Chapter 2

Machine Learning for Protein Function Prediction

Yi-Heng Zhu, Zi Liu, Yu Ding, Zhiwei Ji, and Dong-Jun Yu

Abstract

Knowledge of protein functions is crucial to understanding and investigating cellular functions across all organisms. Accurate annotations of protein functions are also useful for the elucidation of mechanisms of various diseases and can be used to guide target-based drug design efforts. Although biological experiments are the most precise way for functional annotation of proteins, they are often time-consuming, laborious, and expensive. Therefore, there is an urgent need to develop efficient and accurate computational approaches for protein function prediction. This chapter comprehensively reviews and categorizes prominent computational predictors of protein functions, which are defined by the Gene Ontology (GO) terms, including template detection-based methods, statistical machine learning-based methods, deep learning-based methods, and composition methods. Applications of those protein function prediction methods are also discussed.

Key words Protein function prediction, Gene Ontology, Machine learning, Deep learning, Template detection

1 Introduction

Proteins perform various cellular functions including catalyzing biochemical reactions, transporting substances, transmitting signals, regulating metabolism, and providing immune protection [1]. Accurate annotation of protein functions is critical to reveal molecular-level details of these functions and elucidate mechanisms underlying diseases, thereby guiding targeted drug design [2, 3]. In view of this, protein function annotation has become one of the primary tasks in the post-genomic era [4].

In the early research, protein functions were mainly annotated via experimental methods [5]. Although these methods are the most precise way for function annotation, they tend to be time-consuming, laborious, and expensive, leading to relatively slow growth of the functionally annotated protein data. In stark

The authors Yi-Heng Zhu, Zi Liu, and Yu Ding are first senior authors on this work.

Lukasz Kurgan and Daisuke Kihara (eds.), *Protein Function Prediction: Methods and Protocols*, Methods in Molecular Biology, vol. 2947, https://doi.org/10.1007/978-1-0716-4662-5_2,

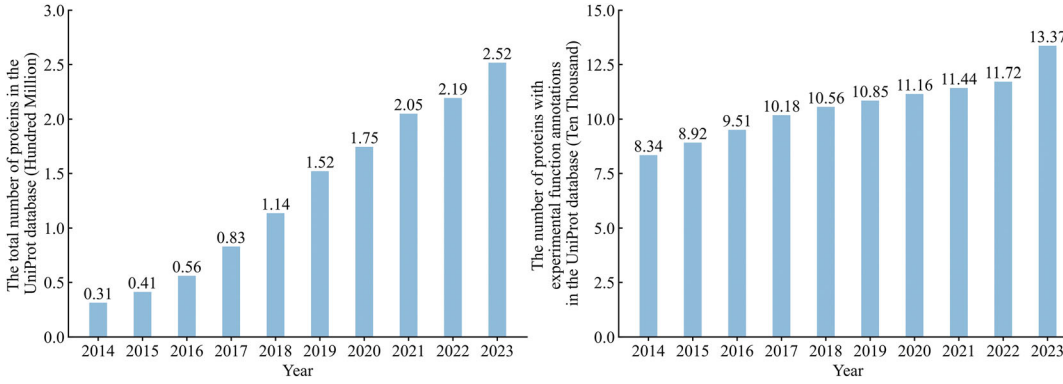

Fig. 1 The growth trends over the past decade in the total number of sequences and the number of sequences with biological-experimental function annotations in the UniProt database

contrast, protein sequences are produced at a very high pace, leading to their rapid accumulation in public databases. Figure 1 shows the corresponding growth trends over the past decade in the total number of sequences and the number of sequences with experimental function annotations in the UniProt protein database [6]. As of January 2024, the UniProt database provided access to 252 million protein sequences, yet fewer than 0.1% of them had function annotations. To fill this gap, it is urgent to develop efficient computational methods to rapidly and accurately predict functions from sequences.

Recently, a large number of computational methods have emerged for protein function prediction [7–9]. These methods typically rely on knowledge-based models that are trained/generated from the protein data with known functions and which can be subsequently used to infer the functions directly from sequences of proteins. Development of these tools is cross-disciplinary including computer science, statistics, and molecular biology. The current function prediction methods could be roughly divided into three categories including template detection-based, machine learning-based, and composition methods. This chapter presents a comprehensive review of representative methods across the three categories.

2 Protein Function Annotation Database

2.1 Gene Ontology

The protein functions are mainly annotated by Gene Ontology (GO) [10], which is an important bioinformatics initiative that standardizes the representation of attributes for genes and gene products (e.g., proteins and RNA molecules) across all species. For a given target protein, the corresponding functions are divided into three aspects using GO annotations, including molecular

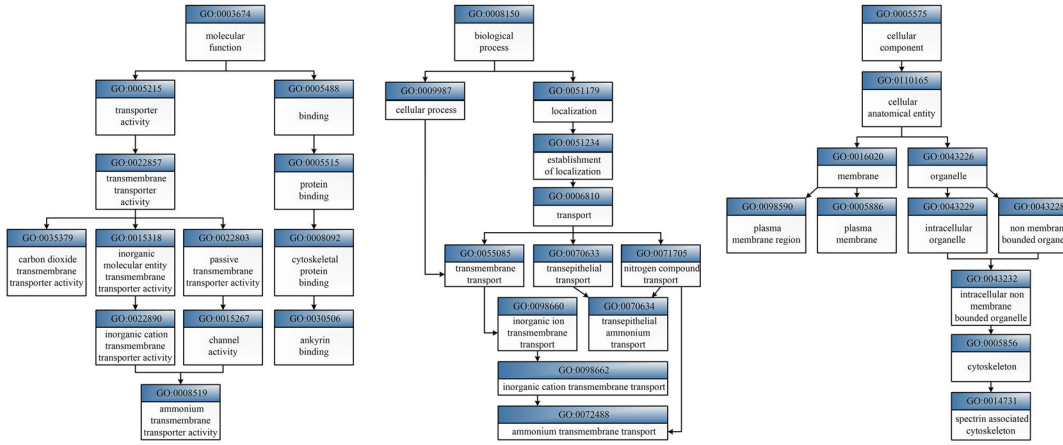

Fig. 2 The function annotations of ammonium transporter Rh type B protein (UniProt ID: Q9H310) in three GO aspects

function (MF), biological process (BP), and cellular component (CC) [11]. MF describes the elemental activities of proteins at the molecular level, such as ligand-binding and catalysis. BP refers to the complete biological process accomplished by several molecular activities in which proteins participate, such as signal transduction and metabolism. CC captures the subcellular location where the proteins are active, such as the cell nucleus and mitochondria.

