Review

Jiaxiao Chen, Zhonghui Gu, Luhua Lai and Jianfeng Pei*

In silico protein function prediction: the rise of machine learning-based approaches

https://doi.org/10.1515/mr-2023-0038
Received August 14, 2023; accepted October 11, 2023;
published online November 29, 2023

Abstract: Proteins function as integral actors in essential life processes, rendering the realm of protein research a fundamental domain that possesses the potential to propel advancements in pharmaceuticals and disease investigation. Within the context of protein research, an imperious demand arises to uncover protein functionalities and untangle intricate mechanistic underpinnings. Due to the exorbitant costs and limited throughput inherent in experimental investigations, computational models offer a promising alternative to accelerate protein function annotation. In recent years, protein pre-training models have exhibited noteworthy advancement across multiple prediction tasks. This advancement highlights a notable prospect for effectively tackling the intricate downstream task associated with protein function prediction. In this review, we elucidate the historical evolution and research paradigms of computational methods for predicting protein function. Subsequently, we summarize the progress in protein and molecule representation as well as feature extraction techniques. Furthermore, we assess the performance of machine learning-based algorithms across various objectives in protein function prediction, thereby offering a comprehensive perspective on the progress within this field.

Keywords: protein function prediction; pre-training models; protein interaction prediction; protein function annotation; biological knowledge graph

Introduction

Proteins, intricate biomolecules and macromolecules, are composed of one or more elongated chains of amino acid residues, synthesized via dehydration condensation. They represent the ultimate outcome of genetic information expression through processes of transcription and translation, playing a pivotal role as carriers and enactors of vital biological activities. Polypeptide chains synthesized within organisms spontaneously adopt zigzag conformations and fold into stable three-dimensional structures. It is widely acknowledged that the functionality of a protein heavily depends on its specific three-dimensional structure. Consequently, the discernment of protein structure and function has emerged as a paramount pursuit within the realm of life sciences [1]. The advancement of structural biology techniques, coupled with breakthroughs in deep learning-driven protein structure prediction methodologies exemplified by AlphaFold2 and RosseTTAFold, has propelled significant leaps in the identification of protein structures [2, 3]. These advancements will propel the annotation of protein functionality and the comprehension of the intricate mechanisms underlying it, which is the ultimate goal of protein research.

The exploration of protein functionality encounters challenges of greater complexity compared to the study of protein structure, as proteins exhibit diverse forms of functionality. Proteins serve not only as individual entities catalyzing chemical reactions but also engage in multifaceted interactions with other proteins or molecules. For instance, transcription factors and RNA-binding proteins exert their influence by binding to nucleotide chains [4]. Entities like the proteasome and inflammasome function as intricate complexes [5]. Furthermore, current understanding acknowledges that proteins frequently participate in elaborate interaction networks, thereby complicating the precise delineation of their functional attributes [6].

***Corresponding author: Jianfeng Pei**, Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China; and Research Unit of Drug Design Method, Chinese Academy of Medical Sciences (2021RU014), Beijing, 100871, China, E-mail: jfpei@pku.edu.cn. https://orcid.org/0000-0002-8482-1185
Jiaxiao Chen, Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China
Zhonghui Gu, Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China
Luhua Lai, Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China; Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China; BNLMS, College of Chemistry and Molecular Engineering, Peking University, Beijing, China; and Research Unit of Drug Design Method, Chinese Academy of Medical Sciences (2021RU014), Beijing, China

Traditionally, ascertaining protein function necessitates a sequence of molecular biological experiments including gene knockout, protein–protein interaction (PPI) experiments, drug–protein interaction investigations, and other methodologies [7, 8]. The experimental findings coupled with manual annotation has long served as the gold standard for ascertaining protein functionality. However, with the development of omics techniques and progress in protein structure research, the number of discovered protein sequences and structures is increasing exponentially every year [9–11]. Given the substantial time and resource costs, the current annotation of protein functions by experimental is unable to keep pace with the rate of natural protein discovery.

Hence, the utilization of computational methodologies for predicting protein function has evolved over a span exceeding two decades. Within this domain, the scope of protein function prediction encompasses two overarching research objectives (Figure 1). The first objective revolves around the anticipation of proteins possessing specific attributes or engaging with interacting partners. Notably, investigations have been directed towards predicting PPIs DNA-binding proteins, and RNA-binding proteins. A noteworthy instance is the critical assessment of prediction of interactions (CAPRI), a community-wide competition centering on PPI prediction [12]. The second objective pertains to the prediction of protein function annotations. Illustratively, the gene ontology (GO) database contains an extensive repository of functional annotations for proteins, prompting certain studies towards predicting GO annotations [13]. Focus on this goal, the critical assessment of functional annotation (CAFA) is an ongoing and global competition to improve the computational annotation of protein function [14]. In the early stage, most of the methods were based on interpretable physical and chemical properties or analysis of interprotein relationships. These empirical and manual feature extraction-based approaches are susceptible to bottlenecks, due to some underlying assumptions on which the algorithm relies that are not always hold true. By contrast, data-driven methods do not rely on empirical knowledge but are mainly affected by data quantity and noise levels. In recent years, the accumulation of data and the development of machine learning have provided significant impetus and opportunities in this domain. The accumulation of data and the advancement of machine learning have furnished significant impetus and opportunities within this field. The

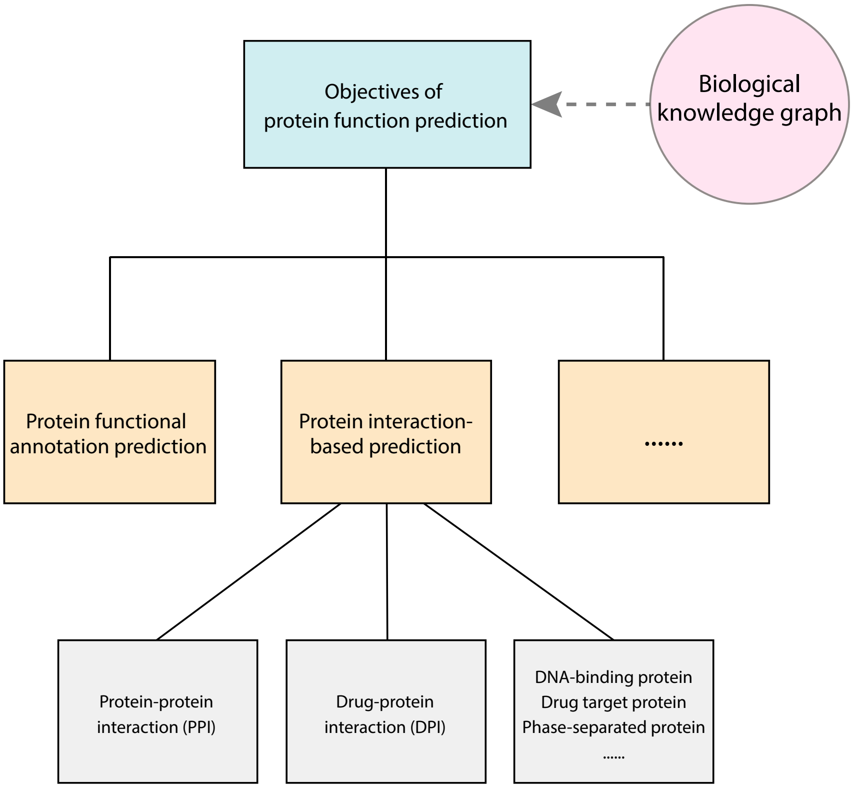
Objectives of protein function prediction
Biological knowledge graph
Protein functional annotation prediction
Protein interaction-based prediction
......
Protein-protein interaction (PPI)
Drug-protein interaction (DPI)
DNA-binding protein
Drug target protein
Phase-separated protein
......

Figure 1: Multiple objectives of protein function prediction.

substantial discovery of sequences and structures has yielded abundant samples for machine learning. Concurrently, the development of transformer models and graph neural networks (GNN) on the foundation of extensive samples holds the potential to provide enhanced structural and sequential feature inputs for function prediction. Consequently, machine learning-based data-driven approaches have gained prominence [15].

We have emphasized the significance of protein function prediction within the scope of life sciences and highlighted the potential inherent in computational methods for predicting protein function. This field has witnessed significant advancements over the years, with researchers continuously striving to improve accuracy and efficiency. In the subsequent sections, we will delve into the progression of research paradigms in protein function prediction. We will trace the evolution from traditional methodologies to more advanced machine learning-based approaches that have revolutionized this field. These modern techniques leverage vast amounts of data and powerful algorithms to extract meaningful insights from complex biological systems. To provide a comprehensive understanding, we will review classical research workflows involved in predicting protein function. This includes exploring different strategies for representing proteins, extracting relevant features that capture their functional characteristics, selecting appropriate frameworks or models for analysis, and training these models using suitable datasets.

Notably, recent years have witnessed remarkable breakthroughs in large-scale pre-training models in natural language processing (NLP). These models have demonstrated exceptional capabilities in understanding and generating human-like text by learning from massive amounts of textual data. The application of such pre-training techniques holds great promise for advancing our understanding of proteins as well. By incorporating concepts from NLP into protein function prediction research, scientists are exploring new avenues to enhance predictions based on similarities between language structures and protein sequences or structures. This interdisciplinary approach opens up exciting possibilities for improving accuracy and expanding our knowledge about how proteins perform their vital functions within living organisms. Consequently, pre-training models have progressively assumed a pivotal role across various domains [16]. These developments have revolutionized computational protein research by providing innovative solutions for representing proteins and molecules. One of the key contributions of NLP frameworks is their ability to extract meaningful information from vast amounts of unstructured text data related to proteins and molecules. By leveraging techniques such as named entity recognition, relation extraction, and semantic parsing, these frameworks enable researchers to automatically annotate and categorize protein-related information. This not only saves significant time and effort but also enhances the accuracy and comprehensiveness of protein representation. Moreover, pre-training models play a crucial role in capturing intricate patterns within protein sequences or molecular structures. Through unsupervised learning on large-scale datasets, these models learn rich representations that encode both local structural features and global contextual information. As a result, they can effectively capture the complex relationships between amino acids or atoms in proteins or molecules. The combination of NLP frameworks with pre-training models has opened up new avenues for exploring diverse research prospects in computational biology. For instance, researchers can now leverage these methodologies for tasks such as protein structure prediction, drug discovery, functional annotation of genes/proteins, and analysis of genetic variations associated with diseases. Our focus will be directed towards scrutinizing the influence of NLP frameworks and pre-training models on computational protein research, with a particular emphasis on innovations concerning the representation of proteins and molecules. Finally, we will undertake a comparative analysis of the research prospects presented by machine learning-based methodologies across various tasks.

Paradigms for in silico protein function prediction

The realm of protein function prediction has developed concomitantly with progress in omics technology, structural biology research, and machine learning theory. Thus, in this progression, the paradigm of protein function prediction remains dynamic and adaptable.

In the early stage of protein function prediction, traditional algorithms were employed to predict protein function based on sequence information. In this stage, “inheritance through homology” serves as the primary foundation [17]. For example, PSI-BLAST is a classical method that can be used as a fast and sensitive tool for protein sequence alignment, which can extract the functional signals with certain noise from protein sequences [18]. There are also algorithms to classify protein families by analyzing the differences and similarities in protein sequences. For instance, BLAST is a sequence homology search algorithm that has been widely used since its emergence. And the combination of Markov clustering and pairwise similarity relationship algorithms with BLAST enables rapid and accurate detection of protein families [19, 20]. Furthermore, owing to the evident coevolutionary pattern observed between interacting proteins, several studies have employed the computation of distance matrices derived from phylogenetic trees of two protein families to extract coevolution information for accurate protein function prediction [21–23].

With the development of structural biology, numerous structure-based approaches have been developed. Protein structures exhibit greater conservation than sequences, thus studies based on structural similarity yield more precise results. Since 2000, various attempts have been made to predict enzyme classification (EC) based on structural similarities [24]. Furthermore, the Protein Structure Classification Database (SCOP) has been utilized in studies aiming to predict cytokine families or subfamilies [25]. Additionally, there were also studies to predict the interaction between proteins based on the similarity of protein surface [26]. It is worth noting that at this stage, in addition to inferring functional similarity based on structural similarity, the increase in the number of protein structures also greatly promotes the study of PPIs. For example, employing a fast fourier transform (FFT) algorithm for the spatial conformation matching in protein–protein docking has demonstrated exceptional performance, which often ranks among the top contenders in protein–protein docking competition (CAPRI) [27]. Furthermore, molecular dynamics simulation has been utilized for studying PPIs. However, the parameters used in these methods are mostly based on experience rather than first principles. Therefore, they are limited by computational resources and the lack of a detailed understanding of mechanisms.

