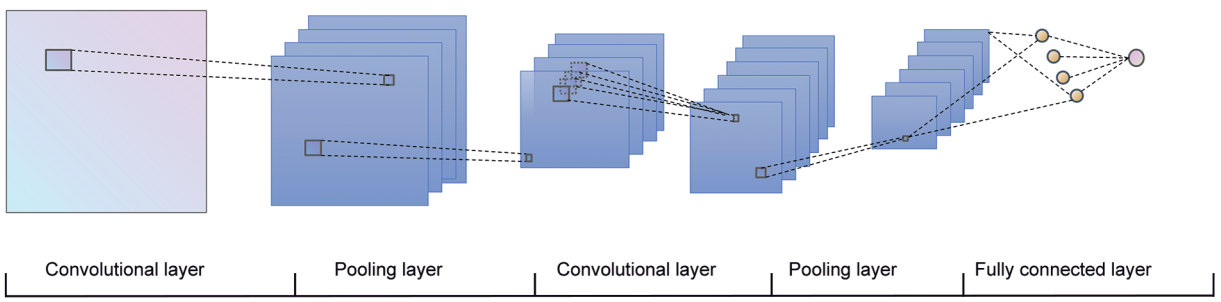

Fig. 3 The structure of CNN. CNN, convolutional neural network.

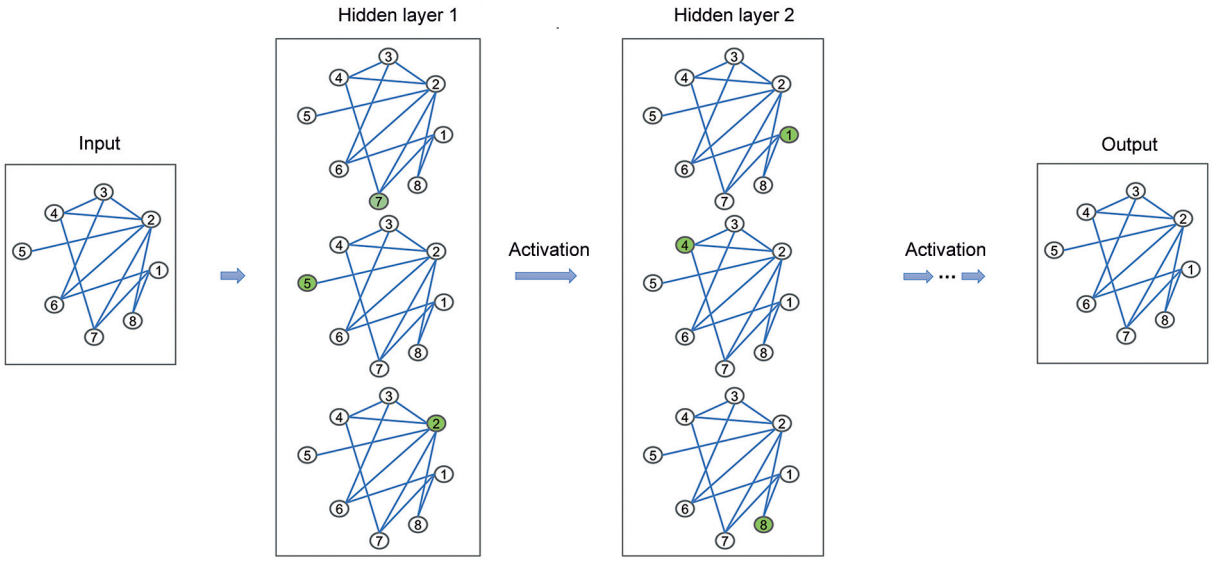

Fig. 4 The structure of GNN. GNN, graph neural network.

protein structures in protein function prediction and protein folding classification.[30]

Variational Autoencoder

VAE is a generative model for *a priori* data distribution.[31] VAE consists of two parts: an encoder and a decoder (►**Fig. 5**). The encoder embeds the input data x into a low-dimensional space, and the decoder reconstructs the original input data from the low-dimensional features. The low-dimensional features are referred to as the hidden features. VAE appends additional distribution assumptions to the hidden features, enabling the sampling of hidden features from the low-dimensional feature distribution and then generating new data samples using the decoder. The learning criterion of VAE minimizes the difference between the original and reconstructed data. VAEs have been used to generate protein sequences that can then be used to generate mutant libraries for enzyme evolution.[32]

Generative Adversarial Network

GAN, a generative model, learns data distribution using an adversarial approach.[33] GAN consists of two main parts: generative and discriminative networks (►**Fig. 6**). The generative network generates samples from a random distribution, and the discriminative network identifies whether the samples originate from the generative network or training samples. When the discriminant network cannot distinguish between the training samples and samples from the generator, the training of the model converges, leaving the trained generative network to generate new data samples. GAN does not need to explicitly model any data distribution to generate realistic samples. GANs have been widely used in many fields, such as imaging, text, speech, and mutant library generation for enzyme design.[34]

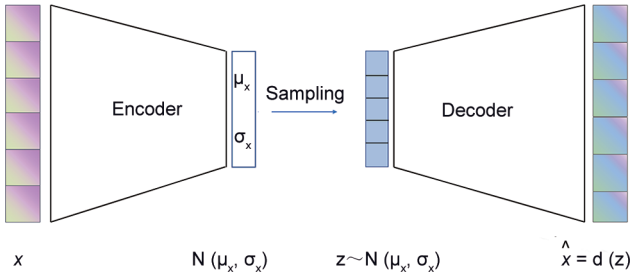

Fig. 5 The structure of VAE. VAE, variational autoencoder.

Transfer Learning

Transfer learning draws on the idea that humans develop knowledge in tasks and reuse existing knowledge in new and relevant tasks.[35] The information that can be transferred includes data samples, feature information, model parameters, and variable relationships. In transfer learning, a pretrained model in the relevant domain is typically used as a starting point to train the target task model on the target dataset. The advantage of transfer learning is that it does not require designing and training a completely new network for the target task, which can reduce the required data volume for the target task, shorten model development time, and improve the model performance. Transfer learning has been successfully used in the image and text domains and for the generalization improvement of the conditional protein sequence generation model.[36–38]

Embedding Technology

The embedding technology focuses on transforming high-dimensional sparse variables into dense vectors to facilitate downstream task processing. This technology was first applied to the textual domain for transforming words from one-hot-encoded vectors into dense D-dimensional vectors. In subsequent scenarios, the vectors were used as static vectors to represent words, involving models such as Word2-Vec, GloVe, and FastText.[39–42] The static vector representation, in which each word is represented as a fixed vector,