Figure 2 illustrates the function annotations in three GO aspects for the human ammonium transporter Rh type B protein (UniProt ID: Q9H310) [12]. In each aspect, the functions of this protein are represented as a directed acyclic graph, where nodes denote GO terms and edges indicate the parent-child relationship between GO terms. Protein function prediction methods aim to accurately predict the GO terms with the corresponding parent-child relationship from the protein sequences.

2.2 Gene Ontology Annotation Databases

Gene Ontology Annotation (GOA) database (https://www.ebi.ac.uk/GOA/) [13] is the most commonly used function annotation database, which was established by the European Bioinformatics Institute (EMBL-EBI) in 2004. This database provides high-quality functional annotations (i.e., GO annotations) for proteins from the UniProt database, RNA molecules from the RNACentral database [14], and protein complexes from the Complex Portal database [15]. Each record in GOA consists of the name or identifier of protein/RNA/complex, reference database, GO aspect, GO terms, annotation date, and annotation method. Among these, eight annotation methods derived from biological experiments are considered as gold standard, including inference from experiment (EXP), inference from direct assay (IDA), inference from physical

interaction (IPI), inference from mutant phenotype (IMP), inference from genetic interaction (IGI), inference from expression pattern (IEP), traceable author statement (TAS), and inference by curator (IC). As of January 2024, the GOA database contains 133.7 thousand protein sequences with experimental GO annotations. There are also several other protein function databases including InterPro [16], Reactome [17], KEGG [18], and BRENDA [19].

3 Computational Methods for Protein Function Prediction

Since 2010, much progress has been made in the protein function prediction field, primarily driven by the Critical Assessment of Functional Annotation (CAFA) challenges [20]. CAFA is a global competition that evaluates function predictors, which is held every 3 years. CAFA provides a fair and objective competition platform, attracting numerous experts to develop a series of efficient function prediction methods. These methods could be categorized into the three groups that we discuss in the following subsections.

3.1 Template Detection-Based Methods

In the early stages of the protein function prediction field, template detection-based methods were predominant [21, 22]. The foundational principle for these methods is that if two proteins show similarity in certain biological attributes (such as sequence and structure), they are likely to perform similar functions. The key principles of the template detection-based methods are as follows. For a given query protein, they identify the corresponding functional templates from a public database that share similar biological attributes; these templates are then used to infer the biological functions of the query protein. According to the employed biological attributes, the template detection-based methods could be roughly divided into five categories, including sequence alignment-based, structure alignment-based, interaction network-based, family transference-based, and multi-attribute fusion-based methods, with the details in Table 1.

Sequence alignment-based methods utilize sequence alignment tools, such as PSI-BLAST [38] and HHblits [39], to measure the sequence similarity between proteins, which is further used as the metric to select functional templates. The representative examples include Gotcha [23], Blast2GO [24], and GoFDR [25], where GoFDR was ranked as 4, 2, and 2 at the MF, BP, and CC predictions, respectively, in the second CAFA competition (CAFA2) [40].

Structure alignment-based methods use structure alignment tools, such as TM-align [41] and DALI [42], to quantify the similarity of three-dimensional structures between query and candidate proteins to select templates. If the structure of the query is

Table 1
Summary of 16 state-of-the-art template detection-based methods for protein function prediction

Type	Method	Ref[a]	Year	Availability
Sequence alignment	Gotcha	[23]	2004	NA[b]
	Blast2GO	[24]	2005	NA
	GoFDR	[25]	2017	NA
Structure alignment	ProFunc	[26]	2005	NA
	FINDSITE	[27]	2009	NA
	COFACTOR	[28]	2012	https://zhanggroup.org/COFACTOR/
Interaction network	Letovsky's method	[29]	2003	NA
	Vazquez's method	[30]	2003	NA
	Chua's method	[31]	2004	NA
Family transference	MultiPfam2GO	[32]	2008	NA
	dcGO	[33]	2013	https://supfam.org/SUPERFAMILY/dcGO/
	FunFams	[34]	2015	http://www.cathdb.info/search/by_sequence
Multi-attribute fusion	MS-kNN	[35]	2013	NA
	INGA	[36]	2015	http://protein.bio.unipd.it/inga
	MetaGO	[37]	2018	https://zhanggroup.org/MetaGO/
	QAUST	[21]	2021	http://www.cbrc.kaust.edu.sa/qaust/submit

[a] *Ref* reference
[b] *NA* not available

unavailable, then the structure prediction tools, such as AlphaFold2 [43] and I-TASSER [44], are utilized to predict the corresponding structure from the sequence. There are several representative examples including ProFunc [26], FINDSITE [27], and COFACTOR [28].

Interaction network-based methods search for proteins that interact with the query protein based on the protein–protein interaction network databases (e.g., STRING [45] and PrePPI [46]), and use them as templates. Some of the key examples are Letovsky's method [29], Vazquez's method [30], and Chua's method [31].

Family transference-based methods integrate the hidden Markov models [47] with multiple sequence alignments to inter the family of a given query protein, and functions of this protein family are transferred into the query. A few illustrative methods in this category include MultiPfam2GO [32], dcGO [33], and FunFams [34], where FunFams achieved the third rank overall at the predictions of three GO aspects in the fourth CAFA competition (CAFA4) [48].

Multi-attribute fusion-based methods first design function prediction sub-methods using different biological attributes and the function predictions are derived by fusing their predictions

together. For instance, the INGA method [36] was ranked 5, 3, and 2 for the MF, BP, and CC predictions, respectively, in the third CAFA competition (CAFA3) [20], combining results of three sub-methods that rely on the sequence similarity, structural domain similarity, and interaction network. Other example methods in this category are MS-kNN [35], MetaGO [37], and QAUST [21].

We note an inevitable drawback for the template detection-based methods, the fact that their prediction performance highly depends on the availability and quality of the functional templates. If the templates with high quality are unavailable, the corresponding prediction accuracy would significantly decline.

3.2 Machine Learning-Based Methods

To overcome the drawbacks of the template detection-based methods, machine learning algorithms are used as an alternative to develop protein function prediction methods [49]. These methods aim to encode proteins as feature vectors or matrices from biological views, which are processed with machine learning algorithms to train function prediction models. Machine learning-based methods could be further divided into two groups that we discuss in the following two subsections.

3.2.1 Statistical Machine Learning-Based Methods

For earlier predictors, researchers manually designed feature representations for proteins, such as position-specific scoring matrices, physicochemical property vectors, and secondary structure matrices, which are then combined with statistical machine learning algorithms (e.g., support vector machines [50] and Bayesian estimation [51]) to implement function prediction models. Representative examples include CHUGO [52], Lee's method [53], FFPred [54], GOPred [55], Jeong's method [56], TMEC [57], HMC-LMLP [58], GOLabeler [59], and MLC [60], where GOLabeler achieved the first rank in all three GO predictions of CAFA3 [20] through integrating logistic regression model with multiple sequence-based feature representations. Table 2 summarizes the further details of the above-mentioned methods.