With the development of machine learning, a novel research paradigm has emerged, shifting its reliance from understanding or assuming protein interaction mechanisms to data-driven approaches and feature extraction. In the early years, this research paradigm actually had to discard some of its interpretability, even if it achieved good performance for some tasks. In recent years, due to the availability of large-scale protein data, advancements in deep learning frameworks, and improved computational hardware support, the research paradigm based on machine learning is also changing gradually. Along with the improvement of accuracy, it also facilitates comprehension of fundamental principles underlying protein function. In the following sections, we provide an overview of protein function prediction methods based on machine learning.

Protein representation and feature extraction

In general, the research flow in machine learning can be divided into two steps. The first step involves encoding the data as input, followed by the subsequent training of the model through diverse algorithms or frameworks to facilitate forthcoming prediction tasks (Figure 2). Notably, a primary challenge encountered when applying machine learning to protein investigations pertains to the digital representation of proteins and their effective utilization as inputs within machine learning models. Although the function of a protein is inherently dictated by its sequence and structure, the encoding of these attributes alongside the extraction of other key features has consistently constituted a significant theme in this field. Despite the absence of a universal approach that comprehensively addresses this problem, researchers persistently refine protein representation and feature extraction methodologies tailored to varying data types and downstream objectives.

Traditional protein representation methods

During the preliminary phase of machine learning-driven protein function prediction, the prevailing constraints encompassing algorithmic limitations, data volume restrictions, and computational hardware barriers frequently necessitated a manual one-step feature extraction process. Instead of solely relying on direct sequence similarity or clustering methodologies, certain algorithms embraced the integration of proteins' chemical properties as inputs for machine learning models. To exemplify, select studies incorporated amino acid composition, hydrophobicity, solvent-accessible surface area, and polarizability as input features. Then these inputs were combined through a support vector machine (SVM) classifier to solve the binary classification problem of DNA-binding protein or RNA-binding protein [28–30]. Correspondingly, akin methodologies have found application within studies aiming at enzyme family classification [24].

The recognition has gradually emerged that manual feature summarization alone is insufficient to comprehensively address the intricate challenges inherent in protein function prediction. Acknowledging the potential of machine learning in facilitating feature extraction, the incorporation of comprehensive sequence information of proteins assumes significance. Within this context, the most straightforward

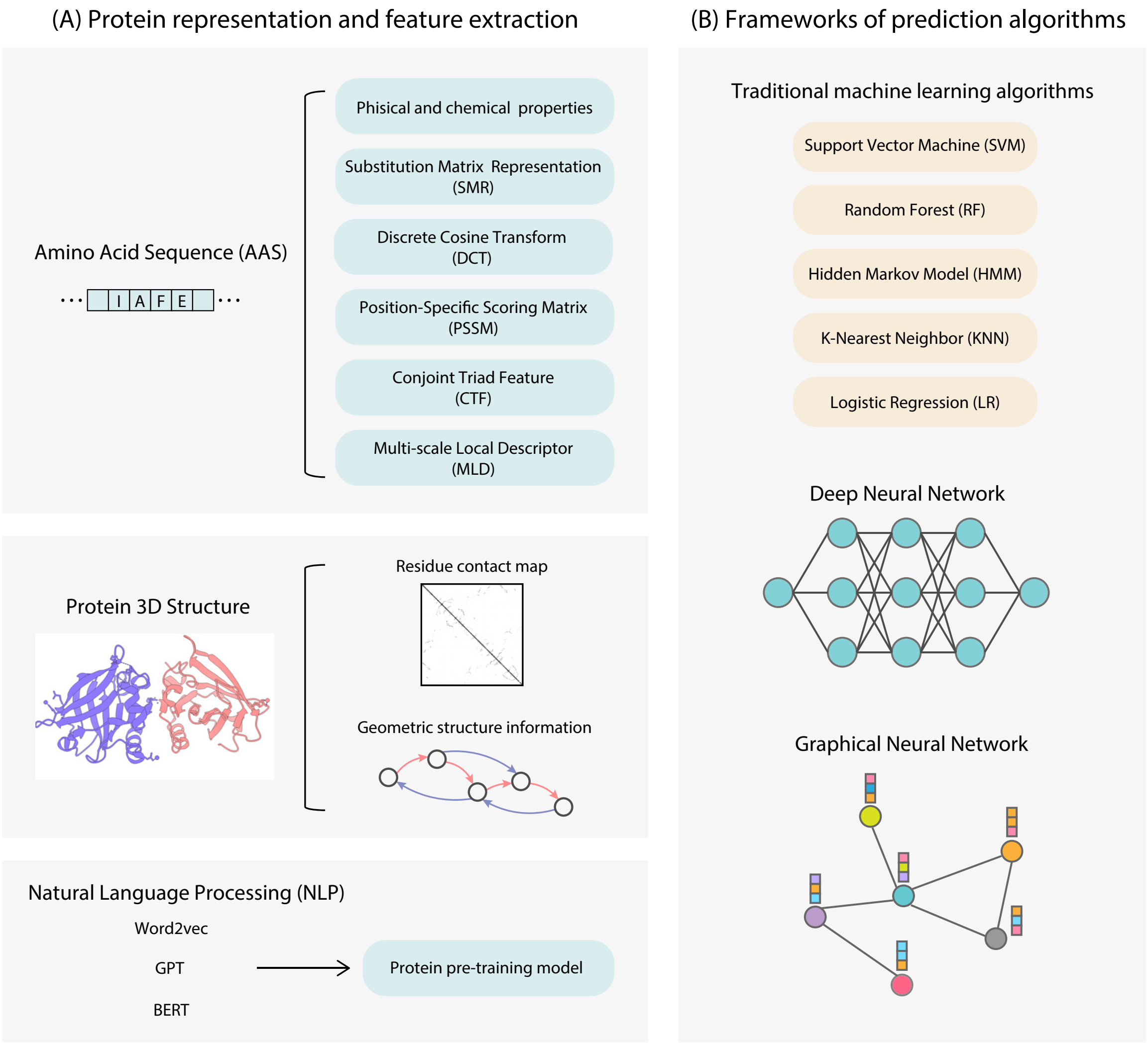
(A) Protein representation and feature extraction
Amino Acid Sequence (AAS)
··· I A F E ···
Phisical and chemical properties
Substitution Matrix Representation (SMR)
Discrete Cosine Transform (DCT)
Position-Specific Scoring Matrix (PSSM)
Conjoint Triad Feature (CTF)
Multi-scale Local Descriptor (MLD)
Protein 3D Structure
Residue contact map
Geometric structure information
Natural Language Processing (NLP)
Word2vec
GPT
BERT
Protein pre-training model
(B) Frameworks of prediction algorithms
Traditional machine learning algorithms
Support Vector Machine (SVM)
Random Forest (RF)
Hidden Markov Model (HMM)
K-Nearest Neighbor (KNN)
Logistic Regression (LR)
Deep Neural Network
Graphical Neural Network

Figure 2: The research paradigm of protein function prediction. (A) Multiple approaches of protein representation and feature extraction. There are several protein-representation approaches for sequence and structure data. Protein encoders based on pre-trained models are also developing rapidly. (B) Machine learning frameworks for subsequent prediction. GPT, generative pre-trained transformer; BERT, bidirectional encoder representations from transformers.

approach for encoding amino acid sequences (AAS) involves the sequential arrangement of amino acids, accompanied by the specification of amino acid types at respective positions. Previous studies have demonstrated that combining AAS with certain physical or chemical descriptors can yield informative protein representations. Research groups have developed several servers for computing these descriptors, such as PROFEAT, which calculates 6 feature groups composed of 10 features, including 51 descriptors and 1,447 values [31]. The features calculated by these servers include amino acid composition, dipeptide composition, normalized Moreau–Broto autocorrelation, Moran autocorrelation, Geary autocorrelation, number of sequence-order coupling, quasi-sequence descriptors, and distributions of various structural and chemical properties. This method for protein encoding and feature extraction has been widely used in a variety of downstream tasks related to protein function, such as predicting drug–protein interactions (DPI), anti-hypertensive peptides, and RNA–protein interactions [32–34].

In order to enhance the feature extraction of protein sequences, various protein encoding methods have been proposed. To better facilitate amino acid alignment and incorporate evolutionary information, substitution matrix representation (SMR) was developed [35]. It calculates the probability that amino acid at each position mutates into another type of amino acid and represents any given protein sequence with length N as an $N \times 20$ substitution matrix, where the sequential similarity depends on the divergence time and substitution rate in the matrix. This approach is often applied to the prediction of interactions between proteins and biomolecules. For example, some studies added discrete cosine transform (DCT) on the basis of SMR for protein interaction prediction in various species, and the average accuracy is up to 96.28, 96.30, and 86.74 % for different species, respectively, which is significantly better than previous methods [36]. Meanwhile, this approach has also been applied to predict drug–protein interaction. For example, a study from Huang et al. encoded protein sequences with SMR descriptors, which achieved more than 80.00 % accuracy on multiple benchmark datasets for the prediction of drug–protein interaction [37].

In order to emphasize the specificity of different positions on the protein sequence, the position-specific scoring matrix (PSSM) method was proposed. This method utilizes PSI-BLAST (Position-specific Iterative BLAST) to calculate the percentage of different residues at each position [38], employing sequence alignment extract evolutionarily relevant feature information. It was applied in the PPlevo algorithm for predicting PPIs [39]. Furthermore, PSSM encoding method has been combined with various classifiers to predict protein function classification in yeast [40]. PSSM can also be integrated with other methods such as orthogonal local preservation projection (OLPP) to encode protein as a feature vector of fixed length and then combined with a RoF classifier to identify non-interacting and interacting protein pairs, with an accuracy of more than 90.00 % in yeast [41]. Autocovariance based on PSSM is another effective sequence-based protein representation method. This method extracts features from PSSMs by considering proximity effects, enabling the highlighting of some specific patterns in the whole sequence, which is also widely employed in protein classification tasks [42–46]. Following the principles in PSSM, SPRINT (Protein interaction Score) and PIPE (Protein Interaction Prediction Engine) determine the interaction between protein pairs by searching for similar pairwise regions among known protein complexes [47, 48].

The Conjoint Triad Feature (CTF) encapsulates not only the attributes of the target amino acid but also those of its neighboring amino acids. By treating any three consecutive amino acids as an entity, it extracts the intrinsic characteristics of a protein. Consequently, it possesses the capability to encompass both the protein sequence's compositional information and the interconnected relationships among adjacent amino acids. The application of CTF extends across various domains, encompassing the prediction of PPI, RNA–protein interactions, and enzyme function [49, 50]. For example, Dey et al. used CTF protein representation methods combined with supervised machine learning methods (SVM, KNN, NB) to predict the interaction between the DENV virus and human proteins, as well as further predict the GO and KEGG pathway [51]. Wang et al. combined CTF and chaos game representation (CGR) with a random forest model to predict RNA–protein interactions [52]. Another study developed an SVM-based method to predict Enzyme Commission (EC), which used CTF to represent a given protein sequence [53].

Another notable approach is the multi-scale local descriptor (MLD), which partitions the protein sequence into segments of varying lengths to capture multi-scale local insights. These methodologies have showcased pronounced efficacy in the encoding of protein sequences and have found widespread application across diverse domains connected to protein function prediction [54, 55]. Despite their divergence in processing techniques for protein sequences, these methodologies collectively share a common hallmark: the integration of essential empirical features and the application of artificial feature extraction processes based on AAS. While these methods demonstrate superior performance compared to the mere encoding of protein sequences and amino acid types, the integration of their respective strengths becomes intricate when confronted with the intricate downstream task of protein function prediction.

NLP-based protein representation methods

The original object of NLP is human language, which shares analogous data structures with protein sequences [56]. Both utilize discrete units to construct structures endowed with specific attributes, ultimately express specific semantics or functions from this specific coding method. Experimental and computational biology have provided a large amount of protein-related data. In recent years, drawing inspiration from the paradigms of NLP, pre-training models tailored for protein encoding have emerged. In 2015, Asgari et al. introduced word2vec to the realm of biomolecules, pioneering the protein representation method provec [57]. This representation method focused on the first-order and second-order information in the protein sequences, generated vectors in protein space, and extracted corresponding protein properties in the embedding space. Combined with the SVM classifier, it was used for the classification of protein families.