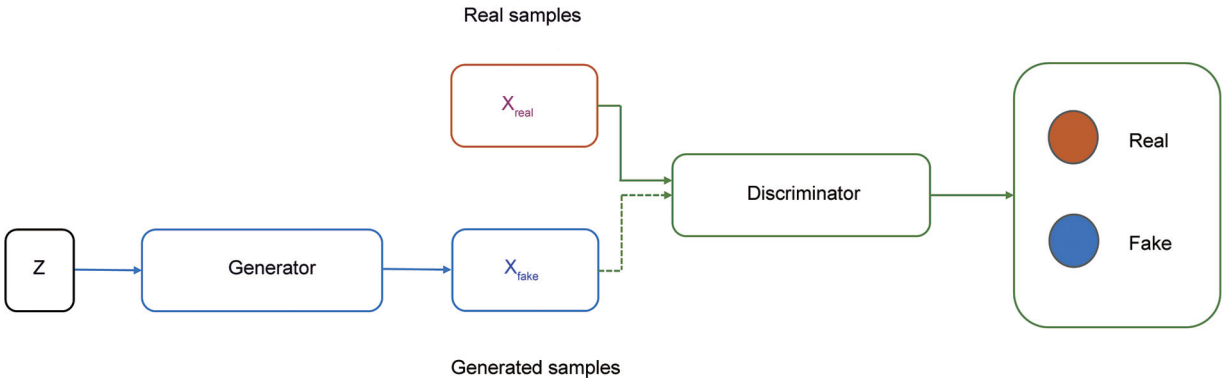

Fig. 6 The structure of GAN. GAN, generative adversarial network.

cannot effectively distinguish the semantics of the same word in different contexts; thus, a dynamic vector representation is then developed. The models involved in dynamic vector representation are Embeddings from Language Model,[43] Generative Pre-trained Transformer,[44] and Bidirectional Encoder Representations from Transformers (BERT).[45] The embedding technology can be used to represent discrete words, discrete nodes in graph data, amino acid residues in protein sequence data, and nodes in 3D protein structure data. For protein sequence representation, UniRep was trained based on approximately 2.4 million sequence data in the UniRef50 database, which can be used for downstream protein function prediction tasks.[26] In terms of 3D protein structure representation, the embedding technology is used to train the Uni-Mol model for predicting protein-binding pocket structure based on 3 million protein-binding pocket data.[46] The embedding technology has a strong comprehensive information representation capability and a low online deployment threshold, making it widely used in the industry.

Deep-Learning-Assisted Adaptive Evolution Strategy for Enzyme Engineering

Enzyme engineering promotes the evolution of natural enzymes to meet synthetic production requirements. However, the sequence space of enzymes is large. Rational design approaches that require protein structures and catalytic mechanisms are difficult to scale up efficiently. Recently, deep-learning methods have been applied in several studies to predict protein structures,[47] such as AlphaFold and Uni-Fold, as well as to learn sequence–function relationships from experimental data,[20,48–50] such as DLKcat.[27] This adaptive evolution strategy is consistent with wet characterization, shows great potential to facilitate automated and intelligent protein design, and effectively aids protein design and evolution. The adaptive evolution strategy contains the following four modules: sequence–function data, sequence–function relationship model, virtual mutant generation, and mutant selection (►**Fig. 7**).[51]

The sequence–function data module is a database for adaptive evolution strategies. This module collects sequence–

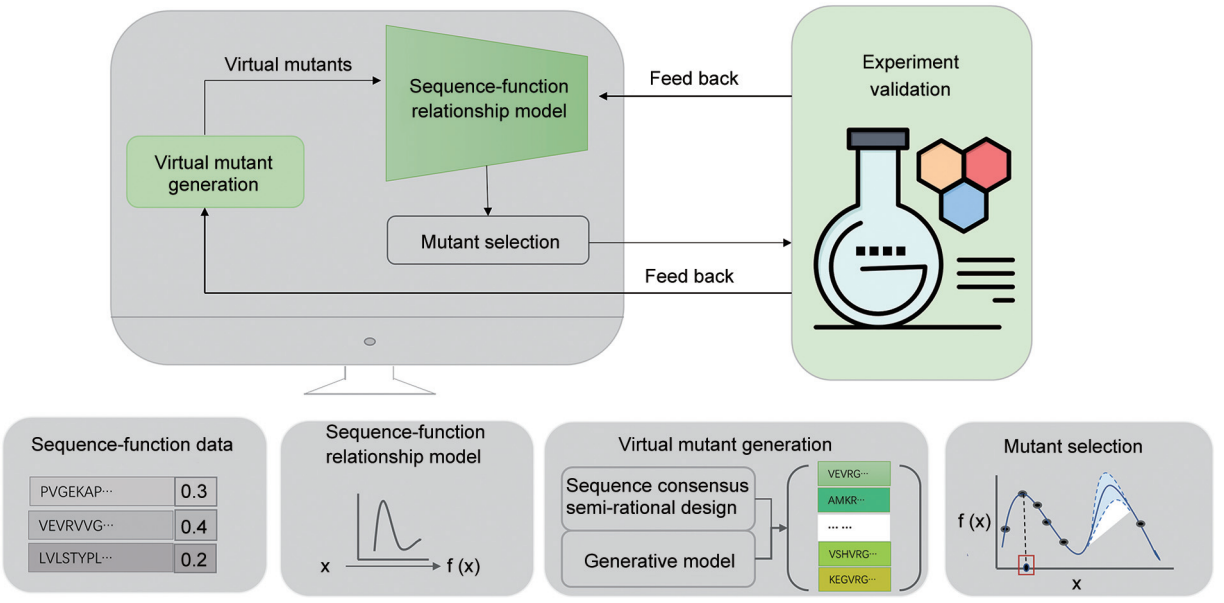

Fig. 7 Overview of adaptive protein engineering strategy.

function data of proteins, such as catalytic activity of enzymes, regioselectivity, and stereoselectivity, from public databases (►**Table 1**)[52–64] and self-produced laboratory data to build a new database that acts as data foundation for deep-learning modeling.

The sequence–function relationship modeling module determines whether an input sequence has a target function or not, and is, therefore, a key modeling process of the adaptive evolution strategy. This module establishes a relational model by applying deep-learning algorithms like MLP and RNN to learn the sequence–function relationship. The model outputs the predicted values related to the target functions based on the input protein sequence. The representation of protein sequences and deep-learning methods are crucial in this module. One-hot sparse representation and embedding-technology-based dense representation are the two main methods for representing protein sequences. Sparse representation learns limited sequence information; thus, it is difficult to represent inter-sequence relationships effectively. Embedding-based dense representation is borrowed from the field of natural language processing, which can effectively learn semantics and bring similar protein sequences closer to the encoding space. This feature is consistent with the fact that similar protein sequences have similar functions. Therefore, a dense representation is more conducive to predicting downstream protein functions. In addition, dense representations can make full use of several unlabeled protein sequences for characterization learning, effectively alleviating the problem of limited labeled data. Deep-learning methods, such as MLP, RNN, and CNN, can achieve end-to-end learning without manual feature processing. However, these are black-box models with data volume requirements.