Although machine learning methods complement the template detection-based methods, their prediction accuracy could be unsatisfactory. The major reason for this is that the feature representations could be poorly designed and/or relatively simple, failing to extract relevant and useful knowledge from the input sequences.

3.2.2 Deep Learning-Based Methods

To solve potential shortcomings of manually designed feature representations, deep learning techniques, which have been transferred from the field of computer vision, were applied in recent years [61]. An advantage of deep learning algorithms lies in their ability to design complex deep neural network architectures tailored for different data structures that can be used to describe proteins, such as one-dimensional sequences, two-dimensional contact maps, and three-dimensional atomic coordinates. This

Table 2
Summary of nine statistical machine learning-based methods for protein function prediction

Method	Ref[a]	Year	Classifier	Availability
CHUGO	[52]	2005	Support vector machine	NA[b]
Lee's method	[53]	2006	Kernel-based logistic regression	NA
FFPred	[54]	2008	Support vector machine	http://bioinfadmin.cs.ucl.ac.uk/downloads/ffpred/
GOPred	[55]	2010	Support vector machine	http://kinaz.fen.bilkent.edu.tr/gopred
Jeong's method	[56]	2011	Support vector machine	NA
TMEC	[57]	2013	Directed birelational graph	https://sites.google.com/site/guoxian85/tmec
HMC-LMLP	[58]	2016	Multilayer perceptron	http://sites.google.com/site/cerrirc/downloads
GOLabeler	[59]	2018	Logistic regression	http://datamining-iip.fudan.edu.cn/golabeler
MLC	[60]	2020	K-nearest neighbors	www.github.com/stamakro/MLC

[a] *Ref* reference
[b] *NA* not available

capability allows for deep and multi-view extraction of the knowledge that can be derived from the input sequences, thereby potentially enhancing the scope of feature representation. According to whether the pretrained is involved, deep learning-based methods can be roughly classified into three categories, including direct training-based, pretrained language model-based, and biological large language model-based methods.

Early deep learning methods directly trained function prediction models on the protein datasets with known functional annotations through integrating deep neural networks (e.g., convolutional neural network [62] and recurrent neural network [63]) with sequence encoding. Representative examples include DeepGO [61], DeepGOA [64], FFPred-GAN [8], TALE [65], DeepPFP-CO [66], and DeepGOZero [67]. DeepGO is known as the first deep learning-based method for functional prediction, achieving the third rank in the MF prediction of CAFA3 [20]. In addition, several methods incorporate structure and interaction network knowledge into sequence data to train prediction models, with selected examples of MultiPredGO [68] and DeepGraphGO [69], and further details in Table 3.

Compared to the statistical machine learning-based methods, the above-mentioned deep learning methods on average tend to produce more accurate predictions. However, there is still room for

Table 3
Summary of eight state-of-the-art direct training-based methods in deep learning-based protein function prediction

Method	Ref[a]	Year	Deep neural network model	Availability
DeepGO	[61]	2018	Convolutional neural network	https://deepgo.cbrc.kaust.edu.sa/
DeepGOA	[64]	2020	Long- and short-time memory	https://github.com/CSUBioGroup/DeepGOA
FFPred-GAN	[8]	2020	Generative adversarial network	https://github.com/psipred/FFPredGAN
MultiPredGO	[68]	2020	Residual neural network	https://github.com/SwagarikaGiri/MultiPredGO
DeepGraphGO	[69]	2021	Graph convolutional network	https://github.com/yourh/DeepGraphGO
TALE	[65]	2021	Attention network	https://github.com/Shen-Lab/TALE
DeepPFP-CO	[66]	2022	Graph convolutional network	https://csuligroup.com/DeepPFP/
DeepGOZero	[67]	2022	Fully connect network	https://github.com/bio-ontology-research-group/deepgozero

[a] *Ref* reference

further improvement. Since these deep-learning algorithms train prediction models on the protein datasets with known function annotations, their performance highly depends on the scale of these training datasets. If the amount of training data is limited, deep learning models cannot comprehensively learn relationships between protein sequences/structures and their functions, potentially leading to unsatisfactory prediction performance. As of January 2024, the GOA database has ~133,700 proteins with function annotations via biological experiments, where 81.52% of function terms are associated with fewer than 50 protein entries. This amount of training data might be insufficient to train high-accuracy deep learning models, especially if the corresponding networks are excessively large.

To address the issues that stem from an insufficient amount of training data, the pretraining strategy has been utilized in the protein function prediction field. First, deep learning techniques are used to train an unsupervised language model on a large number of protein sequences without functional annotations by considering available evolutionary, structural, and functional knowledge. Then, these language models are used to encode protein sequences as feature embeddings, which are fed into supervised deep neural networks to train function prediction models. Taking DeepFRI [7]

Table 4
Summary of eight popular pretrained language model-based methods in deep learning-based protein function prediction

Method	Ref[a]	Year	Network models[b]	Availability
deepNF	[70]	2018	FCN (3, 0.02 M) + SVM	https://github.com/VGligorijevic/deepNF
DeepFRI	[7]	2021	LSTM (2, 10 M) + GCN	https://beta.deepfri.flatironinstitute.org/
MGEGFP	[71]	2022	GCN (3, 0.02) + LightGBM	https://github.com/zhanglabNKU/MGEGFP
Domain-PFP	[72]	2023	FCN (2, 0.53 M) + FCN	https://github.com/kiharalab/Domain-PFP
CFAGO	[73]	2023	AN-MLP (1, 2, 0.01 M) + MLP	http://bliulab.net/CFAGO/
HiFun	[74]	2023	FCN (2, 0.34 M) + CNN-LSTM-AN	http://www.unimd.org/HiFun
PFmulDL	[75]	2022	CNN (2, 0.03 M) + CNN	https://github.com/idrblab/PFmulDL
AnnoPRO	[76]	2024	CNN-FCN (1, 5, 0.09 M) + LSTM	https://github.com/idrblab/AnnoPRO

FCN Fully connected network, *SVM* support vector machine, *LSTM* long- and short-time memory, *GCN* graph convolution network, *LightGBM* light gradient boosting machine, *AN-MLP* the combination of attention network with one layer and multilayer perceptron with two layers, *MLP* multilayer perceptron, *CNN-LSTM-AN* the combination of convolution neural network, long- and short-time memory, and attention network, *CNN* convolution neural network, *CNN-FCN* the combination of convolution neural network with one layer and fully connected network with five layers
[a] *Ref* reference
[b] Network models consist of a pretrained language model (number of layers, number of training sequences) for feature embeddings and a supervised training model for function prediction

as an example, it utilizes the long short-term memory network to train an unsupervised protein language model on more than 10 million sequences, which are integrated with the supervised graph convolutional neural network that implements the function prediction model. Other representative examples include deepNF [70], MGEGFP [71], Domain-PFP [72], CFAGO [73], HiFun [74], PFmulDL [75], and AnnoPRO [76], which we summarize in Table 4.