Over the past two years, several sequence-based protein pre-training models have emerged. For example, Elnaggar et al. used six mature models in the field of NLP, such as Transformer-XL, XLNet, BERT, Albert, Electra, and T5, to train 393 billion amino acids in UniRef [58]. They tried to capture the biophysical features of protein sequences and verified the advantages of these embedded features on downstream tasks such as protein secondary structure prediction and protein subcellular localization prediction [59]. Brandes et al. presented ProteinBERT, which is a deep language model specifically designed for proteins [60]. The framework used in ProteinBERT is smaller and faster to train, and it achieves near-state-of-the-art performance across multiple benchmarks covering a variety of protein properties including protein structure, post-translational modifications, and biophysical properties. Roshan et al. trained millions of protein sequences using a self-supervised protein language model [61], which showed excellent generalization capabilities with parametric efficiency far higher than previous protein language models. It is worth mentioning that the basic architecture of ESM-1b is transformer, which is a common model in the field of NLP.

Protein pre-training models based on sequences or MSA have shown great potential, which indicates the rationality of applying the pre-training model in NLP to the field of bioinformatics. Protein structure information is also one of the determinants of protein function, while a large number of structural protein information has not been well utilized in protein pre-training models. In this case, researchers tried to add structural information to the pre-training model to obtain richer protein embedding information. For example, Gligorijević et al. proposed DeepFRI to predict protein function by extracting features from both sequences and structures [62]. DeepFRI utilized the LSTM-LM architecture combined with a large number of available sequences and 3D structural data in the form of contact maps. And the result of protein function prediction based on DeepFRI outperformed sequence-based methods on several tasks.

In addition to introducing 3D information in the form of contact maps, GearNet attempted to encode structure information by directly introducing geometric 3D representation [63]. GearNet leveraged AlphaFold2 predicted protein structure for pre-training through self-supervised contrastive learning and outperformed the previous baselines with its acquired structural embeddings on some metrics in the prediction of EC number and GO. Recently, an energy-based protein pre-training model was proposed and applied to two downstream tasks: Protein structure quality assessment (QA) and PPI assessment [64].

In summary, empirically-driven protein feature extraction methodologies continue to maintain a significant foothold. And to deal with diverse task, a variety of well-crafted designs have emerged, which include sequential adjacency and evolutionary relationships among sequences. On the other hand, in recent years, protein encoding methods based on pre-training models are gradually showing their advantages. When confronted with vast amounts of data and complex features, pre-training models have robust capabilities for feature integration. Efficient protein representation and feature extraction is the core of protein function prediction. Within this framework, we have introduced a variety of protein encoding methods. These encoding methods need to be combined with various classifiers, including traditional machine learning classifiers and deep neural network classifiers, for specific downstream tasks.

Protein interaction prediction

Prediction of protein–protein interaction

The process of protein interaction involves the binding of two or more proteins, which plays a pivotal role in numerous biochemical processes. For example, some signaling molecules transmit extracellular signals into the cell through PPI, which is the basis of many biochemical functions [65]. Another example is that proteins can form complexes through long-term interaction and participate in important biological processes such as transport. Moreover, some transient interactions can add modifications to proteins and regulate their function. Therefore, PPI is the core of cell biochemical reactions, and studies about PPI can enhance the comprehension of the mechanisms behind the disease.

The dataset pertaining to PPIs originates from two primary sources. Firstly, a portion of this data is collected from complexes from the Protein Data Bank (PDB), affording atomic-level insights. Secondly, another segment emerges from PPIs elucidated through high-throughput methodologies, including yeast two-hybrid assays, immuno-precipitation, mass spectrometry-based protein complex identification, and affinity purification. The amalgamation of data from these diverse origins has been curated within the publicly accessible Protein Interaction Database (DIP), encompassing protein interaction records spanning diverse organisms ranging from yeast to humans. This reservoir of data constitutes a valuable resource, furnishing ample material for the application of machine learning techniques in the investigation of PPIs.

Traditional molecular docking algorithms can predict the binding conformations of protein complexes, which are effective approaches to study PPIs [66]. These molecular docking algorithms mainly consist of two steps. The first step involves employing a spatial search algorithm such as FFT algorithm to search the spatial conformation of two proteins bound to each other. The second step is to evaluate the affintity of protein binding through scoring function. These traditional molecular docking algorithms often require the spatial conformational coordinates of the protein as input and the three-dimensional space lattice. The traditional molecular docking algorithm has the advantage of being able to obtain multiple candidates binding conformations for any two protein pairs [67, 68]. These algorithms also exhibit certain limitations. First, the prediction of PPIs using molecular docking algorithms heavily relies on the spatial conformation of proteins, which is hampered by the fact that the number of proteins with experimentally-resolved structures is far less than that of protein sequences. Secondly, the interaction prediction of molecular docking algorithm also depends on the scoring function, which is based on experience, physical and chemical laws [69, 70]. The scoring function itself still has considerable potential for improvement. The inputs of molecular docking algorithms are often the rigid conformations of individual proteins, which may undergo flexible backbone changing during the process of interaction. Therefore, to optimize the docking poses, some molecular docking algorithms have to allocate more computational time by introducing local molecular dynamics simulations or flexible conformational libraries [71, 72]. How to deal with flexible docking remains an unresolved issue in this field. Molecular docking algorithms necessitate performing conformational searches for each input protein pair, and occasionally even require conducting conformational modeling from protein sequences. Consequently, the execution of molecular docking algorithms on a large-scale screening basis could potentially be hampered by computational time constraints.

In recent times, machine learning-based methodologies have in part addressed the limitations inherent in traditional approaches concerning the prediction of PPIs. Noteworthy studies in recent years using machine learning-based methods to predict PPI are listed in Table 1. First, machine learning-based prediction methods can directly treat the target of the task as binary classification, which means that the input protein representation could be more flexible. Beyond characterizing protein sequences and conformations, the integration of effective feature extraction founded on physicochemical priori knowledge can be incorporated into the model. Since 2001, computational methods have been employed in attempts to predict PPIs [73–75]. Until recent years, various protein representation methods have been applied to this objective. For example, Carlos et al. applied six different new features to represent proteins [76]. Sun et al. applied the Autocovariance method with Stacked autoencoder (SAE) to study sequence-based human PPI predictions [77]. Bryant et al. used multiple sequence alignment (MSA) as input [78]. Beyond augmenting flexibility in protein representation and prediction targets, machine learning-grounded PPI prediction algorithms pivot around a data-driven paradigm rather than relying on prior knowledge. Therefore, the use of extensive PPI datasets significantly enhances the precision of machine learning-based PPI prediction. For example, Hanggara et al. obtained a large number of PPI datasets based on string-DB, and the validation accuracy was close to 90 % [79]. Machine learning-based prediction methods have also contributed to the exploration of fundamental principles underlying PPI. Methods have been devised to discern akin targets within protein interaction networks. For instance, Zhou et al. employed a PPIs network between SARS-CoV-2 and human, constructed through high-throughput yeast experiments and mass spectrometry, to unveil 361 new host factors, including proteins devoid of specific experimental structures such as BAG3—an entity implicated in diverse diseases like heart disease and cancer [80]. Kovács et al. delved into the role of BAG3 in bacterial infections through the lens of a PPI network. These network-based prediction methods even have the potential to challenge the conventional wisdom that interacting proteins are not necessarily similar, and that similar proteins do not necessarily interact with each other [81].

AlphaFold2 greatly facilitated protein structure prediction [2, 3], making it feasible to achieve structure information from mere protein sequences. Since protein spatial structure information is difficult to extract directly from the embedding of protein sequences, integrating the predicted protein spatial structure in the PPI prediction process can effectively improve the accuracy in the post-AlphaFold era. For example, TAGPPI relied solely on sequences and performs PPI prediction end-to-end, without additional input of protein 3D structure [87]. TAGPPI used AlphaFold2 in the algorithm to construct the residue contact map of the protein, which contained precise spatial structure information, thus effectively improving the prediction ability of PPI. For the protein complex prediction task, the DeepMind team retrained AlphaFold-multimer for the protein complexes [88]. They linked multiple proteins into single chains with cross-chain positional encodings as input to AlphaFold2. This approach demonstrated a great improvement in heterologous complex structure prediction. Bryant

Table 1: Algorithms for protein–protein interactions (PPI) prediction.

Authors	Protein representation	Framework	Advantage	Dataset	Year	Ref.
Juwen Shen et al.	Kernel function, CTF	SVM	To explore any newly discovered protein network of unknown biological relevance	Human Protein References Database (HPRD)	2006	[82]
Fatma-Elzahraa Eid et al.	Doc2vec	SVM	DBNS method to construct negative datasets	VirusMentha	2016	[83]
Tanlin Sun et al.	Autocovariance	Stacked autoencoder	This model is the first PPI prediction model based on deep learning algorithm.	Pan's PPI dataset from [84]	2017	[77]
Somaye Hashemifar et al.	AAS	Siamese-like convolutional neural network	Superior to the stat-of-the-art methods	Profppikernel	2018	[85]
Carlos H.M. Rodrigues	Physical and chemical properties, PSSM Score	Graphical neural network	The average Pearson Correlation of 0.82 ± 0.06 is better than the previous method	SKEMPI 2.0	2019	[76]
Stván A. Kovács1 et al.	–	L3 (length three) link prediction methods	Significantly better than all the existing link prediction methods	HI-tested, a subset of the human interaction dataset HI-II-145	2019	[81]
Faruq Sandi Hanggara et al.	CTF	Stacked-autoencoder and stacked-randomized autoencoder	The average validation accuracy was 0.89 % ± 0.02 %	STRING-DB	2020	[79]
Patrick Bryant et al.	Several MSAs	–	Use Alphafold2 to predict heterodimeric protein complexes	CASP14 set, 216 novel protein complexes	2021	[78]
Yang Xue et al.	AAS, function tokens embeddings, the vectorized Rips complex barcodes, and Alpha complex barcodes	The single-stream multimodal transformer, Residual CNN	A multimodal protein pretraining model with three modes: Sequence, Structure, and Function	CATH, PDB	2022	[86]

SVM, support vector machine; AAS, amino acid score; CTF, the conjoint triad feature; MSA, multiple sequence alignment.

et al. employed AlphaFold2 to incorporate species-specific multiple sequence alignment (MSA), thereby enhancing the precision of protein complex prediction [78]. AF2Complex enables the structural inference of a polymeric protein complex using a single protein sequence without necessitating retraining of AlphaFold2. In contrast to other approaches, this method integrates MSA regions from diverse proteins by means of sequence alignment. Leveraging these sequence and template features, the AF2Complex model generates a comprehensive complex model by iteratively computing its interface score S for ranking the confidence level [89].

In summary, PPI prediction is a crucial research direction for protein function prediction, with significant implications for comprehending protein interaction network and identifying disease targets. Treating PPI prediction as a classification task or directly predicting protein complex binding conformations are both meaningful prediction targets. With the accumulation of protein data, data-driven machine learning prediction algorithms are playing an increasingly important role in PPI prediction. How to extract features and integrate information more effectively remains a problem.

Prediction of drug–protein interaction

The development of algorithms for predicting small molecule-protein interactions is crucial not only for drug screening, but also for identifying potential drug targets. Additionally, these algorithms can be utilized to predict the interactions between endogenous metabolic molecules and proteins, including sugars, bioactive peptides, endogenous regulatory factors, signaling molecules, etc., thereby shedding light on cellular regulatory mechanisms. Similar to the encoding of macromolecular proteins, representation of small molecular compounds has experienced a paradigm shift from traditional molecular descriptors to machine learning training for embedding.

As shown in Figure 3, we will introduce the following forms of molecular representation: one-dimensional linear inputs (such as SMILES or selfies, inches), structural or path-based fingerprints, and two-dimensional graphical structures (atoms and bonds) that involving topological information.

