Communications in Information and Systems
Volume 23, Number 2, 151–184, 2023

A review of computational models for predicting protein-protein interaction and non-interaction

Nan Zhao and Xinqi Gong*

Predicting potential protein-protein interaction and non-interaction are vital to study the mechanism of protein function. Traditional experimental technologies show their disadvantages of being expensive, time-consuming and laborious. Numerous computational methods have been developed to detect potential interacting and non-interacting protein partners. This paper reviews recent advancements in effective computational models for protein-protein interactions and non-interactions prediction. We classified the computational methods based on the protein information types into five different categories and introduced the main ideas, advantages and disadvantages of algorithms in each category. To obtain a high-quality dataset, we analyzed the collection methods and composition of positive and negative samples in detail and described some applications of real non-interacting protein pairs. Finally, we summarized some challenges and open issues in the future.

1. Introduction

Protein is the direct executor of life activities, and its structure and dynamic properties are closely related to physiological functions. Protein-protein interactions (PPIs) are a fundamental component of intracellular and intercellular communication, which play a vital role in all cellular functions, such as cell development, cell metabolism, signal transduction, or cell apoptosis. Consequently, correctly detecting PPIs is helpful in understanding protein functions in-depth and essential for advancing structure-based drug design and disease treatment. The opposite of PPIs is protein-protein non-interactions (PPNIs), where the number of non-interacting protein pairs far exceeds that of interacting protein pairs. Identifying PPNIs facilitates the collection of high-quality negative samples for more efficient characterization of protein pairs. Furthermore, knowledge of PPNIs contributes to protein

*Corresponding author.

complex non-contact chains and is critical for predicting the structure of super-large protein complexes.

Many high-throughput technologies have been developed to determine protein partners, including yeast two-hybrid [1, 2], mass spectrometry [3, 4], protein chips [5] and tandem affinity purification [6, 7]. While these techniques can detect PPIs on a large scale, they exhibit time-consuming, labor-intensive, and have a high fraction of false positive rates and low agreements with each other [8, 9]. Furthermore, even well-studied organisms have a protein-protein interaction network (PPI network) that is sketchy at best [10]. This highlights that various computational methods are still needed for PPIs and PPNIs prediction complementary to experimental methods.

Dataset quality, feature representation and classifier selection are three major determinants of the generalizability of predictive models. For classifier selection, Hu et al. [11] and Soleyman et al. [12] focused on the computational model of deep learning frameworks. The former highlighted the diverse learning architectures, benchmarks and extended applications, while the latter focused on reviewing recent deep learning models applied to protein tasks such as predicting PPIs, protein functions and protein design. Chakraborty et al. [13] summarized some support vector machine (SVM)-based PPIs prediction computational models and challenges incurred in applying the SVM method. For feature extraction, Hu et al. [14] focused on introducing the algorithms used for predicting PPIs and different validation schemes and metrics to evaluate the predictive performance. These studies focused less on the quality of the dataset and did not delve into the PPNIs.

This review is structured as follows. We first outline protein-protein interaction architectures in Section 2. Section 2.1 discusses protein-related databases and datasets composition used in training and testing. Section 2.2 introduces diverse computational models, including sequence-based, structure-based, genomic-based, network-based and other computational methods. Next, we describe protein-protein non-interaction architectures in Section 3. Section 3.1 briefly explains negative sample constructed methods to obtain high-quality negative datasets. Section 3.2 analysis selection criteria for positive and negative examples and unbalanced dataset construction. Section 3.3 describes applications of real non-interacting protein pairs. Section 4 discusses the challenges and future works of computational methods for identifying PPIs and PPNIs.