The virtual mutant generation module was used to build a virtual mutant library to create important reservoirs for an adaptive evolution strategy. For protein sequences of length N, the module decides which mutants are selected from the sequence space of 20^N to build a virtual mutant library for the next step of function prediction. There are two common methods for building virtual mutant libraries. The first is based on sequence consistency and semi-rational design. The second is to use generative models to generate mutant libraries. The available neural network generative models include VAE and GAN.

The mutant selection module is used to identify candidate mutants for further characterization by wet experiments. The simplest method is to select the top N based on the ranked output of the sequence–function relationship model. However, this approach tends to select similar sequences, which is not conducive to improving the generalization of the model. Another method is to select N sequences from multiple local optima with high confidence, based on the prediction confidence given by the model. This approach increases the diversity of sequences and improves the generalization ability of the model.

In the adaptive evolution strategy, relevant sequence–function data are first collected for the target function to build a database. Based on the established database, the sequence–function relationship and virtual mutant generation models are established using a deep-learning method. The virtual mutant generation model is used to generate the mutant library, and the sequence–function relationship model is used to predict its function. Finally, the mutant selection module is used to select candidate mutants according to the predicted value. The following wet experiment provides feedback results for the above strategy and

Table 1 Commonly used database

No	Database name	Database type	Ref.
1	UniProt	Protein sequences annotated with functional information	52
2	Protein Data Bank	3D macromolecular structure data	53
3	ProThermDB	Thermodynamic database for proteins and mutants, including protein information, structural information, experimental conditions, literature information, and experimental thermodynamic data	54
4	FireProtDB	Experimental thermostability data for single-point mutants	55
5	IntEnzyDB	Structure and kinetics enzymology database	56
6	CAZy database	Functions and literature information for carbohydrate-active enzyme	57
7	CasPEDIA Database	Protein sequences and function information for class 2 CRISPR-Cas enzymes	58
8	GotEnzymes	Turnover numbers of enzyme–compound pairs	59
9	INTEDE	Interactome of drug-metabolizing enzymes	60
10	M-CSA	Enzyme reaction mechanisms and active sites	61
11	MetaCyc database	Metabolic pathways and enzymes	62
12	SABIO-RK	Biochemical reactions and their reaction kinetics	63
13	EzCatDB	Enzyme reactions, active-site structures, catalytic mechanisms, literature information, protein sequences, and structures	64

continuously optimizes the sequence–function relationship, the virtual mutant generation modules, and the mutant selection strategy. In this repeated cycle, enzyme evolution proceeds adaptively.

Applications of Deep Learning in Enzyme Evolution

Deep learning is a useful tool in enzyme engineering. In particular, an adaptive evolution strategy based on the synergy between deep learning and wet experiments is expected to drive enzyme evolution toward automation and intelligence. The sequence–function relationship model is the core module of the adaptive evolution strategy, which determines whether a sequence with a target function can be recognized. The mutant generation module establishes a virtual mutant library that provides an important reservoir for the adaptive evolution strategy and determines the upper limit of the quality of the final mutants obtained. The sequence–function relationship model and mutant generation module are crucial to the effectiveness of the adaptive evolution strategy. This section focuses on the progress of deep-learning applications in both the sequence–function relationship model and protein sequence generation.

Deep Learning in Protein Sequence–Function Relationship Modeling

The performance of sequence–function relationship models depends on whether the protein representation model can effectively extract the relevant feature information from the protein sequence and structure, and if the deep-learning model can effectively learn the relationships from protein representation.

Representation of Protein Sequence and Structure

The protein sequence determines its high-level structure, which, in turn, determines its function. In deep-learning-based sequence–function relationship modeling, the protein sequence and structure are represented in vector form as the input data of the deep-learning model. Representation of protein sequences and structures is crucial in modeling of sequence–function relationship. Effective extraction of protein information from sequences and structures has become a popular research topic.

Protein sequences are chains consisting of 20 essential amino acids that are very similar in form to natural languages. Protein sequences contain different functional information similar to the semantic concepts of natural language. Therefore, natural language processing can be used to effectively extract protein sequence information.[65–67] Wang and Zhao processed protein sequences using the word separation technology in natural language, and constructed a protein sequence dictionary containing 630,598 words using 100,000 protein sequences in the PDB database as a corpus and nine amino acids as the maximum word length.[68] Comparison of subword-based protein sequence segmentation and protein secondary structures showed that subword-based protein sequence segmentation was more efficient at the level of information representation. Asgari et al used a subword-based method to process protein sequences and found that this approach is more efficient than traditional methods in protein functional module discovery and protein sequence classification.[69] Due to technological developments, the bag-of-words (BoW) model has been applied to protein sequence processing. The basic assumption of the BoW model is that articles using a similar vocabulary have similar topics, and the BoW model focuses on the frequency of occurrences of vocabulary in an article. Based on 524,529 unannotated sequences in the UniProt database, Arnold and colleagues used the BoW model to learn protein sequence representation and used it for protein function classification tasks, such as chiral selectivity.[70] The BoW model-based protein representation is more effective than the traditional amino acid representation.

The BoW model mainly provides statistical information about the data and does not provide information about the contextual association of protein sequences, whereas the protein functions are affected by the relationship of the anterior–posterior position of amino acids in the sequences and the physicochemical properties of amino acids. Protein sequence representation based on embedding technology can extract contextual information from protein sequences and learn evolutionary information embedded in billions of protein sequences across various species.[71,72] AlQuraishi and colleagues used an RNN model to learn protein sequence representation based on approximately 24 million protein sequence data in UniRef50,[26] to obtain the protein sequence representation model UniRep, which was used for tasks such as protein stability prediction. Elnaggar et al used autoregressive models such as Transformer-XL and XLNet, autoencoding models such as BERT, Albert, Electra, and T5, as well as protein sequences from the UniRef and Big Fantastic Database as training data, to obtain the protein sequence representation model ProtTrans,[73] which performed well in the subsequent amino acid residue prediction tasks. Rao et al developed a pretrained protein embedding model called TAPE.[74] Pfam, a database of 31 million protein domains, was used as the pretraining corpus for TAPE, and TAPE was tested using five protein biology tasks. The pretrained model TAPE outperforms other models without self-supervised pretraining on almost all tasks. Self-supervised pretrained embedding is helpful, especially in protein engineering tasks. Min et al developed a novel pretraining model, PLUS-RNN, consisting of masked language modeling and a protein-specific pretraining task.[75] The PLUS-RNN was tested using seven widely used protein biology tasks and outperformed other language models (LMs) without protein-specific pretraining in six tasks.