The characterization of small molecules through string representation is widely employed in scientific research. Molecular structures can be translated into machine-

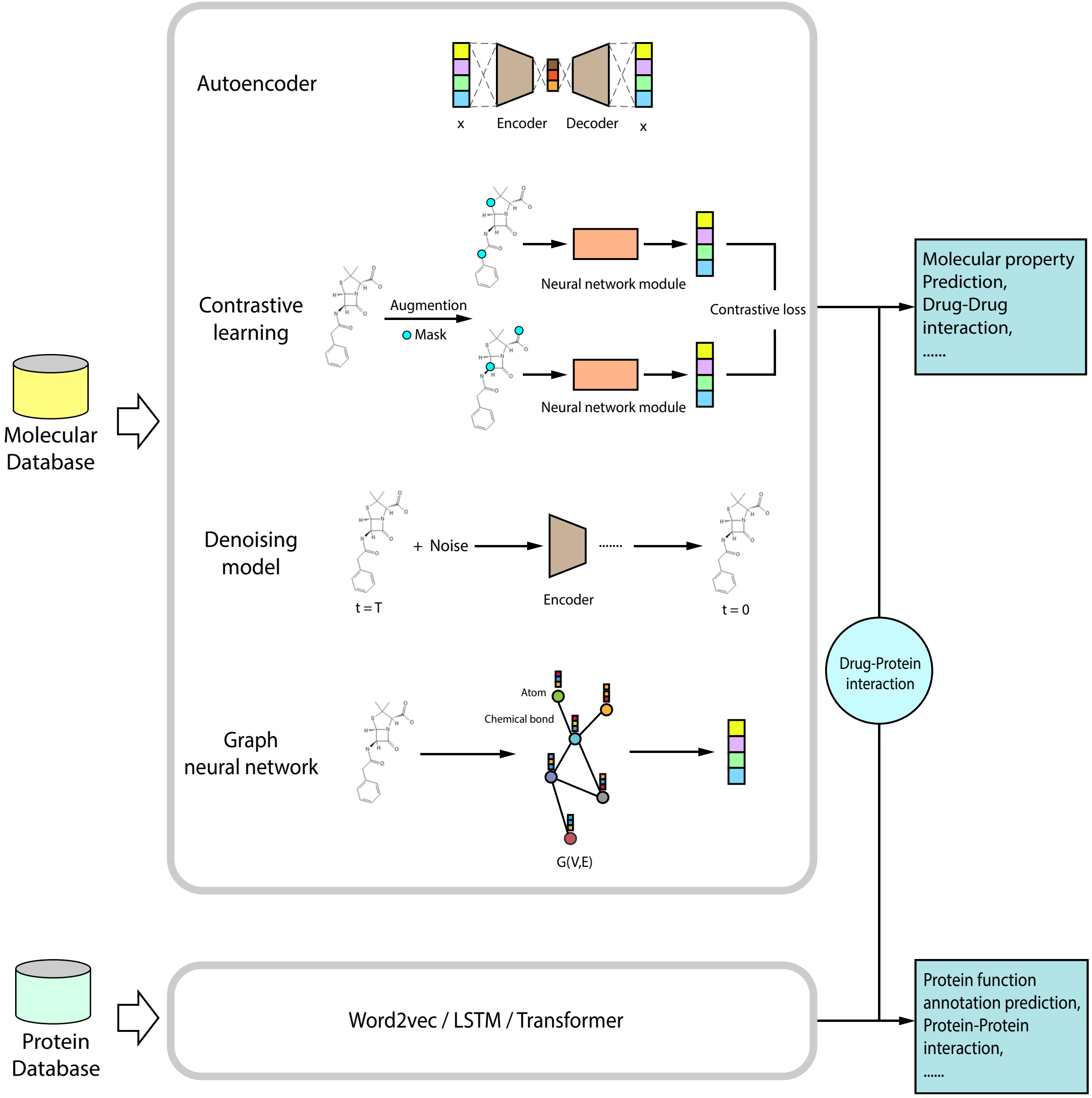

Figure 3: Small molecule and protein representation based on machine learning pre-training models. LSTM, long short-term memory.

readable string representations that are more suitable for NLP, among which, SMILES is typical molecular string representation. Several deep generative models were developed to learning the distribution of SMILES representation [90–92]. It is worth noting that the SMILES string is non-unique, often leading to multiple encoded representations for a single molecule. In this context, certain deep generative models have proposed enhancements to the traditional SMILES format. An illustrative example is SELFIES, which serves as an advanced alternative to SMILES. Particularly in the context of Pangu-based models, SELFIES is preferred over SMILES as input. This preference stems from findings that molecules generated using SELFIES exhibit an efficacy level of up to 100 %.

The internal topological structure of the small molecule naturally allows the molecule to be represented as a two-dimensional graph. The atoms of a molecule are mapped to nodes of a graph containing information such as atomic type, chirality, etc. The edges are linked when there exists a covalent bond between two atoms, and the edge attributes include the types of chemical bonds. Such a graph structure is often represented as inputs of GNN. Combined with deep neural networks such as transformer, the topological structure of molecules can be better extracted [93–96].

Recent molecular characterization methods based on deep learning pre-training models aim to add molecular structure, molecular properties and other information into the training process to generate efficient embedding.

Self-supervised frameworks are frequently employed in small molecule pre-training. Given the substantial data requirements of pre-trained models, leveraging contrastive learning for data augmentation represents an effective strategy. After data augmentation, the consistency between similar inputs is maximized in the feature space and the differences between different classes of data are enlarged. At present, the prevailing approach to enhance small molecules is to randomly mask the atoms, chemical bonds, and subgraphs of molecules. MolCLR, for example, constructed molecular maps using extensive unlabeled data and developed graph neural network encoders to learn molecular properties, which performed impressively in the benchmark test [97]. iMolCLR reduced negative pairs between similar molecules [98]. In addition to directly comparing the degree of similarity between molecules, ATMOL compared the molecular map with masked attention matrices generated by graph attention networks (GAT), which improved the performance on downstream tasks. There are also been some studies trying to combine 3D information of molecules with generative models. GraphMVP, for example, used 2D topologies and 3D geometric views for sub-supervised learning. GraphMVP used accurate 3D molecular conformations from the GEOM dataset to do more discriminating 3D geometric enhancements than the 2D molecular map encoder [99]. The success of denoising in image generation has led to its application in molecular characterization tasks. A recent work utilized denoising-based autoencoders to learn molecular force fields for pre-training, and improved molecular property prediction performance on multiple benchmark datasets [100].

Most small molecular drugs exert their efficacy by interacting with their target proteins, such as enzymes, ion channels, and G-protein-coupled receptors. Therefore, identifying DPI is an important prerequisite for drug discovery, pharmacology, drug side effects, and other studies [101]. Biochemical assays for experimentally undiscovered DPI are costly and time-consuming. In the face of a large number of potential unpaired small molecule compounds and drug target proteins, large-scale virtual screening by computational methods can provide a very valuable reference and guidance for experimental verification.

Similar to the prediction of PPI, there are three main methods to predict DPI, the first of which is based on molecular docking. Conformation search and molecular dynamics simulation are combined to reconstruct the contact relationship between small drug molecules and proteins in 3D space, with the goal to find the best binding pose. The disadvantage of molecular docking-based methods is that they require accurate protein structure as input and are time-consuming [102–104]. The second approach is to predict interactions based on drug–protein association networks [105, 106]. The underlying principle here is that proteins sharing similar structures and exhibiting close relationships are more likely to interact with the same drug. This methodology typically involves the establishment of a network encompassing existing drugs and proteins, followed by the computation of similarity scores for both drug pairs and protein pairs. However, given the absence of a standardized protein similarity score, a drawback of this approach pertains to the accuracy of similarity scoring, particularly when dealing with rare or novel proteins. It is imperative to acknowledge that network-based methodologies hinge upon assumptions that may not universally hold true, as not all similar proteins necessarily interact with similar drugs. The third approach involves a data-driven method rooted in learning, which operates independently of *a priori* assumptions but demands substantial data quantity and quality to be effective [107]. Several databases, such as PubChem, ChEMBL, DrugBank, and DUD-E contain a large amount of information on the interaction of ligand molecules with target proteins. The PDB database also contains a large amount of structural data. The collective information within these databases underpins the utilization of machine learning techniques for the prediction of DPI. In this context, machine learning algorithms have been widely used in the field of computer-aided drug design (CADD) to predict DPI.

We have already reviewed the methods of small molecule characterization commonly used in machine learning. Small molecule compounds can be naturally described in computer-readable formats, such as strings, graphs, etc. The Simplified Molecular Input Line Entry System (SMILES) is the most widely used string format. It is worth noting that both small molecules and proteins are essentially composed of atoms and chemical bonds, and are therefore very easy to represent in the form of a connection graph. Representing atoms as nodes and chemical bonds as edges, GNN are natural small molecule machine learning network frameworks. Multiple variants of graph networks achieved state of the art (SOTA) performance in multiple machine learning domains, such as graph convolutional networks (GCNs), graph attention networks (GATs), and graph isomorphic networks (GINs). These network architectures can be efficiently employed for various downstream tasks concerning small molecules. Furthermore, the utilization of pre-training models specifically designed for small molecules to achieve effective embeddings is an emerging concept. This representation approach, rooted in pre-training, is relatively novel and currently lacks comprehensive evaluation on downstream tasks, particularly in the context of DPI.

Typical studies in recent years using machine learning-based methods to predict DPI are listed in Table 2. The prediction of DPI is quite different from the prediction of PPI. DPI prediction requires not only integrating drug molecule and protein databases but also embedding the chemical space and protein space into a unified space. In this case, the deep neural network framework is gradually showing more advantages over traditional machine learning methods. A straightforward idea for embedding small molecules and proteins into the same hidden space is to use simple concatenation to combine protein and small molecule representations. However, this approach has limitations as simple concatenation does not comprehensively capture the intricate interactions between the two compound types. Recent efforts have aimed at achieving more robust information interactions that effectively capture the nuanced connections between small molecules and proteins. For instance, the Perceiver CPI model integrated a cross-attention module to compel the model to discern the

Table 2: Algorithms for drug–protein interactions (DPI) prediction.

Author	Algorithm	Protein representation	Framework	Advantage	Year	Ref.
Nobuyoshi Nagamine et al.	MDMA	AAS, Chemical Structure, and Mass Spectrometry	SVM	One of the earliest methods to apply machine learning to the study of small protein molecules	2007	[115]
Yoshihiro Yamanishi et al.	–	Chemical structure and genome sequence	A bipartite graph	The 3D structural information of the target protein is not required, and the chemical and genomic Spaces are integrated into a unified space	2008	[116]
Ming Wen et al.	DeepDTIs	ECFP + PSC	Deep belief network	The first to use deep learning	2017	[117]
Hakime Öztürk et al.	DeepDTA	SMILES sequence	1D CNN	Only sequence information was used to predict binding affinity	2018	[118]
Qing Ye et al.	KGE_NFM	DistMult	Neural factorization machine	Pre-training model based on knowledge graph	2021	[119]
Gengmo Zhou et al.	Uni-Mol	SE (3)-equivariant transformer architecture	Additional 4-layer Uni-Mol and a simple differential evolution algorithm to sample and optimize the complex	The first general 3D molecular pre-training framework	2022	[120]
Vineeth R. Chelur et al.	BiRDS	The MSAs features, Token embedding, position Embedding, Segment Embedding	ResNet	BiRDS can accurately predict the most active binding site of a protein using only sequence information	2022	[121]
Ngoc-Quang Nguyen et al.	Perceiver CPI	Molecular: Molecular Graph + ECFP Protein:1D sequence	D-MPNN, MLP, 1D CNN	Cross-attention module	2022	[108]
Jian Wang et al.	Yuel	Employ rdkit to represent SMILES by a graph (N, V, E) Protein sequence	GCN, FC	Predict interactions between unknown compounds and unknown proteins	2022	[109]
Penglei Wan et al.	STAMP-DPI	Molecular: Mol2vec Protein: TAPE	Transformer decoder	More attention is paid to protein structural features	2022	[112]
Qichang Zhao et al.	HyperAttentionDTI	Molecular: SMILES Protein: Protein sequence	CNN, attention mechanism	Focus on complex noncovalent intermolecular interactions between atoms and amino acids	2022	[113]
Yifan Wu et al.	BridgeDPI	Molecular: Morgan fingerprint + physicochemical Protein: one-hot + 1,2,3-mer	CNN, FNN, GNN	Capture network-level information between molecules and proteins	2022	[114]
Mehdi Yazdani-Jahromi et al.	AttentionSiteDTI	Molecular: SMILES Protein: Protein sequence	GAT	Works well on new proteins	2022	[110]

AAS, amino acid score; GCN, graph convolutional network; FNN, feedforward neural network; CNN, convolutional neural network; GAT, graph attention networks; D-MPNN, directed message passing neural network; MLP, multi-layer perception.

impact of compound information on protein information [108]. Other investigations have focused on predicting interactions between previously uncharacterized proteins and unknown small molecules. Yuel, for instance, prominently incorporated a characterized FC layer alongside an attention-based affinity prediction module that employed the outer product to combine protein and small molecule features [109]. Additionally, a noteworthy example emphasizing the prediction of DPI involving new proteins is AttentionSiteDTI, drawing inspiration from sentence classification models [110]. Here, the drug target complex was likened to a sentence with meaningful connections between its biochemical entity (referred to as the protein pocket) and the drug molecule. The authors highlighted that unlike previous studies, this model demonstrated exceptional performance when applied to novel proteins. Furthermore, to address potential information loss when characterizing molecular graphs through graph convolutional networks, SSGraphCPI introduces a comprehensive approach by incorporating both 1D SMILES representations and 2D molecular graphs, thereby incorporating sequence and structural features [111]. In contrast to many other methods for predicting DPI, which primarily emphasize molecular representation, STAMP-DPI places a stronger focus on protein representation and the higher-level relationships between distinct instances [112]. STAMP-DPI employs TAPE encoding to integrate contact maps and GNN as a protein representation and establishes GalaxyDB, an esteemed benchmark dataset specifically designed for DPI prediction. HyperAttentionDTI highlights the incorporation of intricate non-covalent interactions between atoms and amino acids by employing an attention mechanism that assigns an attention vector to each atom and amino acid [113]. HyperAttentionDTI has exhibited noteworthy performance improvements on benchmark datasets. Notably, BridgeDPI merges network-based and learning-based concepts [114]. It constructs a drug–protein association network by introducing a class of virtual nodes designed to bridge the gap between drugs and proteins. Furthermore, it leverages drug molecules and protein sequences as prior knowledge to generate features for interaction prediction.