Despite the encouraging performance of these methods, some challenges remain. Specifically, the size of datasets and the scale of neural networks are relatively small in the pretraining phase (see Table 4), thereby restricting further performance improvement. In recent years, a series of biological large language models, including SeqVec [77], TAPE [78], ESM-1b [79], ProtTrans [80], ESM2 [81], SaProt [82], Ankh [83], and CARP [84] (see details in Table 5), have emerged, achieving outstanding performance in numerous bioinformatics tasks, such as protein structure prediction [43, 81] and ligand-binding prediction [85]. Their performance advantages are primarily attributed to the large-scale training

Table 5
Summary of eight biological large language models employed in protein function prediction

Model	Ref[a]	Year	(Layers, Params)[b]	Availability
SeqVec	[77]	2019	(3, 93 M)	https://github.com/Rostlab/SeqVec
TAPE	[78]	2019	(12, 38 M)	https://github.com/songlab-cal/tape
ESM-1b	[79]	2021	(33, 650 M)	https://github.com/facebookresearch/esm
ProtTrans	[80]	2021	(24, 3B)	https://github.com/agemagician/ProtTrans
ESM2	[81]	2023	(48, 15B)	https://github.com/facebookresearch/esm
SaProt	[82]	2023	(33, 650 M)	https://github.com/westlake-repl/SaProt
Ankh	[83]	2023	(48, 1.15B)	https://github.com/agemagician/Ankh/tree/main
CARP	[84]	2024	(56, 640 M)	https://github.com/microsoft/protein-sequence-models

[a] *Ref* reference
[b] Layers, params: The number of layers and hyper-parameters for neural networks in biological large language models

datasets and highly complex neural networks. More specifically, most of the above-mentioned language models are deep neural networks with over 20 layers, trained on hundreds of millions of protein sequences. They have the capacity to effectively learn from these large sequence datasets, encoding sequences into highly dimensional feature representations that aim to encapsulate structural and functional patterns. In light of this, several protein language models have been utilized for protein function prediction, where they are used directly as pretrained language models to encode protein sequences. This has enhanced the accuracy of the function prediction. For instance, DeepGO-SE utilizes the ESM2 to extract the discriminative feature embeddings, which are further fed to multiple GO semantic entailment models for high-accuracy function prediction [86]. Other key examples include GOPredSim [87], PANDA2 [88], ATGO [89], GAT-GO [90], GNNGO3D [91], Struct2GO [92], MMSMAPlus [93], SPROF-GO [94], HNetGO [95], HEAL [96], TransFun [97], TEMPROT [98], PredGO [99], and DeepGOMeta [100], which we summarize in Table 6.

In summary, deep learning-based methods have become the mainstream approach in the field of protein function prediction, and their prediction performance often surpasses that of the template detection-based and statistical machine learning-based methods. However, their drawback is the heavy dependency on large-scale training data and huge computational resources.

3.3 Composition Methods

Composition methods are designed by ensembling the prediction results of multiple template detection and machine learning-based methods, with the underlying aim to further improve prediction

Table 6
Summary of 15 state-of-the-art biological large language model-based methods in deep learning-based protein function prediction

Method	Ref[a]	Year	Network models[b]	Availability
GOPredSim	[87]	2021	SeqVec + KNN	https://embed.protein.properties/
PANDA2	[88]	2022	ESM-1b + GNN	http://dna.cs.miami.edu/PANDA2/
ATGO	[89]	2022	ESM-1b + FCN	https://zhanggroup.org/ATGO/
GAT-GO	[90]	2022	ESM-1b + GAT	NA[c]
GNNGO3D	[91]	2023	ESM-1b + GAT-GCN	NA
Struct2GO	[92]	2023	SeqVec + GCN	https://github.com/lyjps/Struct2GO
MMSMAPlus	[93]	2023	ProtTrans + CNN-AN	https://github.com/wzy-2020/MMSMAPlus
SPROF-GO	[94]	2023	ProtTrans + AN	https://github.com/biomed-AI/SPROF-GO/
HNetGO	[95]	2023	SeqVec + AN	https://github.com/BIOGOHITSZ/HNetGO
HEAL	[96]	2023	ESM-1b + GCN-AN	https://github.com/ZhonghuiGu/HEAL
TransFun	[97]	2023	ESM-1b + EGNN	https://github.com/jianlin-cheng/TransFun
TEMPROT	[98]	2023	ProtTrans + MLP	https://github.com/gabrielbianchin/TEMPROT/
PredGO	[99]	2023	ESM-1b + EGNN-AN	http://predgo.denglab.org/
DeepGO-SE	[86]	2024	ESM2 + MLP	https://github.com/bio-ontology-research-group/deepgo2
DeepGOMeta	[100]	2024	ESM2 + MLP	https://github.com/bio-ontology-research-group/deepgometa

KNN K-nearest neighbors, *GNN* graph neural network, *FCN* fully connected network, *GAT* graph attention network, *GAT-GCN* the combination of graph attention network and graph convolution network, *GCN* graph convolution network, *CNN-AN* the combination of convolution neural network and attention network, *AN* attention network, *GCN-AN* the combination of graph convolution network and attention network, *EGNN* Equivariant graph neural network, *MLP* multilayer perceptron, *EGNN-AN* the combination of equivariant graph neural network and attention network

[a] *Ref* reference

[b] Network models consist of a biological large language model for feature embeddings and a supervised training model for function prediction

[c] *NA* not available

accuracy [101, 102]. In the last CAFA competition (CAFA4), composition methods have shown great potential [48]. The representative example is NetGO, which achieves the first rank among all the competing methods for MF, BP, and CC predictions [103]. This method uses the rank learning algorithm to ensemble the prediction results of six template detection and machine

learning methods, which rely on different types of information that include sequence similarity, sequence composition, amino acid physicochemical property, protein family coding, naïve probability of GO terms, and protein–protein interaction network. Other examples of composition methods include DeepText2GO [104], DeepGOPlus [105], TALE+ [65], ATGO+ [89], TransFun + [97], and TEMPROT+ [98], which we detail in Table 7.