Natural language processing technology can effectively extract protein sequence information for subsequent sequence–functional relationship learning. However, the function of a protein is determined by its high-level structure. Since only a few protein crystal structures have been resolved through an enormous experimental effort, learning the representation of high-level protein structures is much more difficult than learning protein sequences.

Computational methods for predicting 3D protein structures from protein sequences provided useful alternatives when protein crystallization is not possible. However, traditional computational methods, such as MODELLER,[76] fall far short of atomic accuracy, especially when no homologous structure is available. Deep learning–based structure-prediction tools provide accurate computational approaches even when similar structures are not known, which makes large-scale structural bioinformatics possible.[77] AlphaFold2, developed by DeepMind, predicts protein structures from amino acid sequences with much higher accuracy than previous methods.[78] AlphaFold2 views the prediction of protein structures as a graph inference in a 3D space, where the edges of the graph are defined by residues in proximity. The AlphaFold2 network was trained on the structural data from the PDB, which contains 200,000 crystal structures of proteins and nucleic acids. AlphaFold2 uses attention-based deep neural networks to extract spatial and evolutionary relationships from amino acid sequences. Currently, the predicted structure database using AlphaFold2 (https://alphafold.com/) contains over 200 million entries and is continuously growing. However, despite its breakthrough accuracy and performance to predict protein structures, AlphaFold2 model still has limitations. It is difficult to predict the structures with metal ions, cofactors, and other ligands, or posttranslational modifications, such as glycosylation, methylation, and phosphorylation.[79] Another important aspect is that the use of evolutionary information from larger multiple sequence alignments (MSAs) requires powerful computing processors and is time-consuming to predict the structure of proteins as their length increases.[80]

Recently, new modeling methods have been developed to overcome some of the limitations of AlphaFold2. Lin et al developed ESMfold, a masked transformer-based protein language model that operates without the use of MSAs. This omission significantly simplifies the neural architecture required for inference.[81] Compared with AlphaFold2 without MSA, ESMFold performed better on TM scores and achieved comparable accuracy to AlphaFold2 when predicting structures with high confidence. The approach significantly improves prediction speed while maintaining resolution and accuracy, as it does not require the construction of an MSA. The ESM Metagenomic Atlas (https://esmatlas.com/) is a database of predicted structures using ESMfold, which contains more than 617 million structures and 225 million structures predicted with high confidence from metagenomic databases. Recent advances have taken the problem of protein structure prediction to another level, in some cases to an experimental-like level of accuracy. Nevertheless, improvements are needed to overcome the limitation to predict conformer with ligands, and the inability to predict the effects of mutation on protein structure.

Protein sequence combined with high-level structural information is more efficient for functional prediction. Recently, deep learning networks have been trained to extract information from high-level protein structures, called high-level structural representation. Zhou et al obtained a protein representation model, Uni-Mol, using a transformer architecture to train on protein-binding pockets.[46] Uni-Mol performed well in predicting the protein receptor–ligand binding conformation. Gao et al developed a co-supervised pretraining (CoSP) model, a representative model of high-level protein structures, using a GNN to train on the binding pockets of protein receptors and small-molecule ligands.[82] CoSPs can be used for protein-binding pocket search and virtual screening. Zhang et al used GNN to train 3D protein structures and obtained GearNet, a representative model of high-level protein structures.[83] GearNet outperforms the best models based on protein sequences and requires less data for tasks such as the prediction of protein function and protein folding classification. Torng and Altman demonstrated a general framework that applied 3D CNNs (3DCNN) to detect protein functional sites from protein structures.[84] This framework can automatically extract task-dependent features from raw atom distributions, and be tested using the PROSITE family, nitric oxide synthase, and trypsin-like enzymes. The model can discover features from raw data that outperform predefined features and can be generally applied to any functional site, given the available data, without manual adjustments.

The high-level structure of a protein more directly determines its function, thus, incorporating high-level structure information into protein sequence models can effectively improve the performance of the model. Koohi-Moghadam et al developed a deep-learning approach to predict mutations occurring at the metal-binding sites of metalloproteins.[85] The approach uses five types of probes to generate energy-based grid maps from the 3D structures of metal-binding sites. The spatial and sequential features of the metal-binding sites were fed into multichannel CNNs (MCCNNs). The MCCNN model was trained and evaluated using integrated data from MetalPDB, CancerResource2, ClinVar, and UniProt Humsavar. The MCCNN could predict mutations in both the first and second spheres of metals in metalloproteins. The spatial characteristics of the metal-binding sites improved the performance of MCCNN. Mansoor et al found a protein-embedding approach using joint training on protein sequences and structures.[86] Pretrained Evolutionary Scale Modeling-1b was used to generate one-dimensional (1D) and two-dimensional features from the masked sequence. The SE(3) transformer was trained to output a 128-dimensional embedding for generating the final 1D representation from the masked structure and sequence representation. trRosetta2 was used as the training and validation dataset. A subset of the ProTherm dataset, consisting of 1,042 mutants from 126 wild-type proteins, was used to fine-tune the single-mutant effect prediction. Joint training with sequence and structural information improved the prediction of the effect of single mutations on thermal stability. Wang et al described a novel LM-GVP method composed of a protein LM and a GNN.[87] The protein LM was used to extract information from 1D amino acid sequences, and the GNN was used to obtain information from the 3D protein structures. The LM-GVP was tested using various property prediction tasks, including fluorescence, protease stability, and protein functions from gene ontology. LM-GVP outperformed the protein LMs in all tasks.

Protein Sequence–Function Relationship Learning

Efficient protein representation of sequences and structures combined with deep-learning models can help learn protein sequence–function relationships to guide enzyme evolution. Artificial intelligence-based methods for enzyme function evolution reduce experimental workload and improve the success rate of experiments.