In the last two years, with the development of protein and small molecule databases, a few researchers tried to build pre-training models of proteins and small molecules for DPI prediction [119, 120]. One such instance is Uni-Mol, a 3D molecular pre-training model. Differing from its counterparts, Uni-Mol directly employs the 3D molecular structure as input to the model, thereby eschewing the use of 1D sequence or 2D graph structure representations. Uni-Mol draws upon its own expansive dataset encompassing 3D structural information of organic small molecules and protein pockets. This model was trained via a unified pre-training framework and strategic tasks on a large-scale distributed cluster. The utilization of 3D information in representation learning empowered Uni-Mol to yield remarkable performance across multiple downstream tasks, while simultaneously facilitating 3D conformation-related endeavors like molecular conformation prediction and protein–ligand binding conformation prediction [112]. To further explore a more comprehensive representation of molecules and proteins, STAMP-DPI employed a pre-training approach to encode the semantic information of small molecules and proteins within an end-to-end deep learning architecture [112]. Protein representation was accomplished via a hybridization of structural topology mapping and Tape Embedding pretraining features, while drug molecules were concurrently represented using molecular mapping and Mol2vec Embedding pretraining features. Leveraging an attention mechanism, STAMP-DPI captured the intricate interaction information between molecules and proteins, ultimately realizing the prediction of DPI.

In recent years, the proliferation of research on DPI prediction has witnessed an escalation in methodological intricacy, paralleled by an augmentation in predictive accuracy. This trend has prompted a thoroughgoing evaluation of this domain. For instance, they highlighted that excessive similarity among samples in the validation set could lead to inflated accuracy levels, while the spurious negative samples could compromise the model's generalizability. In addition to the network architecture, equal emphasis should be placed on the composition of the training dataset. This is especially significant for big data-driven models which rely heavily on data quality.

Prediction of proteins with specific properties

There is also a need to predict specific functional proteins in certain research scenarios [122–124]. The prediction of proteins with specific properties relies on specifically collected datasets, which also has considerable value in application.

There are proteins such as transcription factors or RNA-binding proteins that function by binding to DNA or RNA [125]. Recognizing these kinds of proteins is of great significance for understanding transcriptional and translational regulation. Therefore, studies have been devoted to predicting DNA-binding proteins based on machine learning [126–130]. A number of methods using deep multi-task architectures to predict protein and DNA or RNA binding were published in 2022. DeepDISOBind implemented intrinsically disordered residues (IDR) that predicted the interaction between proteins and DNA as well as RNA. It used common input layers that are followed by different layers that distinguish between DNA and RNA interactions [131]. The classifier architecture mainly consisted of CNN and FNN. Using PSSM, HMM, DSSP and AlphaFold2 predicted structures to jointly construct amino acid features, GraphSite not only improved the performance of predicting protein binding to nucleic acids, but also had the potential to identify binding sites [132]. In predicting proteins of particular functional categories, the utilization of protein structure prediction tools, exemplified by AlphaFold2, offers significant advantages. Huang et al., for instance, implemented a high-throughput protein clustering approach relying on tertiary structural information. They harnessed structural clustering of proteins to identify deaminase functionality. This method facilitated the identification of a deaminase protein strongly amenable to editing in soybean plants, an achievement unattainable through cytosine base editing (CBE) alone. Moreover, their efforts yielded a suite of novel base editing tools endowed with autonomous intellectual property right [133].

There were also studies that attempted to summarize the properties of drug target proteins and identified drug target proteins based on machine learning methods (Table 4). Most of the studies utilized traditional protein representation methods, while in recent years, NLP-based protein representation methods have also been used [134, 135]. Sun et al. evaluated the performance of various combinations of machine learning algorithms for predicting druggable proteins, utilizing Word2Vec to characterize protein sequences and showcasing its potential in this regard [135]. Chen et al. integrated ESM1b, a sequence-based self-supervised pre-trained protein language model, with a graph convolutional neural network classifier to develop an enhanced sequence-based identification method for drug target proteins. The comprehensive model, named QuoteTarget, successfully identified 1,213 potential untapped drug targets when applied to all *Homo sapiens* proteins. Additionally, the authors employed the gradient-weighted class-activation Mapping (Grad-Cam) algorithm to infer residual binding weights from well-trained networks [136]. In terms of classification algorithms, most drug–target protein

Table 3: Algorithms for predicting protein functional gene ontology (GO) annotations.

Author	Algorithm	Protein representation	Pre-training	Advantage	Year	Ref.
Domenico Cozzetto et al.	FFPred 3	258 sequence-derived features	F	Representative functional predictors	2016	[137]
Maxat Kulmanov et al.	DeepGO	AAS, the notion of dense embeddings	F	DeepGO is one of the first DL-based models	2018	[138, 139]
Fuhao Zhang et al.	DeepFunc	Long sparse binary vectors of domains, families, and motifs + Two layers neural network	F	DeepFunc outperforms DeepGO, FFPred3, and GOPDR in effects	2019	[140]
Nils Strodthoff et al.	UDSMProt	RNN, based on AWD-LSTM	T	Achieving advanced performance in many protein classification tasks makes NLP a new paradigm	2019	[141]
Fuhao Zhang et al.	NA	Word2vec, InterPro, Bi-LSTM, multi-scale CNN	F	Combining the local and global semantic features of protein sequences	2020	[142]
Vladimir Gligorijević et al.	DeepFRI	PDB structure, protein domain sequence	T	Structure-based	2021	[62]
Amelia Villegas-Morcillo et al.	NA	Amino acid features, distance maps	T	Combining sequence representation with 3D structural information of proteins does not lead to performance improvement	2021	[143]
Mateo Torres et al.	S2F	HMMER and InterPro	F	S2F introduces a novel label diffusion algorithm to interpret overlapping communities of proteins with related functions	2021	[144]
Boqiao Lai et al.	GAT-GO	RaptorX Inter-Residue Contact, ESM-1b Residue-level Embedding 1D features	T	Protein embedding is performed using sequence and predicted structural information	2022	[145]
Weiqi Xia et al.	PFmulDL	one-hot strategy	F	A transfer learning method and the latest data of GO	2022	[146]
Qianmu Yuan et al.	SPROF-GO	ProtT5-XL-U50	T	Sequence-based pre-trained model and the label diffusion algorithm	2023	[147]
Zhonghui Gu et al.	HEAL	ESM-1b	T	Hierarchical graph converter combined with graph contrast learning	2023	[148]

LSTM, long short-term memory; CNN, convolutional neural network.

prediction algorithms used traditional machine learning algorithms or simple neural networks. And the highest accuracy in these studies was above 90 %. It is worth noting that fewer studies used deep neural network framework in this objective, probably due to the fact that there are not many drug target protein datasets available for training.

Liquid-liquid phase separation is a key principle of intracellular organization in biological systems and has been implicated in a variety of biological processes as well as a range of neurodegenerative diseases. In recent years, there have been many in-depth studies on the LLPS phenomenon of biomolecules [149, 150]. Liquid condensates formed by LLPS are generally thought to be the result of multivalent weak interactions of multiple interacting moieties in multiple folded regions or intrinsically disordered regions (IDRs) [151–153]. Because of this special property, many traditional protein-coding methods may no longer be suitable. The PLAAC tool is web to retrieve protein sequence domains and extract pertinent information encompassing prion-like amino acid compositions [154]. CatGRANULE is an algorithm for a single species [155]. Based on the previously published phase separate protein database PhaSepDB, Chen et al. divided phase separate proteins into two sets of spontaneous phase separate proteins (hSaPS) and interaction dependent phase separate proteins (hPdPS). The distribution of the two phase isolated protein sets and the background protein sets were significantly different from each other by comparing the multimodal characteristics [156]. Chu et al. has developed PSPredictor, a sequence-based protein prediction tool for liquid-liquid phase separation (LLPS), which integrates compositional and sequence information during the protein embedding stage and employs a machine learning algorithm to yield accurate predictions [157].

Prediction of protein function annotation

In the post-AlphaFold era, great progress has been made in predicting protein structures from protein sequences. As a follow-up task of protein structure prediction, protein function identification is the ultimate goal of protein research. The relationship between protein sequence and protein function is a long-standing question in biology.

Since 2000, many researchers have aimed to promote the usage of unified descriptions to annotate the functions of gene products and to assist computational studies [13, 158, 159]. To a certain extent, GO achieves the goal of functional annotation and provides computer-readable functional annotation. GO terms consist of three ontologies: molecular function (MF), biological process (BP), and cell component (CC). GO database comprehensively annotates gene products at multiple levels and is still being updated. The decline of sequencing costs and the development of genome sequencing projects results in a drastic increase in the number of known protein sequences each year, while the functional database corresponding to protein sequences is growing slowly. By integrating a large amount of protein sequence and structural information, combined with comprehensive functional annotation, researchers have attempted to directly predict protein functional annotation, which provided a rapid and accurate reference for a large number of newly discovered proteins.

Protein function annotation algorithms are also of great interest for protein design. Function prediction can provide guidance for unconditional protein generation models to explore the functional space of proteins. For instance, Lisanza et al. developed a diffusion model in sequence space to generate protein structures. In this model, during each round of denoising, a sequence-based function prediction model is employed to compute the gradient of sequence features related to the target function. This process incorporates function-guided gradient descent alongside denoising, progressively making the sequence features to cater to the requirements of the target function. They trained a predictive model for recognizing Immunoglobulin folds to guide the unconditional generation model. Remarkably, 68.7 % of the generated protein structures can be categorized under the same protein fold as existing Immunoglobulin structures [160].

Traditional methods for functional prediction of protein sequences usually require alignment of sequences with large annotated sequence databases using BLASTp or other algorithms. Using the pHMM models constructed by sequence family information provided by Pfam is also a method to predict protein function. However, the search time of the whole dataset is linear with the dataset size, and it is very time-consuming to identify the function of a new protein sequence. Therefore, it is particularly important to use machine learning to predict protein sequence function more quickly and accurately. Although machine learning-based protein function prediction models didn't emerge until 2015, it is developing at a rapid pace (Table 3). Unlike the prediction of protein interactions, the prediction of protein function tends to be a multi-label classification problem. Hence, a distinctive hallmark of this domain lies in the utilization of extensive annotation data in conjunction with deep neural networks. Furthermore, the prediction of protein function annotations draws parallels with research methodologies employed in NLP. In recent times, numerous investigations have employed protein pre-training models to attain commendable performance [62, 141, 143]. ProteinBERT was a deep protein language model that combined language models in pre-training for GO annotation prediction. There were also studies that did not depend on the physical and chemical properties of proteins but used unsupervised label propagation algorithms to predict protein function from the interaction network, which has also achieved good results [144]. SPROF-GO used sequence-based protein pre-training language model to extract sequence information and combined with the label diffusion algorithm to make function prediction [147]. PANNZER was a protein function prediction web server that can be used to predict the functional representation of new genomes [161]. In addition to web servers, there are also open source software that can be installed locally, such as Wei2GO [162]. HEAL utilized a hierarchical graph transformer combined with graph contrastive learning to maximize the similarity between different views represented by the graph, and outperforms DeepFRI on the PDBch test set. In the absence of an experimental protein structure, HEAL outperformed DeepFRI and DeepGOPlus on the AFch test set by utilizing the structure predicted by AlphaFold2. HEAL exhibits proficiency in identifying crucial functional sites through class activation mapping [148].