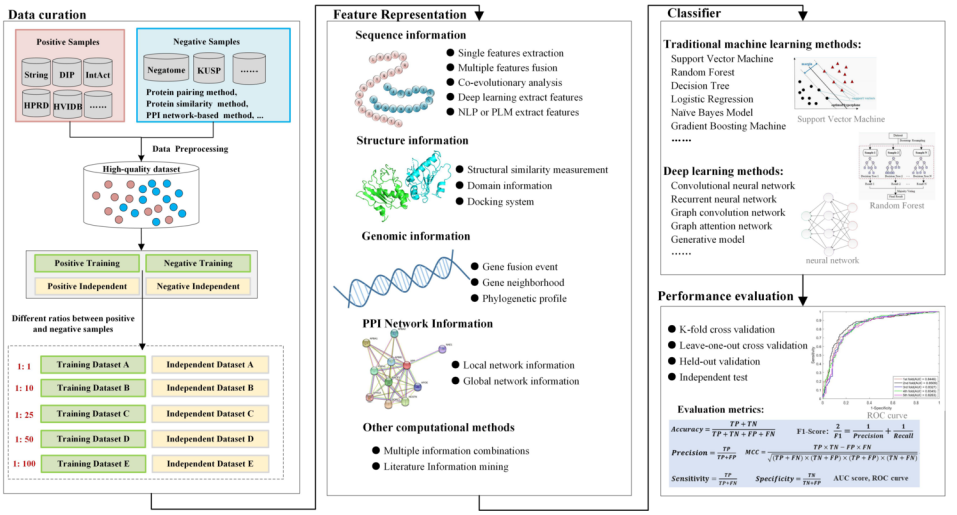

Figure 1: The overall framework of computational methods for PPIs prediction.

2. Protein-protein interaction

With the enrichment of data and the continuous improvement of machine learning algorithms, computational models have developed rapidly. Diverse protein data can be used as input data for downstream analysis of different protein tasks, such as primary sequence, protein structure, gene expression and network topology. The framework of the model evaluation study is presented in Figure 1. It shows a clear data processing process, followed by feature representation for diverse types of protein data, classifiers and performance evaluation. Previous studies have introduced the machine learning and deep learning algorithms used to predict PPIs in detail, as well as model evaluation, so this paper focuses on the data analysis and feature representation.

2.1. Dataset

2.1.1. Data source Many reliable biological data have been obtained with the development and maturity of experimental techniques and computational methods. Bioinformatics research collates and stores these biological data and establishes the database of the gene, protein sequence, protein structure, protein interaction, and other aspects. The details are

as follows: (i) Protein sequences. Protein can obtain sequence information from UniProt [15], SWISS-PROT [16], PIR [17], NCBI [18] and NRL3D [19] databases. (ii) Protein structures. Protein can collect structure information from PDB [20], SCOP [21], Pfam [22], InterDom [23] and 3did [24]. (iii) Gene ontology. Protein can derive gene ontology information from GO Database [25] and QuickGO [26]. (iv) Genomic information. Protein can obtain genomic information from MIPS mammalian protein-protein interaction database (MPPI) [27] and Candida Genome Database (CGD) [28]. (v) Protein-protein interactions. The databases include DIP [29], IntAct [30], BIND [31], MINT [32], BioGRID [33], HPRD [34], STRING [35], HVIDB [36], VirusMentha [37], HIPPIE [38], PRIN [39], DIPOS [40], PHISTO [41], VirusMINT [42], VirHostNet [43], etc. (vi) Protein-protein non-interactions. The databases include Negatome [44, 45], KUPS [46] and negative interactions from large-scale two-hybrid experiments [47].

2.1.2. Data composition This section summarizes some typical computational methods for PPIs prediction, focusing on the dataset species, the collecting methods of positive samples and negative samples, and the ratio of positive samples and negative samples (see Tables 1 and 2).

We analyze it from the following three aspects: (1) Species. Table 1 shows that computational methods for predicting intra-species PPIs are mostly trained and tested on *S. cerevisiae* and *H. sapiens* datasets, while Table 2 shows that the host of inter-species PPIs mainly considers human. (2) Sample collection. Positive samples are mainly obtained from DIP, IntAct and other PPIs databases, while negative samples are obtained by random pairing or pairing proteins with different subcellular locations. (3) Proportion of positive and negative samples. A balanced dataset is one in which the number of positive and negative samples is equal, i.e., the ratio is 1:1. The number of negative samples in an unbalanced dataset is usually K times the number of positive samples ($K = 2, 3, 5, 10, \ldots$). As indicated in Table 1, most researchers use balanced datasets for intra-species PPIs prediction, and some will additionally train and test on unbalanced datasets. Table 2 shows that inter-species PPIs are usually performed on unbalanced datasets.