Arnold colleagues used algorithms such as MLP to model the sequence–function relationship of nitric oxide dioxygenase,[88] and after only two rounds of screening for mutation site prediction, two target mutants with seven-site co-mutations were obtained, which were then used to achieve carbon–silicon bond formation in *R*- and *S*-conformations. Carbon–silicon bond-forming enzymes, which do not exist, were synthesized. The application of artificial intelligence in assisted enzyme-directed evolution helps to create new enzymes, reduce the number of rounds of directed evolution, and reduce the cost and time of research and development. Enzyme conversion rate (K_{cat}) is a core indicator of an enzyme's catalytic performance. Nielsen and colleagues used a GNN and CNN to establish a K_{cat} prediction model based on substrate structure and enzyme sequence, known as DLKca.[28] The model used GNN to encode the substrate structure, and CNN to encode the protein sequence, and subsequently used the attention mechanism to learn the relationship between each amino acid residue concerning the conversion rate. DLKcat can be used to predict the conversion rates of enzymes and their mutants into substrates. This study provided important information for enzyme mutation design and performance prediction. Shroff et al used a CNN to build a mutant prediction model based on the 3D protein structure.[89] This model was used to assist in the evolution of blue fluorescent proteins, phosphomannose isomerase, and TEM-1 β-lactamase, with a 6- to 30-fold functional improvement. Alper and colleagues developed a 3DCNN-based mutant prediction model, MutCompute, to assist in evaluating polyethylene terephthalate (PET) hydrolase.[90] Then, a highly efficient PET hydrolase variant was obtained, with a 38-fold increase in plastic hydrolysis capacity at 50°C, capable of degrading 51 different unprepared thermoformed PET products within 1 week. Wong and colleagues used various neural network models to assist directed evolution and increase the gene-editing capability of the clustered regularly interspaced short palindromic repeats (CRISPR)-associated protein 9 (Cas9) of CRISPR-Cas9.[91] The use of artificial intelligence models reduced the wet experimental workload by 95%. This study demonstrates the potential of artificial intelligence-supervised learning models for various enzyme evolution applications.

Deep-learning models trained using protein sequences with functional tags have achieved many results in protein evolution. However, there are a few protein sequences containing functional tags. Thus, adopting a transfer learning strategy to make full use of the existing resources without tags in large databases such as GenBank and UniProt, to reduce the reliance on tagged data, and to improve the generality of models has become a new trend in artificial intelligence-assisted protein evolution research.

Church and colleagues used a transfer learning strategy to learn the general features of proteins based on >20 million protein sequences in UniRef50, followed by fine-tuning the target protein sequence–function data to learn global and local features.[92] The strategy uses only a small amount of identified protein sequence–function data to learn sequence–function relationships and obtain protein representation models for building sequence-based protein function prediction models. With this strategy, green fluorescent protein and TEM-1β-lactamase achieved an increase in protein activity through only one round of mutation, reducing experimental workload and cost. Zhao and colleagues used a transfer learning strategy to develop a protein function prediction model, ECNet.[36] For amino acid representation, the pretrained model TAPE, based on the attention mechanism, was fine-tuned to represent the global evolutionary information, which generated 768-dimensional representation vectors for each amino acid. For sequence–function relationship learning, a bidirectional long short-term memory network (BiLSTM) algorithm followed by a two-layer fully connected neural network was used to build a functional prediction model. Amino acid residues were one-hot-encoded into the embedding layer to output a 20-dimensional representation vector, which was then spliced as input to the functional prediction model using the evolutionary information generated by the TAPE model. Evolution of TEM-1 β-lactamase using ECNet yielded mutants that were eightfold more functional than the wild type.

Combining protein sequence structure representation and supervised learning models to aid in enzyme evolution has shown high potential for application, but it currently covers only a small number of enzyme families. The introduction of transfer learning can improve the generalization of the model and is expected to be used for more enzyme families.

Deep Learning in Protein Sequence Generation

The protein sequence–function relationship model allows rapid screening for efficient target function mutants. Because protein sequences are large, protein sequence functional relationship model screening using all protein sequences is computationally expensive and inefficient. The protein sequence generation models provide a new solution to this problem. They learn data distribution features from several protein sequences and automatically generate candidate sequences,[34,50,93,94] forming a virtual mutant candidate pool for functional screening.

The generative model samples the sequence space where the target function is located to generate multi-residue mutants with high target performance. This method has a higher sampling efficiency compared with traditional methods based on sequence alignment. The generative models have been used for the sequence generation of many proteins.[95–98] The main generative models currently used in protein engineering are VAE, GAN, and autoregressive models. Lobzaev et al used the VAE model and natural language processing technology to address the problems of low activity and instability in the blood, and susceptibility to immune

reactions of human-derived sphingosine-1-phosphate lyases as therapeutic enzymes.[99] A protein sequence generative model based on 1,147 protein sequences of sphingosine-1-phosphate lyase was developed to generate a library of sphingosine-1-phosphate lyase mutants, which were evaluated in terms of hydrophobicity, isoelectric point, stability, and structure, and possessed the target properties. Hawkins-Hooker et al trained a VAE based on 70,000 luciferase protein sequences to generate protein sequences with luminescence functions.[32] Giessel et al used generative models to assist in the functional evolution of human-derived ornithine transcarbamylase (OTC) to improve the enzyme activity and thermal stability.[100] They collected 3,818 OTC sequences with 45% sequence similarity and trained a VAE generative model. The model generated 87 OTC sequences with an average mutation of eight residues compared with the wild type, 86% of the mutants were more active than the wild type, and the average melting temperature was 12°C higher than the wild type. In contrast, only 42.5% of the mutants obtained by the conventional consensus method were more active than the wild type, and the average melting temperature was only 8°C higher than that of the wild type. In addition, the mutants produced by the generative model improved both activity and thermal stability, whereas the mutants produced by the conventional consensus method improved only thermal stability but reduced activity. This result shows that the generative model can sample the target protein sequence space more efficiently. Repecka et al developed ProteinGAN,[101] a protein sequence generation model based on a GAN, by introducing an attention mechanism.[102] The model learned information about protein sequence evolutionary relationships from a complex amino acid sequence space and generated new sequences with natural properties such as solubility and physicochemical activity. Using this model to learn the malate dehydrogenase protein (MDH) sequences, 24% of the new sequences generated were soluble and had MDH catalytic activity. Surprisingly, a mutant with 106-residues mutation, equivalent to 34% of the wild-type protein sequence, still maintained catalytic activity compared with the wild type, whereas 50% of the protein sequence was usually co-mutated resulting in its inactivation. This study showed that protein sequence generation models can learn key features of target protein sequences and explore the sequence mutation space that is difficult to reach by traditional methods. However, the methods mentioned above are mainly applicable to certain classes of protein families and lack generality.