For protein function prediction, in addition to introduce NLP methods to encode protein sequences, some researchers have made specific exploration of protein characteristics. For example, PFmulDL was proposed to solve the problem that existing prediction methods often misclassify protein families in “rare classes” [146]. PFmulDL combined recurrent neural network with convolutional neural network to expand the number of annotated protein families and improved the performance of protein function prediction for rare categories. Some researchers have also explored whether the addition of protein structure information can improve the prediction accuracy. For example, GAT-GO found that predicted protein contact map can improve the results of protein function prediction. The LM-gvp approach harnessed both one-dimensional protein AAS and three-dimensional structural information for its prediction [167]. This method combined a protein language model with a graph neural network, and demonstrated impressive prediction performance.

Prediction of protein function by biological knowledge graph

The current landscape of protein function prediction models is not without its challenges, as many existing methods struggle to comprehensively capture and effectively leverage biological knowledge. Knowledge graphs present a promising avenue to address these limitations, as

Table 4: Algorithms for drug target protein prediction.

Author	Protein representation	Method	Accuracy	Year	Ref.
Lian Yi Han et al.	A descriptor encoding the structural and physicochemical properties of a protein	SVM	83.70 %	2007	[163]
Qingliang Li et al.	Composition of the amino acid residues, Hydrophobicity, Polarity, polarizability, Charge, Solvent accessibility, Normalized van der Waals volume	SVM	84.00 %	2007	[164]
Ali Akbar Jamali1 et al.	Three different sets of physicochemical properties	SVM	89.78 %	2016	[134]
Tanlin Sun et al.	Word2vec, auto covariance, Cojoint Triad	CNN and Traditional machine learning methods	89.55 %	2018	[135]
Phasit Charoenkwan et al.	Amino acid composition, amphiphilic pseudo-amino acid composition, dipeptide composition, Composition-Transition-Distribution, pseudo amino acid composition	SVM	91.90 %	2022	[165, 166]
Rahu Sikander et al.	Grouped amino acid composition (GDPC), reduced amino acid alphabet (RAAA), novel encoder pseudo amino acid segmentation (S-PseAAC)	ERT, XGB, RF	93.78 %	2022	[167]
Lezheng Yu et al.	Dictionary, dipeptide composition, tripeptide composition, Composition-Transition-Distribution	CNN-RNN	92.40 %	2022	[122]
Jiaxiao Chen et al.	ESM1b, predicted contact map	GCN	95.00 %	2023	[136]

SVM, support vector machine; GCN, graph convolutional network; CNN, convolutional neural network; RNN, recurrent neural network; XGB, eXtreme gradient boosting; ERT, ensemble of regression tress; RF, random forest.

they possess the capacity to amalgamate information from extensive biomedical knowledge databases through a graph-based representation. This framework is particularly relevant for tasks involving the prediction of protein properties.

The construction of biomedical knowledge graph relies on a variety of data sources, including unstructured and structured databases. Currently, prominent knowledge repositories compile information centered around proteins, each database emphasizing distinct data types that contribute to the formulation of the knowledge graph. For example, DrugBank [168] and SuperTarget [169] mainly contain pharmaceutical properties, and PubChem [170] and ChEMBL [171] mainly contain functional and biological activities of compounds. KEGG [172] mainly includes genome, biochemical reaction information and pathway information. InterPro [173] integrates multiple databases to summarize protein sequences into protein families.

Knowledge graphs combine these data sources to model complex associations between different types of biological entities, such as drugs, proteins, antibodies, etc. In modeling, various types of relationships between entities are included, expressed as different association semantics. Traditional biological networks help to recognize network topological relationships and clarify associations between entities, but their learning depends on path exploration processes, with high computational and spatial costs and limited scalability. In recent years, with the development of computer technology, new methods for mining and graph modeling of high dimensional biomedical networks have emerged. Entities and associations are projected into low-dimensional spaces by a knowledge graph embedding (KGE) model, and low-rank vector or matrix representations of graph nodes and edges are learned to preserve the inherent structure of the graph.

Knowledge graph can be constructed using a variety of methods, such as Translation-based models, Tensor factorization-based models, Neural network-based models. Methods of tensor factorization include local linear embedding (LLE) and Laplacian feature mapping (LE), which build networks from non-relational data. The embedding vector is obtained by factorization of the adjacency matrix between nodes and adjacent nodes. Neural network-based models use deep architectures, such as SDNE and DNGR, which are based on deep autoencoder architectures.

At present, many researchers have constructed knowledge graphs related to biomedicine, such as GNBR and DRKG, Hetionet, CBKH [174]. They all got information from known publicly available data sets or from the biomedical literature. PharmKG is a biomedical knowledge graph that connects over 500,000 individuals related to genes, drugs, and diseases. It contains diverse information specific to the domain of biomedicine derived from various omics data sources such as gene expression, chemical structure, and disease word embeddings while maintaining semantic and biomedical features [175]. PrimeKG is a comprehensive knowledge graph designed for precise analysis in the field of precision medicine. These scales include perturbations in disease-associated proteins, biological processes and pathways, anatomical and phenotypic aspects, as well as an extensive collection of approved drugs along with their therapeutic effects [176]. According to the GO and the Uniprot knowledge base, ProteinKG65 incorporates diverse information by aligning descriptions and protein sequences into GO terms and protein entities [177]. Biswas et al. suggested a technique for constructing a biological knowledge graph using tensor factorization. The approach involves incorporating complex-valued embeddings into the knowledge graph, which includes information on disease gene associations and relevant contextual details [178].

Entities and associations of biological networks are represented as matrices and vectors through KEGs, which allows traditional machine learning methods to be applied to downstream tasks related to embedding of biological entities, such as link prediction and node classification. More specifically, a combination of network embedding techniques and machine learning methods can cluster proteins, drugs, or study drug–gene–disease correlations. KEG provides an effective paradigm for promoting data integration in the biomedical field.

In order to conduct a comprehensive evaluation of various graph embedding techniques, Yue et al. selected 11 representative approaches and systematically compared their performance across three crucial biomedical tasks: prediction of drug–disease associations (DDA), drug–drug interactions (DDI), and PPI. Additionally, they also performed two node classification tasks involving the categorization of medical terms based on semantic types and the prediction of protein functions. The experimental findings suggest that graph embedding methods have yielded promising results. The study conducted by Vlietstra et al. aimed to assess the feasibility of utilizing protein knowledge maps for identifying targeted genes associated with disease-related non-coding SNPs, achieved through a comprehensive evaluation and comparison of six established methodologies for protein knowledge mapping [179].

In addition to the comprehensive evaluation of knowledge graphs constructed by previous researchers, there are also researchers who construct their own knowledge graphs for the discovery of potential drug targets or the calculation of drug–target interactions. Himmelstein et al. systematically simulated the efficacy of 755 existing treatments using

Hetionet, a model that integrates knowledge from millions of biomedical studies and connects various entities such as compounds, diseases, genes, anatomical structures, pathways, biological processes, molecular functions, cell components, pharmacological classes, side effects and symptoms. The predicted results were validated with two external treatment groups [180]. The TriModel, a knowledge graph embeddings model, constructs the knowledge base using multi-part embeddings. It generates vector representations for all drugs and targets in the knowledge graph to compute candidate drug target interactions [181]. KGE_NFM was a new method for drug–protein interaction prediction based on knowledge graph and recommendation system. In addition to the traditional representation method, KGE_NFM combined knowledge graph and recommendation system method-neural factorization machine (NFM) to predict drug target interaction, which improved the accuracy and stability in real scenarios [119].

Fernández-Torras et al. constructed a comprehensive knowledge graph, Bioteque, which encompasses over 450,000 biological entities and 30 million relationships among them. Bioteque serves as a valuable tool for scrutinizing high-throughput PPIs data and predicting drug responses. The graph comprises 12 biological entities such as genes, diseases, drugs, drugs used to treat diseases, and 67 types of associations including gene–gene interactions [182]. Nasiri et al. approach the problem of predicting PPIs as a link prediction task in attribute networks, utilizing attribute embedding techniques to forecast interactions between proteins within the PPI network. The key aspect of this method is assigning weights to features based on their significance, enabling differentiation of each feature's contribution [183].

The biological network plays a crucial role in the biomedical field as it serves as the primary source of data for data-driven problems. knowledge graph embedding techniques enable information-rich representations, facilitating knowledge graph-based problem-solving through traditional machine learning methods. These techniques have been extensively employed in various biomedical applications and are instrumental in protein function prediction.

Conclusion and perspective

In this article, we review the development history and research paradigms of computational methods for predicting protein function. We then summarize common approaches to protein and molecular representation and feature extraction. Furthermore, we evaluated the performance of the machine learning-based algorithms in four task objectives of protein function prediction, which provided a comprehensive perspective for understanding the field.

In the realm of protein function prediction, the landscape of downstream tasks has seen a limited evolution in the realm of classification algorithms in recent times. Traditional machine learning techniques such as SVMs and random forests continue to effectively address a substantial portion of prediction requirements. As a result, the necessity for deep neural networks in this context remains somewhat moderate. Conversely, the central focus has shifted to refining protein representation and feature extraction methodologies. The prevalent utilization of feature extraction techniques, which involve the deliberate design of features based on AAS, remains a predominant strategy. This knowledge-based encoding approach offers the flexibility to tailor features to specific task objectives, albeit it may struggle to concurrently address multiple objectives. Significantly, a notable trend has emerged in recent years wherein pre-training models derived from NLP are demonstrating their advantages. Protein pre-training models are now being applied to various protein function prediction tasks, notably in the realm of predicting protein function annotations. Several of these protein pre-training models have exhibited a robust generalization capacity across a spectrum of downstream tasks.

In addition to the aforementioned objectives that we have elucidated, the field of protein function prediction encompasses a multitude of intricate downstream tasks. While traditional computational methodologies retain a significant foothold, they might encounter limitations in accurately identifying certain proteins endowed with specific properties. Phase-separated proteins, for example, often contain many IDRs [184]. Algorithms based on traditional machine learning to predict phase-separated proteins tend to cover only specific scenarios or datasets [154, 156, 184]. In this case, with the accumulation of massive data, data-driven protein pre-training models may have great potential in the prediction of complex protein functions. In addition, combining self-supervised deep learning protein characterization methods with clustering algorithms, researchers have opportunity to identify new protein classes with specific functions [133]. Furthermore, the addition of sequence and structure information makes it easier for the network to learn a relatively complete protein space, which is conducive to rational design and transformation of proteins [185].

For protein function prediction, with the development of ChatGPT, large language models based on protein language have also been developed. For example, ProteinChat and DrugChat are large language models focusing on protein function and small molecule properties respectively. Within the scope of existing literature research, these large language models can effectively answer the questions raised by users in the form of text interaction, including the query of protein function and properties, queries for specific protein–small molecule interactions, etc. However, current large language models may not be able to outperform supervised deep learning algorithms for specific prediction tasks with clear objectives. This performance variance could arise due to differences in training data configurations, feature extraction methodologies, and model architectures. The distinctive value of these large language models, however, might reside in their capability to synthesize the internal logic governing the functioning and interactions of proteins and small molecules. Recently, large language models have been applied to single cell RNA-seq, successfully learning the gene regulatory networks as well as protein interaction networks in cells. The large language model based on biological data can predict gene expression of the perturbed networks without specific supervised training. Thus, these models may harbor the potential to illuminate unexplored inquiries that have yet to be thoroughly investigated [186].

Research ethics: The local Institutional Review Board deemed the study exempt from review.

Author contributions: Jiaxiao Chen, Zhonghui Gu, Luhua Lai and Jianfeng Pei wrote the paper. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Authors state no conflict of interest.

Research funding: This work has been supported in part by the National Natural Science Foundation of China (22033001), the National Key R&D Program of China (2022YFA1303700) and the Chinese Academy of Medical Sciences (2021-I2M-5-014).