2.2. Model

The essential components of proteins are amino acids. Amino acids have various physicochemical properties, making proteins receive different forces during the interaction process. In addition, the types of protein data are rich and diverse, making people propose many computational methods to

Table 1: Summary of computational models for Intra-species PPIs prediction.

Method	Data species	Positive samples	Negative samples	P: N	URL
CT [48]	*H. sapiens*	HPRD	randomly pairing proteins	1:1	N/A
AC+SVM [49]	*S. cerevisiae*	DIP	three strategies: 1. randomly pairing proteins; 2. pairing proteins with different subcellular localization; 3. artificial shuffling of protein sequences.	1:1	http://www.scucic.cn/Predict_PPI/index.htm
DPPI [50]	*H. sapiens, S. cerevisiae*	HIPPIE and DIP	randomly pairing proteins	1:10	https://github.com/hashemifar/DPPI/
	S. cerevisiae core subset	DIP	pairing proteins with different subcellular localization	1:1	
	A large dataset consisting of 11 different species: *H. sapiens*, *S. cerevisiae*, *S. pombe*, *M. musculus*, *D.melanogaster*, *C. elegans*, *A. thaliana*, *B. subtilis*, *B. taurus*, *E. coli* and *R. norvegicus*	HINT	randomly pairing proteins	1:10	
NVDT [51]	*H. sapiens, M. musculus*	DIP	Negatome	1:1	https://github.com/Zhaonan99/NVDT
DeepTrio [52]	*S. cerevisiae*, *H. pylori*, *D. melanogaster* and *M. musculus*	DIP	pairing proteins with different subcellular localization	1:1	https://github.com/huxiaoti/deeptrio.git
	S. cerevisiae, *H. sapiens*	BioGRID and DIP	artificial shuffling of protein sequences	1:2	
	S. cerevisiae	DIP	pairing proteins with different subcellular localization	1:1	
Struct2Graph [53]	*S. cerevisiae*, *H. sapiens*, *E. coli*, *C. elegans* and *S. aureus*	IntAct and STRING	large-scale two-hybrid experiments	1:1, 1:2, 1:3, 1:5, 1:10	https://github.com/baranwa2/Struct2Graph
TAGPPI [54]	Intra-species dataset: *E. coli*, *S. cerevisiae*, *C. elegans* and *D. melanogaster*	DIP	randomly pairing proteins	1:1	https://github.com/xzenglab/TAGPPI
	Multi-species dataset: *E. coli*, *C. elegans* and *D. melanogaster*	DIP	randomly pairing proteins	1:1	
	Multi-class dataset: *H. sapiens*	STRING	N/A	75875 samples	

Note: *Saccharomyces cerevisiae* (*S. cerevisiae*), *Homo sapiens* (*H. sapiens*), *Escherichia coli* (*E. coli*), *Caenorhabditis elegans* (*C. elegans*), *Staphylococcus aureus* (*S. aureus*), *Drosophila melanogaster* (*D. melanogaster*), *Helicobacter pylori* (*H. pylori*), *Mus musculus* (*M. musculus*), *Schizosaccharomyces Pombe* (*S. pombe*), *Arabidopsis thaliana* (*A. thaliana*), *Bacillus Subtilis* (*B. subtilis*), *Bos taurus* (*B. taurus*) and *Rattus norvegicus* (*R. norvegicus*).

Table 2: Summary of computational models for Inter-species PPIs prediction.