A transfer learning strategy can solve the problem of the generality of protein sequence-generative models. Madani et al used a transfer learning strategy to obtain a conditional sequence generation model ProGen, based on 280 million protein sequence data using an autoregressive model and an attention mechanism network.[37,38] The model generates protein sequences with target functions based on input functional tags and is suitable for multiple protein families. ProGen was used to learn the sequences of antibacterial lysozyme superfamily proteins, and 90 newly generated sequences were selected for activity testing. The lysozyme activity was 73%, and some sequences were more active than the control, indicating that the model developed had the potential to generate efficient mutants. Sevgen et al developed ProT-VAE, a generative model, using a transfer learning strategy by combining a pretrained attention mechanism model with the framework of the VAE.[103] ProT-VAE mode comprises three blocks. The first block is a pretrained transformer-based T5 encoder and decoder model called ProtT5nv. ProtT5nv is trained starting from a pretrained T5 model from NLP data and further trained with 46M protein sequences from UniRef50. The second block is a generic dimensionality-reduction block that efficiently compresses a high-dimensional transformer hidden state into a parsimonious intermediate-level representation. This block was pretrained using UniProt's mean-squared error reconstruction objective. The third block is a three-layer fully connected maximum mean discrepancy VAE (MMD-VAE). This block compresses the flattened output of the dimensionality reduction block. This block was initialized and trained from scratch for each target protein family of interest. The ProT-VAE model allows alignment-free training, whereas VAE models frequently require arranging sequences within MSAs. The computational complexity of MSAs grows exponentially with the number of proteins. ProT-VAEP was assayed using phenylalanine hydroxylase (PAH). Of the PAH proteins generated from ProT-VAE, 69 were active, 19 of which were more active than wild-type hPAH. ProT-VAE demonstrated the capacity to learn functional and phylogenetic separation within the latent space without the need for MSAs. The ProT-VAE model can generate highly mutated sequences (>100 mutations; up to 130 mutations for the highest activity) that are still functional. The ProT-VAE is an accurate, generative, fast, and transferable model for data-driven protein engineering.

Efficient protein sequence-generative models can provide high-quality candidate mutants for enzyme evolution. However, the bias of the training dataset causes the generative models to learn biased distributions, thus, the data balance of the training set needs to be considered in practical applications.

Conclusion

Artificial intelligence technology based on deep learning, representation learning, transfer learning, and generative modeling is helping enzyme engineering researchers to learn and apply protein sequence structure–function relationships from large amounts of data, exploring a larger protein sequence space, and providing more diverse and novel possibilities. Adaptive enzyme evolution strategies combining computational and experimental modes have a high potential to improve efficiency and reduce cost and have become a new trend in enzyme evolution. However, the application of artificial intelligence technology in enzyme evolution is still in its early stages, and many issues need to be addressed to fully exploit the potential of artificial intelligence in this field. First, since databases such as GenBank and UniProt were not designed for deep learning when they were

first established, the quality and quantity of data from these databases (especially data with annotations) can hardly meet the requirements of deep learning. To reduce the dependence of deep learning on big data, representation learning, and transfer learning provide more possibilities; however, how to balance global and local features requires further research. Second, due to the lack of experimental data on complex conformation, it remains challenging to represent and understand the association between protein conformation and function. Artificial intelligence technology has studied structure–function relationships much less than sequence–function relationships. Combining data with physical knowledge may provide more possibilities for this study. Finally, considering that most current sequence–function relationship models are built for a certain superfamily, the generality of the deep learning model to cover multiple enzyme families would greatly promote the development of artificial intelligence-based enzyme engineering. With the development of technology and the accumulation of data, artificial intelligence will become a powerful tool for protein engineering, assisting biocatalysis and synthetic biology in solving key problems in the field of biopharmaceuticals, and promoting drug development and production in the future.

Funding
This work was supported by the National Natural Science Foundation of China (Grant No. 22208217) for Jiang, and the Shanghai Pujiang Program (Grant No. 21PJ1423200) and the Program of Shanghai Academic/Technology Research Leader (Grant No. 23XD1435000) for Yi.

Conflict of Interest
None declared.

References

1 Devine PN, Howard RM, Kumar R, et al. Extending the application of biocatalysis to meet the challenges of drug development. Nat Rev Chem 2018;2:409–421
2 Adams JP, Brown MJB, Diaz-Rodriguez A, et al. Biocatalysis: a pharma perspective. Adv Synth Catal 2019;361:2421–2432
3 Stepan AF, Tran TP, Helal CJ, et al. Late-stage microsomal oxidation reduces drug-drug interaction and identifies phosphodiesterase 2A inhibitor PF-06815189. ACS Med Chem Lett 2018;9 (02):68–72
4 Charlton SN, Hayes MA. Oxygenating biocatalysts for hydroxyl functionalisation in drug discovery and development. ChemMedChem 2022;17(12):e202200115
5 Fuchs CS, Farnberger JE, Steinkellner G, et al. Asymmetric amination of α-chiral aliphatic aldehydes via dynamic kinetic resolution to access stereocomplementary brivaracetam and pregabalin precursors. Adv Synth Catal 2018;360(04):768–778
6 Ali M, Ishqi HM, Husain Q. Enzyme engineering: reshaping the biocatalytic functions. Biotechnol Bioeng 2020;117(06): 1877–1894
7 Victorino da Silva Amatto I, Gonsales da Rosa-Garzon N, Antônio de Oliveira Simões F, et al. Enzyme engineering and its industrial applications. Biotechnol Appl Biochem 2022;69(02):389–409
8 Campbell E, Kaltenbach M, Correy GJ, et al. The role of protein dynamics in the evolution of new enzyme function. Nat Chem Biol 2016;12(11):944–950
9 Curado-Carballada C, Feixas F, Iglesias-Fernández J, Osuna S. Hidden conformations in Aspergillus niger monoamine oxidase are key for catalytic efficiency. Angew Chem Int Ed Engl 2019;58 (10):3097–3101
10 Petrović D, Risso VA, Kamerlin SCL, Sanchez-Ruiz JM. Conformational dynamics and enzyme evolution. J R Soc Interface 2018;15 (144):20180330
11 Wrenbeck EE, Azouz LR, Whitehead TA. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. Nat Commun 2017;8:15695
12 Currin A, Swainston N, Day PJ, Kell DB. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. Chem Soc Rev 2015;44(05):1172–1239
13 Obexer R, Godina A, Garrabou X, et al. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. Nat Chem 2017;9(01):50–56
14 Jiménez-Osés G, Osuna S, Gao X, et al. The role of distant mutations and allosteric regulation on LovD active site dynamics. Nat Chem Biol 2014;10(06):431–436
15 Wang Y, Xue P, Cao M, Yu T, Lane ST, Zhao H. Directed Evolution: Methodologies and Applications. Chem Rev 2021;121(20): 12384–12444
16 Longwell CK, Labanieh L, Cochran JR. High-throughput screening technologies for enzyme engineering. Curr Opin Biotechnol 2017;48:196–202
17 Jiang S, Zhang L, Yao Z, et al. Switching a nitrilase from Syechocystis sp. PCC6803 to a nitrile hydratase by rationally regulating reaction pathways. Catal Sci Technol 2017;7(05):1122–1128
18 Ferguson AL, Ranganathan R. 100th anniversary of macromolecular science viewpoint: data-driven protein design. ACS Macro Lett 2021;10(03):327–340
19 Hossack EJ, Hardy FJ, Green AP. Building enzymes through design and evolution. ACS Catal 2023;13(19):12436–12444
20 Paladino A, Marchetti F, Rinaldi S, Colombo G. Protein design: from computer models to artificial intelligence. Wiley Interdiscip Rev Comput Mol Sci 2017;7:e1318
21 Yi D, Bayer T, Badenhorst CPS, et al. Recent trends in biocatalysis. Chem Soc Rev 2021;50(14):8003–8049
22 Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. IEEE Access 2019;7:53040–53065
23 Dongare AD, Kharde RR, Kachare AD. Introduction to artificial neural network. Int J Eng Innov Technol 2012;2(01):189–194
24 Xu Y, Verma D, Sheridan RP, et al. Deep dive into machine learning models for protein engineering. J Chem Inf Model 2020;60(06):2773–2790
25 Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput 2019;31 (07):1235–1270
26 Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 2019;16(12): 1315–1322
27 Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks. Pattern Recognit 2018;77:354–377
28 Li F, Yuan L, Lu H, et al. Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction. Nat Catal 2022;5(08):662–672
29 Zhang S, Tong H, Xu J, Maciejewski R. Graph convolutional networks: a comprehensive review. Comput Soc Netw 2019;6 (01):11
30 Zhang ZB, Xu MH, Jamasb A, et al. Protein representation learning by geometric structure pretraining. *arXiv*. Preprint. September 19, 2022. Available from: https://doi.org/10.48550/arXiv.2203.06125
31 Pu Y, Gan Z, Henao R, et al. Variational autoencoder for deep learning of images, labels and captions. Paper presented at: Proceedings of the 30th International Conference on Neural