Data availability: Data are openly available in a public repository.

References

1. Avery C, Patterson J, Grear T, Frater T, Jacobs DJ. Protein function analysis through machine learning. Biomolecules 2022;12:1246.
2. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
3. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021;373:871–6.
4. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet 2014;15:829–45.
5. Song H, Liu B, Huai W, Yu Z, Wang W, Zhao J, et al. The E3 ubiquitin ligase TRIM31 attenuates NLRP3 inflammasome activation by promoting proteasomal degradation of NLRP3. Nat Commun 2016;7: 1–11.
6. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015; 43:D447–52.
7. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell 2014;157:1262–78.
8. Berggård T, Linse S, James P. Methods for the detection and analysis of protein–protein interactions. Proteomics 2007;7:2833–42.
9. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote) omics data. Nat Methods 2016;13:731–40.
10. Consortium U. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15.
11. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 2007;35:D301–D3.
12. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, et al. CAPRI: a critical assessment of predicted interactions. Proteins 2003;52:2–9.
13. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25: 25–9.
14. Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol 2019;20:244.
15. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015;349:255–60.
16. Zhang S, Fan R, Liu Y, Chen S, Liu Q, Zeng W. Applications of transformer-based language models in bioinformatics: a survey. Bioinform Adv 2023;3:vbad001.
17. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 2007;8:995–1005.
18. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
19. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002;30: 1575–84.
20. Enright AJ, Ouzounis CA. GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics 2000;16:451–7.
21. Jothi R, Cherukuri PF, Tasneem A, Przytycka TM. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein–protein interactions. J Mol Biol 2006;362:861–75.
22. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. Co-evolution of proteins with their interaction partners. J Mol Biol 2000;299: 283–93.
23. Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng 2001;14:609–14.
24. Cai CZ, Han LY, Ji ZL, Chen YZ. Enzyme family classification by support vector machines. Proteins 2004;55:66–76.
25. Huang N, Chen H, Sun Z. CTKPred: an SVM-based method for the prediction and classification of the cytokine superfamily. Protein Eng Des Sel 2005;18:365–8.

26. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A. PRISM: protein interactions by structural matching. Nucleic Acids Res 2005;33: W331–6.
27. Chen R, Tong W, Mintseris J, Li L, Weng Z. ZDOCK predictions for the CAPRI challenge. Proteins: Struct Funct Bioinf 2003;52:68–73.
28. Cai YD, Lin SL. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Biochim Biophys Acta 2003;1648:127–33.
29. Han LY, Cai CZ, Lo SL, Chung MC, Chen YZ. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA 2004;10:355–68.
30. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 2002;18: 147–59.
31. Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 2006;34: W32–7.
32. Yu H, Chen J, Xu X, Li Y, Zhao H, Fang Y, et al. A systematic prediction of multiple drug–target interactions from chemical, genomic, and pharmacological data. PLoS One 2012;7:e37608.
33. Zhang W, Qu Q, Zhang Y, Wang W. The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. Neurocomputing 2018;273:526–34.
34. Manavalan B, Basith S, Shin TH, Wei L, Lee G. mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics 2019;35:2757–65.
35. Nanni L, Lumini A, Brahnam S. An empirical study on the matrix-based protein representations and their combination with sequence-based approaches. Amino Acids 2013;44:887–901.
36. Huang YA, You ZH, Gao X, Wong L, Wang L. Using weighted sparse representation model combined with discrete cosine transformation to predict protein–protein interactions from protein sequence. BioMed Res Int 2015;2015:902198.
37. Huang YA, You ZH, Chen X. A systematic prediction of drug–target interactions using molecular fingerprints and protein sequences. Curr Protein Pept Sci 2018;19:468–78.
38. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–8.
39. Zahiri J, Yaghoubi O, Mohammad-Noori M, Ebrahimpour R, Masoudi-Nejad A. PPIevo: protein–protein interaction prediction from PSSM based evolutionary information. Genomics 2013;102:237–42.
40. cheol Jeong J, Lin X, Chen X-W. On position-specific scoring matrix for protein function prediction. IEEE ACM Trans Comput Biol Bioinf 2010; 8:308–15.
41. Li Y, Wang Z, Li LP, You ZH, Huang WZ, Zhan XK, et al. Robust and accurate prediction of protein–protein interactions by exploiting evolutionary information. Sci Rep 2021;11:1–12.
42. Yu L, Guo Y, Zhang Z, Li Y, Li M, Li G, et al. SecretP: a new method for predicting mammalian secreted proteins. Peptides 2010;31: 574–8.
43. Wen Z, Li M, Li Y, Guo Y, Wang K. Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 2007;32:277–83.
44. Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res 2008;36:3025–30.
45. Wang X, Wang R, Wei Y, Gui Y. A novel conjoint triad auto covariance (CTAC) coding method for predicting protein–protein interaction based on amino acid sequence. Math Biosci 2019;313:41–7.
46. Luo J, Yu L, Guo Y, Li M. Functional classification of secreted proteins by position specific scoring matrix and auto covariance. Chemometr Intell Lab Syst 2012;110:163–7.
47. Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, et al. PIPE: a protein–protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. BMC Bioinf 2006;7:1–15.
48. Li Y, Ilie L. SPRINT: ultrafast protein–protein interaction prediction of the entire human interactome. BMC Bioinf 2017;18:1–11.
49. Wang YC, Wang XB, Yang ZX, Deng NY. Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. Protein Pept Lett 2010;17:1441–9.
50. Wang H, Hu X. Accurate prediction of nuclear receptors with conjoint triad feature. BMC Bioinf 2015;16:1–13.
51. Dey L, Mukhopadhyay A. A classification-based approach to prediction of dengue virus and human protein–protein interactions using amino acid composition and conjoint triad features. In: IEEE region 10 symposium (TENSYMP) 2019. IEEE; 2019.
52. Wang H, Wu P. Prediction of RNA–protein interactions using conjoint triad feature and chaos game representation. Bioengineered 2018;9: 242–51.
53. Wang YC, Wang Y, Yang ZX, Deng NY. Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical context. BMC Syst Biol 2011;5:1–11.
54. You ZH, Chan KC, Hu P. Predicting protein–protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. PLoS One 2015;10: e0125811.
55. You ZH, Zhu L, Zheng CH, Yu HJ, Deng SP, Ji Z. Prediction of protein–protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. In: BMC bioinformatics. Springer; 2014.
56. Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. Comput Struct Biotechnol J 2021;19: 1750–8.
57. Asgari E, Mofrad MRK. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One 2015;10:e0141287.
58. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. Prottrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2021;44: 7112–27.
59. Elnaggar A, Heinzinger M, Dallago C, Rihawi G, Wang Y, Jones L, et al. ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. ArXiv preprint arXiv:2007.06225, 2020.
60. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 2022;38:2102–10.
61. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 2021;118: e2016239118.
62. Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun 2021;12:3168.

63. Zhang Z, Xu M, Jamasb A, et al Protein representation learning by geometric structure pretraining. arXiv preprint arXiv:2203.06125. 2022.
64. Guo Y, Wu J, Ma H, Huang J. Self-supervised pre-training for protein embeddings using tertiary structures. In: Proceedings of the AAAI conference on artificial intelligence; 2022.
65. Sarkar D, Saha S. Machine-learning techniques for the prediction of protein–protein interactions. J Biosci 2019;44:104.
66. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Med Chem 1998;19: 1639–62.
67. Zhou P, Jin B, Li H, Huang SY. HPEPDOCK: a web server for blind peptide–protein docking based on a hierarchical algorithm. Nucleic Acids Res 2018;46:W443–50.
68. Yan Y, Tao H, He J, Huang S-Y. The HDOCK server for integrated protein–protein docking. Nat Protoc 2020;15:1829–52.
69. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–43.
70. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, et al. A critical assessment of docking programs and scoring functions. J Med Chem 2006;49:5912–31.
71. Huber T, Torda AE, Van Gunsteren WF. Local elevation: a method for improving the searching properties of molecular dynamics simulation. J Comput Aided Mol Des 1994;8:695–708.
72. Feig M. Local protein structure refinement via molecular dynamics simulations with locPREFMD. J Chem Inf Model 2016;56:1304–12.
73. Bock JR, Gough DA. Predicting protein–protein interactions from primary structure. Bioinformatics 2001;17:455–60.
74. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. In: Proceedings. IEEE computer society bioinformatics conference. IEEE; 2002;197–206 pp.
75. Deng M, Mehta S, Sun F, et al Inferring domain-domain interactions from protein-protein interactions. In: Proceedings of the sixth annual international conference on Computational biology; 2002: 117–126 pp.
76. Rodrigues CHM, Myung Y, Pires DEV, Ascher DB. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. Nucleic Acids Res 2019;47:W338–44.
77. Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinf 2017; 18:1–8.
78. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein–protein interactions using AlphaFold2. Nat Commun 2022;13:1265.
79. Hanggara FS, Anam K. Sequence-based protein–protein interaction prediction using greedy layer-wise training of deep neural networks. In: AIP conference proceedings. AIP Publishing LLC; 2020.
80. A comprehensive SARS-CoV-2–human protein–protein interactome network identifies pathobiology and host-targeting therapies for COVID-19. Nat Biotechnol 2023;41:1–39.
81. Kovács IA, Luck K, Spirohn K, Wang Y, Pollis C, Schlabach S, et al. Network-based prediction of protein interactions. Nat Commun 2019; 10:1240.
82. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci USA 2007;104:4337–41.
83. Eid FE, ElHefnawi M, Heath LS. DeNovo: virus-host sequence-based protein–protein interaction prediction. Bioinformatics 2016;32: 1144–50.
84. Pan XY, Zhang YN, Shen HB. Large-Scale prediction of human protein–protein interactions from amino acid sequence based on latent topic features. J Proteome Res 2010;9:4992–5001.
85. Hashemifar S, Neyshabur B, Khan AA, Xu J. Predicting protein–protein interactions through sequence-based deep learning. Bioinformatics 2018;34:i802–10.
86. Xue Y, Liu Z, Fang X, et al. Multimodal pre-training model for sequence-based prediction of protein-protein interaction. In: Machine learning in computational biology. PML; 2022;34–46 pp.
87. Song B, Luo X, Luo X, Liu Y, Niu Z, Zeng X. Learning spatial structures of proteins improves protein–protein interaction prediction. Briefings Bioinf 2022;23:bbab558.
88. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021: 2021:463034.
89. Gao M, Nakajima AD, Parks JM, Skolnick J. AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. Nat Commun 2022;13:1744.
90. Cheng Y, Gong Y, Liu Y, Song B, Zou Q. Molecular design in drug discovery: a comprehensive review of deep generative models. Briefings Bioinf 2021;22:bbab344.
91. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 2018;4:268–76.
92. Schwalbe-Koda D, Gómez-Bombarelli R. Generative models for automatic chemical design. Mach Learn Meets Quantum Phys 2020: 445–67. https://doi.org/10.1007/978-3-030-40245-7_21.
93. Thomas N, Smidt T, Kearnes S, Yang L, Li L, Kohlhoff K, et al. Tensor field networks: rotation-and translation-equivariant neural networks for 3d point clouds. ArXiv preprint arXiv:1802.08219, 2018.
94. Kondor R. N-body networks: a covariant hierarchical neural network architecture for learning atomic potentials. ArXiv preprint arXiv: 1803.01588, 2018.
95. Jing B, Eismann S, Suriana P, Townshend RJ, Dror R. Learning from protein structure with geometric vector perceptrons. ArXiv preprint arXiv:2009.01411, 2020.
96. Satorras VG, Hoogeboom E, Welling M. E(n) equivariant graph neural networks. In: International conference on machine learning. PMLR; 2021.
97. Wang Y, Wang J, Cao Z, Barati Farimani A. Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell 2022;4:279–87.
98. Wang Y, Magar R, Liang C, Barati Farimani A. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. J Chem Inf Model 2022;62:2713–25.
99. Liu S, Wang H, Liu W, Lasenby J, Guo H, Tang J. Pre-training molecular graph representation with 3d geometry. ArXiv preprint arXiv: 2110.07728, 2021.
100. Liu S, Guo H, Tang J. Molecular geometry pretraining with se (3)-invariant denoising distance matching. ArXiv preprint arXiv: 2206.13602, 2022.
101. Chen R, Liu X, Jin S, Lin J, Liu J. Machine learning for drug-target interaction prediction. Molecules 2018;23:2208.
102. Jain AN. Scoring functions for protein-ligand docking. Curr Protein Pept Sci 2006;7:407–20.
103. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 2010;31:455–61.