Method	Data species	Positive samples	Negative samples	P: N	URL
(Stefan Wuchtyet, 2011) [55]	Parasite-Human	MINT, IntAct, Reactome and HPRD	randomly pairing proteins	1:1	N/A
(Kshirsagar et al., 2013) [56]	Bacteria-Human	PHISTO	randomly pairing proteins	1:100	http://www.cs.cmu.edu/ mkshirsa/ismb2013_paper320.html
	Human and virus	VirusMINT	randomly pairing proteins	1:1	
Denovo [57]	Human and virus	VirusMentha	randomly pairing proteins	1:1	https://bioinformatics.cs.vt.edu/ alzahraa/denovo
doc2vec+RF [58]	Human and virus	HPIDB	dissimilarity negative sampling method	1:10	http://zzdlab.com/InterSPPI/
LSTM-PHV [59]	Human and virus	HPIDB and IntAct	dissimilarity negative sampling method	1:10	http://kurata35.bio.kyutech.ac.jp/LSTM-PHV
	Human-SARS-CoV-2	BioGRID	dissimilarity negative sampling method	1:10	
(Yang et al., 2021) [60]	Human-HIV, Human-Herpes, Human-Papilloma, Human-Influenza, Human-Hepatitis, Human-Dengue, Human-Zika, Human-SARS-CoV-2	HPIDB, VirHostNet, VirusMentha, PHISTO and PDB	dissimilarity negative sampling method	1:10	https://github.com/XiaodiYangCAU/TransPPI/

predict whether proteins interact with each other from multiple perspectives. The computational methods mainly include four different categories: sequence-based methods, structure-based methods, genomic-based methods and network-based methods. Other computational methods are based on multiple protein information or research in the literature.

2.2.1. Sequence-based computational methods The amino acid arrangement and combination determine the primary structure of proteins. Protein sequences can be inferred from sequenced genomes, the most abundant protein data available. Computational methods based on protein sequence usually vectorize the sequence of the collected dataset and then use appropriate machine learning algorithms to train and predict. Due to the easy availability of protein sequences and the fact that sequence-based schemes do not require prior knowledge, various sequence-based computational models are favored by researchers [61, 62].

1. Single feature extraction In the early development, sequence-based computational models focused on combining the single feature of sequences, like sequence similarity, with the learning ability of traditional classifiers to perform prediction tasks. For instance, Bock and Gough [63] utilized amino acid-associated physicochemical properties to extract sequence feature vectors. The physicochemical properties included charge, hydrophobicity and surface tension. Then, they systematically trained SVM with different kernel functions to recognize interactions. This method provided a new attempt at analyzing the interaction between proteins only using the primary structure of proteins. However, it did not consider the local environment of amino acids, which resulted in the predicted performance being unreliable and robust. Shen et al. [48] expressed the primary protein sequence with the conjoint triad (CT) and trained it with SVM to predict PPIs, in which CT accounted for interactions between adjacent or close residues in the amino acid sequence. To reduce dimension disaster, 20 amino acids were divided into seven groups according to the dipole scale and volume scale. CT regarded the target amino acid and its two proximate amino acids as one unit to statistic their overall electrostatic and hydrophobic properties. However, the convoluted folding of proteins allows residues to interact not only at short distances but also at long distances. Guo et al. [49] characterized the protein primary sequence with auto covariance (AC) and trained with SVM to identify PPIs, in which AC adequately considered the interactions between long-range residues. Seven different physicochemical properties of amino acids

were considered, including hydrophobicity, hydrophilicity, side chain volume, polarity, polarizability, solvent-accessible surface area and net charge index of side chains. Yang et al. [64] represented a computational method that extracted sequence features by utilizing local descriptors (LD, including composition, transition, and distribution), which considered the effect of discontinuous amino acids but ignored global information. Zhou et al. [65] utilized the codon pair frequency difference to identify PPIs and showed comparable performance to other sequence-based methods. Najafabadi et al. [66] exploited a Naive Bayesian classifier to combine the relative codon frequency differences to predict PPIs, which performed well on *S. cerevisiae*, *E. coli* and *Plasmodium falciparum*. This approach demonstrated that the codon usage of functional and physically connected proteins in organisms contained rich sequence information.