Compared to the use of individual predictors, composition methods potentially integrate more biological knowledge and thus could achieve higher prediction performance. Moreover, composition methods can effectively incorporate additional state-of-the-art methods by modifying decision fusion models, offering the advantage of being easy to update. However, if the training data and models of the individual sub-methods are not properly considered, composition methods are prone to parameter overfitting, leading to potentially negative effects on the predictive quality.

3.4 Evaluation Metric

3.4.1 Protein-Centric Metric

Three evaluation metrics have been widely used to evaluate the performance of protein function prediction methods, especially in the CAFA competitions, including the maximum F1-score ($F_{\max}$), minimum semantic distance ($S_{\min}$), and area under the precision-recall curve (AUPRC) [20, 106, 107].

$F_{\max}$ is a basic metric in binary classification, considered the most important metric in CAFA competition with the following definition:

$$F_{\max} = \max_{t} \left\{ \frac{2 \cdot \mathrm{Pre}(t) \cdot \mathrm{Rec}(t)}{\mathrm{Pre}(t) + \mathrm{Rec}(t)} \right\} \tag{1}$$

Table 7
Summary of seven state-of-the-art composition methods for protein function prediction

Method	Ref[a]	Year	(NT, NL)[b]	Availability
DeepText2GO	[104]	2018	(1, 4)	NA[c]
NetGO	[103]	2019	(2, 4)	https://dmiip.sjtu.edu.cn/ng3.0
DeepGOPlus	[105]	2020	(1, 1)	http://deepgoplus.bio2vec.net/
TALE+	[65]	2021	(1, 1)	https://github.com/Shen-Lab/TALE
ATGO+	[89]	2022	(1, 1)	https://zhanggroup.org/ATGO/
TransFun+	[97]	2023	(1, 1)	https://github.com/jianlin-cheng/TransFun
TEMPROT+	[98]	2023	(1, 1)	https://github.com/gabrielbianchin/TEMPROT/

[a] *Ref* reference
[b] NT and NL are the numbers of template dection-based and machine learning-based sub-methods, respectively, in the composition method
[c] *NA* not available

$$\mathrm{Pre}(t) = \frac{1}{m(t)} \cdot \sum_{i=1}^{m(t)} \frac{\sum_{j=1}^{N_G} 1\left(s_{ij} \geq t\right) \cdot I_{ij}}{\sum_{j=1}^{N_G} 1\left(s_{ij} \geq t\right)} \tag{2}$$

$$\mathrm{Rec}(t) = \frac{1}{N_P} \cdot \sum_{i=1}^{N_P} \frac{\sum_{j=1}^{N_G} 1\left(s_{ij} \geq t\right) \cdot I_{ij}}{\sum_{j=1}^{N_G} I_{ij}} \tag{3}$$

where t is a threshold ranging from 0 to 1, $\mathrm{Pre}(t)$ and $\mathrm{Rec}(t)$ are precision and recall, respectively, under the threshold t; s_{ij} is the confidence score that the i-th protein is associated with the j-th GO term by the function prediction model; $1(\cdot) = 1$, if the input is true; otherwise, $1(\cdot) = 0$; $I_{ij} = 1$, if the i-th protein is associated with the j-th GO term in the experimental annotations; otherwise, $I_{ij} = 0$; $m(t)$ is the number of proteins that have at least one GO term with a confidence score higher than t; N_P is the number of all test proteins, and N_G is the number of all GO terms.

AUPRC is a threshold-independent evaluation metric, calculated by the area under the precision-recall curve over all threshold values.

$S_{\min}$ is an information theoretic-based metric, defined as follows:

$$S_{\min} = \min_t \left\{ \sqrt{ru(t)^2 + mi(t)^2} \right\} \tag{4}$$

$$ru(t) = \frac{1}{N_P} \cdot \sum_{i=1}^{N_P} \sum_{j=1}^{N_G} ic(j) \cdot 1\left(s_{ij} < t\right) \cdot I_{ij} \tag{5}$$

$$mi(t) = \frac{1}{N_P} \cdot \sum_{i=1}^{N_P} \sum_{j=1}^{N_G} ic(j) \cdot 1\left(s_{ij} \geq t\right) \cdot \left(1 - I_{ij}\right) \tag{6}$$

$$ic(j) = log_2 \frac{1}{p(j|\mathrm{parent}(j))} \tag{7}$$

where $ru(t)$ and $mi(t)$ are the remaining uncertainty and misinformation, respectively, under the threshold t; $ic(j)$ is the information content of the j-th GO term, and $p(j|\,\mathrm{parent}(j))$ is the conditional probability of the j-th term given its parent terms within the hierarchical GO structure, with additional details in the reference [106].

3.4.2 Term-Centric Metric

The area under the receiver operating characteristic curve (AUROC) is a crucial metric for binary classification and has been extensively employed to assess the performance of function prediction methods at the term-centric level [108]. Specifically, for a

given GO term Q_i, each test protein is labeled as "1" or "0", where "1" indicates this protein is associated with Q_i in the experimental annotation. Next, a confidence score for Q_i is assigned to each protein by the function prediction model. Finally, the AUROC is utilized to assess the prediction performance of Q_i through integrating the confidence scores and labels for all test proteins.

3.5 Applications of Protein Function Prediction

Protein function prediction models have important applications in the following areas:

1. *Deciphering cellular processes.* Accurate protein function predictions can be used to elucidate the roles of proteins in various cellular processes, thereby knowing which proteins are involved in certain cellular pathways [109–111]. This helps to explain functions at the cellular level.
2. *Drug Discovery and Design*. Function annotations can indicate which proteins are involved in disease pathways, leading to the discovery of new drug targets [112–114]. Moreover, a comprehensive analysis of molecular functions (e.g., ligand-binding and enzymatic activity) can guide the drug design for a target protein [115].
3. *Protein Design*. High-accuracy protein function prediction models could identify proteins with functions that are not present or are rare in nature [116–118]. By analyzing the molecular mechanisms of these rare proteins, novel proteins with tailored functions for various applications can be designed in protein engineering.
4. *Functional Genomics.* By identifying the functions of proteins, we can locate the functionally related genes that are involved in the biological pathways or processes [119–121]. This analysis facilitates the discovery of gene relationship networks and regulatory mechanisms underlying complex biological phenomena.
5. *Evolutionary Study.* By analyzing the genomes of different species, we can identify genes and proteins that are conserved across evolutionary distances [122–124]. Accurate function prediction of these conserved proteins provides insights into fundamental biological processes that have been maintained throughout evolution.

4 Discussions

Protein function prediction, i.e., gene ontology prediction, can be viewed as a multi-label prediction task that can be solved via machine learning. Recently, machine learning, especially deep learning methods have achieved great progress in protein function

prediction. This chapter provides an overview of the protein function predictors, alongside the recent advancements in this field of research, which cover the following observations.