Information Processing Systems, December 2016; Barcelona, Spain: 2360–2368
32 Hawkins-Hooker A, Depardieu F, Baur S, Couairon G, Chen A, Bikard D. Generating functional protein variants with variational autoencoders. PLOS Comput Biol 2021;17(02):e1008736
33 Wang K, Gou C, Duan Y, et al. Generative adversarial networks: introduction and outlook. IEEE CAA J Automatic 2017;4(04): 588–598
34 Wu Z, Johnston KE, Arnold FH, Yang KK. Protein sequence design with deep generative models. Curr Opin Chem Biol 2021; 65:18–27
35 Zhuang F, Qi Z, Duan K, et al. A comprehensive survey on transfer learning. Proc IEEE 2021;109(01):43–76
36 Luo Y, Jiang G, Yu T, et al. ECNet is an evolutionary context-integrated deep learning framework for protein engineering. Nat Commun 2021;12(01):5743
37 Mistry J, Chuguransky S, Williams L, et al. Pfam: the protein families database in 2021. Nucleic Acids Res 2021;49: D412–D419
38 Madani A, Krause B, Greene ER, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol 2023;41:1099–1106
39 Church KW. Word2Vec. Nat Lang Eng 2016;23(01):155–162
40 Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv*. Preprint. September 7, 2013. Available from: https://doi.org/10.48550/arXiv.1301.3781
41 Jeffrey P, Richard S, Manning C. GloVe: global vectors for word representation. Paper presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 2014; Doha, Qatar: 1532–1543
42 Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. Paper presented at: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics; April 3–7, 2017; Valencia, Spain; Volumn 2: 427–431
43 Matthew EP, Mark N, Mohit I, et al. Deep contextualized word representations. Paper presented at: Proceedings of NAACL-HLT; June 1-6, 2018; New Orleans, Louisiana: 2227–2237
44 Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P Language Models are Few-Shot Learners. Paper presented at: Proceedings of the 34th International Conference on Neural Information Processing Systems; December 2020; Vancouver, Canada: 1877–1901
45 Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Paper presented at: Proceedings of NAACL-HLT; June 2–7, 2019; Minneapolis, Minnesota: 4171–4186
46 Zhou GM, Gao ZF, Ding QK, et al. Uni-Mol: A universal 3D molecular representation learning framework. Paper presented at: The Eleventh International Conference on Learning Representations; May 1–5, 2023; Kigali, Rwanda
47 Cramer P. AlphaFold2 and the future of structural biology. Nat Struct Mol Biol 2021;28(09):704–705
48 Feehan R, Montezano D, Slusky JSG. Machine learning for enzyme engineering, selection and design. Protein Eng Des Sel 2021;34:gzab019
49 Ovek D, Abali Z, Zeylan ME, Keskin O, Gursoy A, Tuncbag N. Artificial intelligence based methods for hot spot prediction. Curr Opin Struct Biol 2022;72:209–218
50 Wittmann BJ, Johnston KE, Wu Z, Arnold FH. Advances in machine learning for directed evolution. Curr Opin Struct Biol 2021;69:11–18
51 Hie BL, Yang KK. Adaptive machine learning for protein engineering. Curr Opin Struct Biol 2022;72:145–152
52 UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49(D1):D480–D489
53 Burley SK, Bhikadiya C, Bi C, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res 2021;49(D1):D437–D451
54 Nikam R, Kulandaisamy A, Harini K, Sharma D, Gromiha MM. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Res 2021;49 (D1):D420–D424
55 Stourac J, Dubrava J, Musil M, et al. FireProtDB: database of manually curated protein stability data. Nucleic Acids Res 2021; 49(D1):D319–D324
56 Yan B, Ran X, Gollu A, et al. IntEnzyDB: an integrated structure-kinetics enzymology database. J Chem Inf Model 2022;62(22): 5841–5848
57 Drula E, Garron ML, Dogan S, Lombard V, Henrissat B, Terrapon N. The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res 2022;50(D1):D571–D577
58 Adler BA, Trinidad MI, Bellieny-Rabelo D, et al. CasPEDIA Database: a functional classification system for class 2 CRISPR-Cas enzymes. Nucleic Acids Res 2024;52(D1):D590–D596Erratum in: Nucleic Acids Res 2024;52(02):1002
59 Li F, Chen Y, Anton M, Nielsen J. GotEnzymes: an extensive database of enzyme parameter predictions. Nucleic Acids Res 2023;51(D1):D583–D586
60 Yin J, Li F, Zhou Y, et al. INTEDE: interactome of drug-metabolizing enzymes. Nucleic Acids Res 2021;49(D1):D1233–D1243
61 Ribeiro AJM, Holliday GL, Furnham N, Tyzack JD, Ferris K, Thornton JM. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucleic Acids Res 2018;46(D1):D618–D623
62 Caspi R, Billington R, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res 2020;48(D1):D445–D453
63 Wittig U, Rey M, Weidemann A, Kania R, Müller W. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. Nucleic Acids Res 2018;46(D1):D656–D660
64 Nagano N, Nakayama N, Ikeda K, et al. EzCatDB: the enzyme reaction database, 2015 update. Nucleic Acids Res 2015;43 (Database issue, D1):D453–D458
65 Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. Comput Struct Biotechnol J 2021;19:1750–1758
66 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(15): e2016239118
67 Villegas-Morcillo A, Makrodimitris S, van Ham RCHJ, Gomez AM, Sanchez V, Reinders MJT. Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. Bioinformatics 2021;37(02): 162–170
68 Wang L, Zhao KY. Detecting “protein words” through unsupervised word. F1000Research 2015;4:1517
69 Asgari E, McHardy AC, Mofrad MRK. Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). Sci Rep 2019;9(01):3577
70 Yang KK, Wu Z, Bedbrook CN, Arnold FH. Learned protein embeddings for machine learning. Bioinformatics 2018;34 (15):2642–2648
71 Littmann M, Heinzinger M, Dallago C, Weissenow K, Rost B. Protein embeddings and deep learning predict binding residues for various ligand classes. Sci Rep 2021;11(01):23916
72 Väth P, Münch M, Raab C, et al. PROVAL: a framework for comparison of protein sequence embeddings. J Comput Math Data Sci 2022;3:100044