104. Huang SY, Grinter SZ, Zou X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. Phys Chem Chem Phys 2010;12:12899–908.
105. Guo ZH, Yi HC, You ZH. Construction and comprehensive analysis of a molecular association network via lncRNA–miRNA–disease–drug–protein graph. Cells 2019;8:866.
106. Liu H, Zhang W, Nie L, Ding X, Luo J, Zou L. Predicting effective drug combinations using gradient tree boosting based on features extracted from drug–protein heterogeneous network. BMC Bioinf 2019;20:1–12.
107. Zhao L, Ciallella HL, Aleksunes LM, Zhu H. Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. Drug Discov Today 2020;25:1624–38.
108. Nguyen NQ, Jang G, Kim H, Kang J. Perceiver CPI: a nested cross-attention network for compound–protein interaction prediction. Bioinformatics 2022;39:btac731.
109. Wang J, Dokholyan NV. Yuel: improving the generalizability of structure-free compound-protein interaction prediction. J Chem Inf Model 2022;62:463–71.
110. Yazdani-Jahromi M, Yousefi N, Tayebi A, Kolanthai E, Neal CJ, Seal S, et al. AttentionSiteDTI: an interpretable graph-based model for drug–target interaction prediction using NLP sentence-level relation classification. Briefings Bioinf 2022;23:bbac272.
111. Wang X, Liu J, Zhang C, Wang S. SSGraphCPI: a novel model for predicting compound-protein interactions based on deep learning. Int J Mol Sci 2022;23:3780.
112. Wang P, Zheng S, Jiang Y, Li C, Liu J, Wen C, et al. Structure-Aware multimodal deep learning for drug-protein interaction prediction. J Chem Inf Model 2022;62:1308–17.
113. Zhao Q, Zhao H, Zheng K, Wang J. HyperAttentionDTI: Improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics 2022;38:655–62.
114. Wu Y, Gao M, Zeng M, Zhang J, Li M. BridgeDPI: a novel Graph Neural Network for predicting drug–protein interactions. Bioinformatics 2022;38:2571–8.
115. Nagamine N, Sakakibara Y. Statistical prediction of protein chemical interactions based on chemical structure and mass spectrometry data. Bioinformatics 2007;23:2004–12.
116. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 2008;24:i232–40.
117. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, et al. Deep-learning-based drug–target interaction prediction. J Proteome Res 2017;16: 1401–9.
118. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics 2018;34:i821–9.
119. Ye Q, Hsieh CY, Yang Z, Kang Y, Chen J, Cao D, et al. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. Nat Commun 2021;12:6775.
120. Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, et al. Uni-mol: a universal 3D molecular representation learning framework. In: The eleventh international conference on learning representations; 2023.
121. Chelur VR, Priyakumar UD. BiRDS-binding residue detection from protein sequences using deep ResNets. J Chem Inf Model 2022;62: 1809–18.
122. Yu L, Xue L, Liu F, Li Y, Jing R, Luo J. The applications of deep learning algorithms on in silico druggable proteins identification. J Adv Res 2022;219–31. https://doi.org/10.1016/j.jare.2022.01.009.
123. Vernon RM, Chong PA, Tsang B, Kim TH, Bah A, Farber P, et al. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. Elife 2018;7. https://doi.org/10.7554/elife.31486.
124. Vernon RM, Forman-Kay JD. First-generation predictors of biological protein phase separation. Curr Opin Struct Biol 2019;58:88–96.
125. Hudson WH, Ortlund EA. The structure, function and evolution of proteins that bind DNA and RNA. Nat Rev Mol Cell Biol 2014;15:749–60.
126. Shadab S, Alam Khan MT, Neezi NA, Adilina S, Shatabda S. DeepDBP: deep neural networks for identification of DNA-binding proteins. Comput Biol Med 2020;19:100318.
127. Hu S, Ma R, Wang H. An improved deep learning method for predicting DNA-binding proteins based on contextual features in amino acid sequences. PLoS One 2019;14:e0225317.
128. Ali F, Kabir M, Arif M, Khan Swati ZN, Khan ZU, Ullah M, et al. DBPPred-PDSD: machine learning approach for prediction of DNA-binding proteins using Discrete Wavelet Transform and optimized integrated features space. Chemometrics Intellig Lab Syst 2018;182:21–30.
129. Ali F, Ahmed S, Swati ZNK, Akbar S. DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information. J Comput Aided Mol Des 2019;33: 645–58.
130. Si J, Cui J, Cheng J, Wu R. Computational prediction of RNA-binding proteins and binding sites. Int J Mol Sci 2015;16:26303–17.
131. <Auditory sensitivity provided by self-tuned critical oscillations of hair cells.pdf>.
132. Shi W, Singha M, Pu L, Srivastava G, Ramanujam J, Brylinski M. GraphSite: ligand binding site classification with deep graph learning. Biomolecules 2022;12:1053.
133. Huang J, Lin Q, Fei H, He Z, Xu H, Li Y, et al. Discovery of deaminase functions by structure-based protein clustering. Cell 2023;186: 3182–95.e14.
134. Jamali AA, Ferdousi R, Razzaghi S, Li J, Safdari R, Ebrahimie E. DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. Drug Discov Today 2016; 21:718–24.
135. Sun T, Lai L, Pei J. Analysis of protein features and machine learning algorithms for prediction of druggable proteins. Quantitative Bio 2018;6:334–43.
136. Chen J, Gu Z, Xu Y, Deng M, Lai L, Pei J. QuoteTarget: a sequence-based transformer protein language model to identify potentially druggable protein targets. Protein Sci 2023;32:e4555.
137. Cozzetto D, Minneci F, Currant H, Jones DT. FFPred 3: feature-based function prediction for all Gene Ontology domains. Sci Rep 2016;6: 31865.
138. Kulmanov M, Khan MA, Hoehndorf R, Wren J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics 2018;34:660–8.
139. Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 2020;36:422–9.
140. Zhang F, Song H, Zeng M, Li Y, Kurgan L, Li M. DeepFunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. Proteomics 2019;19: 1900019.
141. Strodthoff N, Wagner P, Wenzel M, Samek W. UDSMProt: universal deep sequence models for protein classification. Bioinformatics 2020; 36:2401–9.
142. Zhang F, Song H, Zeng M, Wu FX, Li Y, Pan Y, et al. A deep learning framework for gene ontology annotations with sequence- and

network-based information. IEEE ACM Trans Comput Biol Bioinf 2021;18:2208–17.

143. Villegas-Morcillo A, Makrodimitris S, van Ham R, Gomez AM, Sanchez V, Reinders MJT. Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. Bioinformatics 2021;37:162–70.

144. Torres M, Yang H, Romero AE, Paccanaro A. Protein function prediction for newly sequenced organisms. Nat Mach Intell 2021;3:1050–60.

145. Lai B, Xu J. Accurate protein function prediction via graph attention networks with predicted structure information. Briefings Bioinf 2022;23:bbab502.

146. Xia W, Zheng L, Fang J, Li F, Zhou Y, Zeng Z, et al. PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. Comput Biol Med 2022;145:105465.

147. Yuan Q, Xie J, Xie J, Zhao H, Yang Y. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. Briefings Bioinf 2023;24:bbad117.

148. Gu Z, Luo X, Chen J, Deng M, Lai L. Hierarchical graph transformer with contrastive learning for protein function prediction. Bioinformatics 2023;39:btad410.

149. Brangwynne CP, Mitchison TJ, Hyman AA. Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. Proc Natl Acad Sci USA 2011;108:4334–9.

150. Hyman AA, Brangwynne CP. Beyond stereospecificity: liquids and mesoscale organization of cytoplasm. Dev Cell 2011;21:14–6.

151. Harmon TS, Holehouse AS, Pappu RV. Differential solvation of intrinsically disordered linkers drives the formation of spatially organized droplets in ternary systems of linear multivalent proteins. New J Phys 2018;20:045002.

152. Alberti S, Halfmann R, King O, Kapila A, Lindquist S. A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. Cell 2009;137:146–58.

153. Lin YH, Forman-Kay JD, Chan HS. Theories for sequence-dependent phase behaviors of biomolecular condensates. Biochemistry 2018;57:2499–508.

154. Lancaster AK, Nutter-Upham A, Lindquist S, King OD. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. Bioinformatics 2014;30:2501–2.

155. Bolognesi B, Gotor NL, Dhar R, Cirillo D, Baldrighi M, Tartaglia GG, et al. A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression. Cell Rep 2016;16:222–31.

156. Chen Z, Hou C, Wang L, Yu C, Chen T, Shen B, et al. Screening membraneless organelle participants with machine-learning models that integrate multimodal features. Proc Natl Acad Sci USA 2022;119:e2115369119.

157. Chu X, Sun T, Li Q, Xu Y, Zhang Z, Lai L, et al. Prediction of liquid–liquid phase separating proteins using machine learning. BMC Bioinf 2022;23:1–13.

158. Dessimoz C, Škunca N. The gene ontology handbook. Humana Press: SpringerOpen, New York; 2017.

159. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 2004;32:5539–45.

160. Lisanza SL, Gershon JM, Tipps SWK, Arnoldt L, Hendel S, Sims JN, et al. Joint generation of protein sequence and structure with RoseTTAFold sequence space diffusion. bioRxiv 2023:2023.05.08.539766.

161. Törönen P, Holm L. PANNZER—a practical tool for protein function prediction. Protein Sci 2022;31:118–28.

162. Reijnders MJ. Wei2GO: weighted sequence similarity-based protein function prediction. PeerJ 2022;10:e12931.

163. Han LY, Zheng CJ, Xie B, Jia J, Ma XH, Zhu F, et al. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. Drug Discov Today 2007;12:304–13.

164. Li Q, Lai L. Prediction of potential drug targets based on simple sequence properties. BMC Bioinf 2007;8:353.

165. Charoenkwan P, Schaduangrat N, Moni MA, Shoombuatong W, Manavalan B. Computational prediction and interpretation of druggable proteins using a stacked ensemble-learning framework. iScience 2022;25:104883.

166. Sikander R, Ghulam A, Ali F. XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. Sci Rep 2022;12:1–9.

167. Wang Z, Combs SA, Brand R, Calvo MR, Xu P, Price G, et al. Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. Sci Rep 2022;12:6832.

168. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074–82.

169. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, et al. SuperTarget and Matador: resources for exploring drug–target relationships. Nucleic Acids Res 2007;36:D919–22.

170. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47:D1102–9.

171. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40:D1100–7.

172. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.

173. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Res 2023;51:D418–27.

174. Zeng X, Tu X, Liu Y, Fu X, Su Y. Toward better drug discovery with knowledge graph. Curr Opin Struct Biol 2022;72:114–26.

175. Zheng S, Rao J, Song Y, Zhang J, Xiao X, Fang EF, et al. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. Briefings Bioinf 2021;22:bbaa344.

176. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data 2023;10:67.

177. Cheng S, Liang X, Bi Z, Zhang N, Chen H. ProteinKG65: a knowledge graph for protein science. ArXiv preprint arXiv:2207.10080, 2022.

178. Biswas S, Mitra P, Rao KS. Relation prediction of co-morbid diseases using knowledge graph completion. IEEE ACM Trans Comput Biol Bioinf 2019;18:708–17.

179. Vlietstra WJ, Vos R, van Mulligen EM, Jenster GW, Kors JA. Identifying genes targeted by disease-associated non-coding SNPs with a protein knowledge graph. PLoS One 2022;17:e0271395.

180. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife 2017;6:e26726.

181. Mohamed SK, Nováček V, Nounu A. Discovering protein drug targets using knowledge graph embeddings. Bioinformatics 2020;36:603–10.

182. Fernández-Torras A, Duran-Frigola M, Bertoni M, Locatelli M, Aloy P. Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque. Nat Commun 2022;13:5304.

183. Nasiri E, Berahmand K, Rostami M, Dabiri M. A novel link prediction algorithm for protein–protein interaction networks by attributed graph embedding. Comput Biol Med 2021;137: 104772.

184. Ray S, Maji SK. Predictable phase-separated proteins. Nat Chem 2020; 12:787–9.

185. Bennett NR, Coventry B, Goreshnik I, Huang B, Allen A, Vafeados D, et al. Improving de novo protein binder design with deep learning. Nat Commun 2023;14:2625.

186. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature 2023;618:616–24.