1. Template detection-based methods played a dominant role in the early stage of protein function prediction. These methods can be further divided into five categories, including sequence alignment, structure alignment, interaction network, family transference, and multi-attribute fusion-based methods.
2. Machine learning-based methods have emerged in recent years, including statistical machine learning- and deep learning-based methods. The latter has been the primary driving force behind the advancement of protein function prediction, with three development stages, i.e., direct training, pretrained language model, and biological large language model-based methods.
3. Composition methods combine the strengths of template detection and machine learning-based methods, further enhancing the accuracy of protein function prediction. In the last CAFA4 competition, the top performers are both composition methods, with the typical example of NetGO [103].

Despite significant progress, some challenges remain. First, most deep learning methods directly perform function prediction from sequence alone. With the rapid development of protein structure prediction models (e.g., AlphaFold2 [43] and ESMFold [81]), the structural information that is often relevant to function prediction should be considered at a greater depth. Moreover, considering that proteins are gene expression products, incorporating gene knowledge into function prediction could be a promising strategy to improve accuracy [125]. Studies along these lines are in progress [126–128].

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62372234, 62072243, 61772273, 62306142, and 62402227), the Natural Science Foundation of Jiangsu (BK20201304 and BK20211210), the Foundation of National Defense Key Laboratory of Science and Technology (JZX7Y202001SY000901), Jiangsu Funding Program for Excellent Postdoctoral Talent (No. 2023ZB224), Agricultural Science and Technology Innovation Foundation of Jiangsu Province (No. CX (23) 3125), the Department of Education of Jiangxi Province (GJJ2400905), and Fundamental Research Funds for the Central Universities (No. YDZX2024009 and No. YDZX2025024).

References

1. Watson JL, Juergens D, Bennett NR et al (2023) De novo design of protein structure and function with RFdiffusion. Nature 620(7976):1089–1100
2. Wang J, Lisanza S, Juergens D et al (2022) Scaffolding protein functional sites using deep learning. Science 377(6604):387–394
3. Baldwin ET, van Eeuwen T, Hoyos D et al (2024) Structures, functions and adaptations of the human LINE-1 ORF2 protein. Nature 626(7997):194–206
4. Eisenberg D, Marcotte EM, Xenarios I et al (2000) Protein function in the post-genomic era. Nature 405(6788):823–826
5. MacBeath G, Schreiber SL (2000) Printing proteins as microarrays for high-throughput function determination. Science 289(5485): 1760–1763
6. Consortium TU (2022) UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res 51(D1):D523–D531
7. Gligorijevi V, Renfrew PD, Kosciolek T et al (2021) Structure-based protein function prediction using graph convolutional networks. Nat Commun 12(1):3168
8. Wan C, Jones DT (2020) Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. Nat Mach Intell 2(9):540–550
9. Kulmanov M, Guzmán-Vega FJ, Duek Roggli P et al (2024) Protein function prediction as approximate semantic entailment. Nat Mach Intell 6(1):220–228
10. Aleksander SA, Balhoff J, Carbon S et al (2023) The gene ontology knowledgebase in 2023. Genetics 224(1):iyad031
11. Huang K-F, Wang Y-R, Chang E-C et al (2008) A conserved hydrogen-bond network in the catalytic centre of animal glutaminyl cyclases is critical for catalysis. Biochem J 411(1):181–190
12. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25–29
13. Huntley RP, Sawford T, Mutowo-Meullenet P et al (2015) The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res 43(D1):D1057–D1063
14. Consortium R (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Res 49(D1):D212–D220
15. Meldal BH, Pons C, Perfetto L et al (2021) Analysing the yeast complexome – the complex portal rising to the challenge. Nucleic Acids Res 49(6):3156–3167
16. Paysan-Lafosse T, Blum M, Chuguransky S et al (2023) InterPro in 2022. Nucleic Acids Res 51(D1):D418–D427
17. Gillespie M, Jassal B, Stephan R et al (2022) The reactome pathway knowledgebase 2022. Nucleic Acids Res 50(D1):D687–D692
18. Kanehisa M, Furumichi M, Sato Y et al (2021) KEGG: integrating viruses and cellular organisms. Nucleic Acids Res 49(D1): D545–D551
19. Chang A, Jeske L, Ulbrich S et al (2021) BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res 49(D1):D498–D508
20. Zhou N, Jiang Y, Bergquist TR et al (2019) The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol 20(1): 1–23
21. Smaili FZ, Tian S, Roy A et al (2021) QAUST: protein function prediction using structure similarity, protein interaction, and functional motifs. Genomics Proteomics Bioinformatics 19(6):998–1011
22. Zhang C, Freddolino PL, Zhang Y (2017) COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. Nucleic Acids Res 45(W1):W291–W299
23. Martin DM, Berriman M, Barton GJ (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 5(1): 1–17
24. Conesa A, Götz S, García-Gómez JM et al (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18): 3674–3676
25. Gong Q, Ning W, Tian W (2016) GoFDR: a sequence alignment based method for predicting protein functions. Methods 93:3–14
26. Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33(suppl_2):W89–W93
27. Skolnick J, Brylinski M (2009) FINDSITE: a combined evolution/structure-based approach to protein function prediction. Brief Bioinform 10(4):378–391
28. Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for