73 Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: towards cracking the language of life's code through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2021;14:8

74 Rao R, Bhattacharya N, Thomas N, et al. Evaluating Protein Transfer Learning with TAPE. Paper presented at: Proceedings of the 33rd International Conference on Neural Information Processing Systems; December 2019; Vancouver, Canada: 9689–9701

75 Min S, Park S, Kim S, et al. Pre-training of deep bidirectional protein sequence representations with structural information. IEEE Access 2021;9:123912–123926

76 Webb B, Sali A. Comparative protein structure modeling using MODELLER. Curr Protoc Bioinformatics 2016;54:5.6.1–5.6.37

77 Buller R, Lutz S, Kazlauskas RJ, Snajdrova R, Moore JC, Bornscheuer UT. From nature to industry: Harnessing enzymes for biocatalysis. Science 2023;382(6673):eadh8615

78 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–589

79 Perrakis A, Sixma TK. AI revolutions in biology: the joys and perils of AlphaFold. EMBO Rep 2021;22(11):e54046

80 Bertoline LMF, Lima AN, Krieger JE, Teixeira SK. Before and after AlphaFold2: an overview of protein structure prediction. Front Bioinform 2023;3:1120370

81 Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379(6637):1123–1130

82 Gao Z, Tan C, Wu L, Stan Z. CoSP: Co-supervised pretraining of pocket and ligan. Paper presented at: the next European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases; September 18–22, 2023; Turin, Italy

83 Zhang Z, Xu M, Jamasb A, et al. Protein representation learning by geometric structure pretraining. Paper presented at: The Eleventh International Conference on Learning Representations; May 1–5, 2023; Kigali, Rwanda

84 Torng W, Altman RB. High precision protein functional site detection using 3D convolutional neural networks. Bioinformatics 2019;35(09):1503–1512

85 Koohi-Moghadam M, Wang H, Wang Y, et al. Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach. Nat Mach Intell 2019;1(12):561–567

86 Mansoor S, Baek M, Madan U, Horvitz E. Toward more general embeddings for protein design harnessing joint representations of sequence and structure. *bioRxiv*. Preprint. September 1, 2021. Available from: https://doi.org/10.1101/2021.09.01.458592

87 Wang Z, Combs SA, Brand R, et al. LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction. Sci Rep 2022;12(01):6832

88 Wu Z, Kan SBJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc Natl Acad Sci USA 2019;116(18):8852–8858

89 Shroff R, Cole AW, Diaz DJ, et al. Discovery of novel gain-of-function mutations guided by structure-based deep learning. ACS Synth Biol 2020;9(11):2927–2935

90 Lu H, Diaz DJ, Czarnecki NJ, et al. Machine learning-aided engineering of hydrolases for PET depolymerization. Nature 2022;604(7907):662–667

91 Thean DGL, Chu HY, Fong JHC, et al. Machine learning-coupled combinatorial mutagenesis enables resource-efficient engineering of CRISPR-Cas9 genome editor activities. Nat Commun 2022;13(01):2219

92 Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. Low-N protein engineering with data-efficient deep learning. Nat Methods 2021;18(04):389–396

93 Strokach A, Kim PM. Deep generative modeling for protein design. Curr Opin Struct Biol 2022;72:226–236

94 Osadchy M, Kolodny R. How deep learning tools can help protein engineers find good sequences. J Phys Chem B 2021;125(24):6440–6450

95 Greener JG, Moffat L, Jones DT. Design of metalloproteins and novel protein folds using variational autoencoders. Sci Rep 2018;8(01):16189

96 Trinquier J, Uguzzoni G, Pagnani A, Zamponi F, Weigt M. Efficient generative modeling of protein sequences using simple autoregressive models. Nat Commun 2021;12(01):5800

97 Shin JE, Riesselman AJ, Kollasch AW, et al. Protein design and variant prediction using autoregressive generative models. Nat Commun 2021;12(01):2403

98 Castro E, Godavarthi A, Rubinfien J, et al. Transformer-based protein generation with regularized latent space optimization. Nat Mach Intell 2022;4(10):840–851

99 Lobzaev E, Herrera MA, Campopiano DJ, et al. Designing human Sphingosine-1-phosphate lyases using a temporal Dirichlet variational autoencoder. *bioRxiv*. Preprint. February 15, 2022. Doi: 10.1101/2022.02.14.480330

100 Giessel A, Dousis A, Ravichandran K, et al. Therapeutic enzyme engineering using a generative neural network. Sci Rep 2022;12(01):1536

101 Repecka D, Jauniskis V, Karpus L, et al. Expanding functional protein sequence spaces using generative adversarial networks. Nat Mach Intell 2021;3(04):324–333

102 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Paper presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; December 2017; Long Beach, CA, United States: 6000–6010

103 Sevgen E, Moller J, Lange A, et al. ProT-VAE: protein transformer variational autoencoder for functional protein design. *bioRxiv*. Preprint. January 23, 2023. Available from: https://doi.org/10.1101/2023.01.23.525232