structure-based protein function annotation. Nucleic Acids Res 40(W1):W471–W477

29. Letovsky S, Kasif S (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19(suppl_1):i197–i204
30. Vazquez A, Flammini A, Maritan A et al (2003) Global protein function prediction from protein-protein interaction networks. Nat Biotechnol 21(6):697–700
31. Chua HN, Sung W-K, Wong L (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics 22(13):1623–1630
32. Forslund K, Sonnhammer EL (2008) Predicting protein function from domain content. Bioinformatics 24(15):1681–1687
33. Fang H, Gough J (2013) A domain-centric solution to functional genomics via dcGO predictor. BMC Bioinformatics 14(1):1–11
34. Das S, Lee D, Sillitoe I et al (2015) Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. Bioinformatics 31(21): 3460–3467
35. Lan L, Djuric N, Guo Y et al (2013) MS-kNN: protein function prediction by integrating multiple data sources. BMC Bioinformatics 14(3):1–10
36. Piovesan D, Giollo M, Leonardi E et al (2015) INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. Nucleic Acids Res 43(W1):W134–W140
37. Zhang C, Zheng W, Freddolino PL et al (2018) MetaGO: predicting gene ontology of non-homologous proteins through low-resolution protein structure prediction and protein–protein network mapping. J Mol Biol 430(15):2256–2265
38. Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17): 3389–3402
39. Remmert M, Biegert A, Hauser A et al (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9(2):173–175
40. Jiang Y, Oron TR, Clark WT et al (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol 17(1):1–19
41. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33(7): 2302–2309
42. Holm L, Laakso LM (2016) Dali server update. Nucleic Acids Res 44(W1):W351–W355
43. Jumper J, Evans R, Pritzel A et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596(7873):583–589
44. Yang J, Yan R, Roy A et al (2015) The I-TASSER suite: protein structure and function prediction. Nat Methods 12(1):7–8
45. Szklarczyk D, Gable AL, Nastou KC et al (2021) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 49(D1):D605–D612
46. Zhang QC, Petrey D, Garzón JI et al (2012) PrePPI: a structure-informed database of protein–protein interactions. Nucleic Acids Res 41(D1):D828–D833
47. Eddy SR (2004) What is a hidden Markov model? Nat Biotechnol 22(10):1315–1316
48. Bonello J (2023) Protein function prediction methods exploiting the CATH protein domain classification. UCL (University College London), London
49. Liu J, Tang X, Guan X (2023) Grain protein function prediction based on self-attention mechanism and bidirectional LSTM. Brief Bioinform 24(1):bbac493
50. Hearst MA, Dumais ST, Osuna E et al (1998) Support vector machines. IEEE Intell Syst 13(4):18–28
51. Zyphur MJ, Oswald FL (2015) Bayesian estimation and inference: a user's guide. J Manag 41(2):390–420
52. Eisner R, Poulin B, Szafron D et al (2005) Improving protein function prediction using the hierarchical structure of the gene ontology. In: IEEE symposium on computational intelligence in bioinformatics and computational biology. IEEE, pp 1–10
53. Hyunju L, Zhidong T, Minghua D (2006) Diffusion kernel-based logistic regression models for protein function prediction. OMICS J 10(1):40–55
54. Lobley AE, Nugent T, Orengo CA et al (2008) FFPred: an integrated feature-based function prediction server for vertebrate proteomes. Nucleic Acids Res 36(suppl_2): W297–W302
55. Saraç ÖS, Atalay V, Cetin-Atalay R (2010) GOPred: GO molecular function prediction by combined classifiers. PLoS One 5(8): e12382

56. Cheol Jeong J, Lin X, Chen XW (2011) On position-specific scoring matrix for protein function prediction. IEEE/ACM Trans Comput Biol Bioinform 8(2):308–315
57. Yu G, Rangwala H, Domeniconi C et al (2013) Protein function prediction using multilabel ensemble classification. IEEE/ACM Trans Comput Biol Bioinform 10(4):1045–1057
58. Cerri R, Barros RC et al (2016) Reduction strategies for hierarchical multi-label classification in protein function prediction. BMC Bioinformatics 17(1):373
59. You R, Zhang Z, Xiong Y et al (2018) GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. Bioinformatics 34(14):2465–2473
60. Makrodimitris S, Reinders MJ, Van Ham RC (2020) Metric learning on expression data for gene function prediction. Bioinformatics 36(4):1182–1190
61. Kulmanov M, Khan MA, Hoehndorf R (2018) DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics 34(4):660–668
62. Li Z, Liu F, Yang W et al (2021) A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans Neural Netw Learn Syst 33(12):6999–7019
63. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681
64. Zhang F, Song H, Zeng M et al (2020) A deep learning framework for gene ontology annotations with sequence-and network-based information. IEEE/ACM Trans Comput Biol Bioinform 18(6):2208–2217
65. Cao Y, Shen Y (2021) TALE: transformer-based protein function annotation with joint sequence–label embedding. Bioinformatics 37(18):2825–2833
66. Li M, Shi W, Zhang F et al (2022) A deep learning framework for predicting protein functions with co-occurrence of GO terms. IEEE/ACM Trans Comput Biol Bioinform 20(2):833–842
67. Kulmanov M, Hoehndorf R (2022) DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms. Bioinformatics 38(Supplement_1):i238–i245
68. Giri SJ, Dutta P, Halani P et al (2020) MultiPredGO: deep multi-modal protein function prediction by amalgamating protein structure, sequence, and interaction information. IEEE J Biomed Health Inform 25(5):1832–1838
69. You R, Yao S, Mamitsuka H et al (2021) DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. Bioinformatics 37(Supplement_1):i262–i271
70. Gligorijević V, Barot M, Bonneau R (2018) deepNF: deep network fusion for protein function prediction. Bioinformatics 34(22):3873–3881
71. Li W, Zhang H, Li M et al (2022) MGEGFP: a multi-view graph embedding method for gene function prediction based on adaptive estimation with GCN. Brief Bioinform 23(5):bbac333
72. Ibtehaz N, Kagaya Y, Kihara D (2023) Domain-PFP allows protein function prediction using function-aware domain embedding representations. Commun Biol 6(1):1103
73. Wu Z, Guo M, Jin X et al (2023) CFAGO: cross-fusion of network and attributes based on attention mechanism for protein function prediction. Bioinformatics 39(3):btad123
74. Wu J, Qing H, Ouyang J et al (2023) HiFun: homology independent protein function prediction by a novel protein-language self-attention model. Brief Bioinform 24(5):bbad311
75. Xia W, Zheng L, Fang J et al (2022) PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. Comput Biol Med 145:105465
76. Zheng L, Shi S, Lu M et al (2024) AnnoPRO: a strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding. Genome Biol 25(1):41
77. Heinzinger M, Elnaggar A, Wang Y et al (2019) Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 20:1–17
78. Lee DD, Pham P, Largman Y et al (2009) Advances in neural information processing systems 22. Tech Rep 1(1):1–11
79. Rives A, Meier J, Sercu T et al (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 118(15):e2016239118
80. Elnaggar A, Heinzinger M, Dallago C et al (2021) Prottrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 44(10):7112–7127
81. Lin Z, Akin H, Rao R et al (2023) Evolutionary-scale prediction of atomic-level

protein structure with a language model. Science 379(6637):1123–1130

82. Su J, Han C, Zhou Y et al (2023) SaProt: protein language modeling with structure-aware vocabulary. bioRxiv:2023.2010.2001.560349
83. Elnaggar A, Essam H, Salah-Eldin W et al (2023) Ankh: optimized protein language model unlocks general-purpose modelling. arXiv preprint:arXiv:2301.06568
84. Yang KK, Fusi N, Lu AX (2024) Convolutions are competitive with transformers for protein sequence pretraining. Cell Syst 15(3):286–294
85. Zhu Y-H, Liu Z, Liu Y et al (2024) ULDNA: integrating unsupervised multi-source language models with LSTM-attention network for high-accuracy protein–DNA binding site prediction. Brief Bioinform 25(2):bbae040
86. Kulmanov M, Guzmán-Vega FJ, Duek Roggli P et al (2024) Protein function prediction as approximate semantic entailment. Nat Mach Intell 6:1–9
87. Littmann M, Heinzinger M, Dallago C et al (2021) Embeddings from deep learning transfer GO annotations beyond homology. Sci Rep 11(1):1160
88. Zhao C, Liu T, Wang Z (2022) PANDA2: protein function prediction using graph neural networks. NAR Genom Bioinform 4(1): lqac004
89. Zhu Y-H, Zhang C, Yu D-J et al (2022) Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction. PLoS Comput Biol 18(12): e1010793
90. Lai B, Xu J (2022) Accurate protein function prediction via graph attention networks with predicted structure information. Brief Bioinform 23(1):bbab502
91. Zhang L, Jiang Y, Yang Y (2023) Gnngo3d: protein function prediction based on 3d structure and functional hierarchy learning. IEEE Trans Knowl Data Eng 36(8): 3867–3878
92. Jiao P, Wang B, Wang X et al (2023) Struct2GO: protein function prediction based on graph pooling algorithm and AlphaFold2 structure information. Bioinformatics 39(10):btad637
93. Wang Z, Deng Z, Zhang W et al (2023) MMSMAPlus: a multi-view multi-scale multi-attention embedding model for protein function prediction. Brief Bioinform 24(4): bbad201
94. Yuan Q, Xie J, Xie J et al (2023) Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. Brief Bioinform 24(3):bbad117
95. Zhang X, Guo H, Zhang F et al (2023) HNetGO: protein function prediction via heterogeneous network transformer. Brief Bioinform 24(6):bbab556
96. Gu Z, Luo X, Chen J et al (2023) Hierarchical graph transformer with contrastive learning for protein function prediction. Bioinformatics 39(7):btad410
97. Boadu F, Cao H, Cheng J (2023) Combining protein sequences and structures with transformers and equivariant graph neural networks to predict protein function. Bioinformatics 39(Supplement_1):i318–i325
98. Oliveira GB, Pedrini H, Dias Z (2023) TEMPROT: protein function annotation using transformers embeddings and homology search. BMC Bioinformatics 24(1):242
99. Zheng R, Huang Z, Deng L (2023) Large-scale predicting protein functions through heterogeneous feature fusion. Brief Bioinform 24(4):bbad243
100. Hoehndorf R, Kulmanov M, Tawfiq R et al (2024) DeepGOMeta: predicting functions for microbes. bioRxiv:2024.2001.2028.577602
101. Wang S, You R, Liu Y et al (2023) NetGO 3.0: protein language model improves large-scale functional annotations. Genomics Proteomics Bioinformatics 21(2):349–358
102. Yao S, You R, Wang S et al (2021) NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. Nucleic Acids Res 49(W1):W469–W475
103. You R, Yao S, Xiong Y et al (2019) NetGO: improving large-scale protein function prediction with massive network information. Nucleic Acids Res 47(W1):W379–W387
104. You R, Huang X, Zhu S (2018) DeepText2GO: improving large-scale protein function prediction with deep semantic text representation. Methods 145:82–90
105. Kulmanov M, Hoehndorf R (2020) DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 36(2): 422–429
106. Clark WT, Radivojac P (2013) Information-theoretic evaluation of predicted ontological annotations. Bioinformatics 29(13):i53–i61
107. Radivojac P, Clark WT, Oron TR et al (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10(3):221–227

108. Xie S, Xie X, Zhao X et al (2023) HNSPPI: a hybrid computational model combing network and sequence information for predicting protein–protein interaction. Brief Bioinform 24(5):bbad261
109. Shvedunova M, Akhtar A (2022) Modulation of cellular processes by histone and non-histone protein acetylation. Nat Rev Mol Cell Biol 23(5):329–349
110. Jovanovic M, Rooney MS, Mertins P et al (2015) Dynamic profiling of the protein life cycle in response to pathogens. Science 347(6226):1259038
111. Chatham JC, Zhang J, Wende AR (2021) Role of O-linked N-acetylglucosamine protein modification in cellular (patho) physiology. Physiol Rev 101(2):427–493
112. Noble ME, Endicott JA, Johnson LN (2004) Protein kinase inhibitors: insights into drug design from structure. Science 303(5665): 1800–1805
113. Śledź P, Caflisch A (2018) Protein structure-based drug design: from docking to molecular dynamics. Curr Opin Struct Biol 48:93–102
114. Luo S, Guan J, Ma J et al (2021) A 3D generative model for structure-based drug design. Adv Neural Inf Proces Syst 34:6229–6239
115. Gurung AB, Ali MA, Lee J et al (2021) An updated review of computer-aided drug design and its application to COVID-19. Biomed Res Int 2021(1):8853056
116. Huang P-S, Boyken SE, Baker D (2016) The coming of age of de novo protein design. Nature 537(7620):320–327
117. Anishchenko I, Pellock SJ, Chidyausiku TM et al (2021) De novo protein design by deep network hallucination. Nature 600(7889): 547–552
118. Pan X, Kortemme T (2021) Recent advances in de novo protein design: principles, methods, and applications. J Biol Chem 296: 100558
119. Hieter P, Boguski M (1997) Functional genomics: it's all how you read it. Science 278(5338):601–602
120. Przybyla L, Gilbert LA (2022) A new era in functional genomics screens. Nat Rev Genet 23(2):89–103
121. Pietzner M, Wheeler E, Carrasco-Zanini J et al (2021) Mapping the proteo-genomic convergence of human diseases. Science 374(6569):eabj1541
122. Harms MJ, Thornton JW (2013) Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nat Rev Genet 14(8):559–571
123. Bepler T, Berger B (2021) Learning the protein language: evolution, structure, and function. Cell Systems 12(6):654–669
124. Waters ER, Vierling E (2020) Plant small heat shock proteins–evolutionary and functional diversity. New Phytol 227(1):24–37
125. Khorkova O, Stahl J, Joji A et al (2023) Amplifying gene expression with RNA-targeted therapeutics. Nat Rev Drug Discov 22(7):539–561
126. Jang YJ, Qin Q-Q, Huang S-Y et al (2024) Accurate prediction of protein function using statistics-informed graph networks. Nat Commun 15(1):6601
127. Chen Z, Luo Q (2024) DualNetGO: a dual network model for protein function prediction via effective feature selection. Bioinformatics 40(7):btae437
128. Yuan Q, Tian C, Song Y et al (2024) GPSFun: geometry-aware protein sequence function predictions with language models. Nucleic Acids Res 52:W248