2. Multiple features fusion Since single-feature extraction methods are challenging to characterize the protein sequence information accurately, researchers have proposed many computational methods integrating various features for predicting PPIs [67, 68]. For instance, Du et al. [69] proposed DeepPPI combining multiple sequence features with deep neural network (DNN) for PPIs prediction. This method extracted features based on common protein descriptors, including amino acid composition, dipeptide composition, LD, quasi-sequence-order descriptors, and amphiphilic pseudo amino acid composition. In addition, DNN can effectively learn the representations of protein pairs rather than directly connecting the feature vectors of two proteins. Ahmed et al. developed [70] a neural network model predicting host-pathogen PPIs based on a combination of features, including amino acid quadruplets, pairwise sequence similarity and human interactome properties. Zhang et al. [71] presented EnsDNN to identify PPIs that combined multiple sequence features and DNN. This model used AC, LD and multi-scale continuous and discontinuous local descriptors to characterize the interaction between sequentially distant but spatially close residues. Each descriptor was trained with a specific configured DNN and then integrated for prediction. Chen et al. [72] proposed StackPPI, a predictive framework for predicting PPIs that combined multiple sequence features and an ensemble classifier. This method used pseudo amino acid composition, Moreau-Broto, Moran and Geary autocorrelation descriptor, position-specific scoring matrix, Bi-gram position-specific scoring matrix and LD to encode biologically relevant sequence information. Multi-information fusion can more fully characterize sequence features, thus improving prediction

accuracy. StackPPI utilized XGBoost for dimensionality reduction and constructed a stacker ensemble classifier of random forest (RF), randomized trees and logistic regression. This model reduced generalization errors and improved prediction accuracy. Zhao et al. [51] developed a gene sequence-based method, NVDT, which counted the number, average position and second normalized central moments of nucleotides, and the frequencies of dinucleotides and triplet nucleotides. NVDT not only combined the advantages of local and global protein information but also had high computational speed with low dimensions, making it a robust and efficient method for PPIs and PPNIs prediction. These feature extraction methods generally summarized the physicochemical properties of amino acids, location distribution and other statistical characteristics, but the massive dimension of feature vectors dramatically increased the computational complexity.

3. Co-evolutionary analysis The studies suggest that co-evolutionary proteins interact more easily with each other. Hu et al. [73] focused on extracting co-evolutionary features from sequence information, which were essential for protein function. They proposed CoFex, a feature extraction approach that took into account the co-evolutionary position of amino acids. Yin et al. [74] used biochemical properties of amino acids to characterize protein sequences and conducted co-evolutionary analysis combined with Fourier transform to predict PPIs. Position Specific Scoring Matrix (PSSM) is a vital scoring matrix generated by sequence similarity comparison, which contains not only the evolutionary information between sequences but also the conserved and mutated information of amino acids. It has been successfully used in computational biologies, such as protein secondary structure prediction [75], protein binding site prediction [76] and disorder region prediction [77]. Yang et al. [60] designed a deep learning framework to predict human–virus PPIs that combined evolutionary sequence profile features with a Siamese-based multi-scale convolutional neural network (CNN) architecture and a multi-layer perceptron (MLP). This model represented interacting proteins by PSSM and introduced two types of transfer learning methods ("frozen" type and "fine-tuning" type). These two transfer learning methods allowed training on a source human-virus domain. They used the data of the target domain to retrain the CNN layer, significantly improving the cross-viral prediction performance.

4. Deep learning extract features In the last few years, deep learning-based approaches have massively impacted the field of protein bioinformatics. Some computational methods do not use proteins' prior information, such

as the physicochemical properties, but directly use deep learning algorithms to extract protein sequence features. Hashemifar et al. [50] reported a deep learning framework, DPPI, to identify PPIs based on the primary protein sequence. DPPI exploited a Siamese-like CNN architecture combined with a random projection module and data augmentation. This framework could efficiently handle large amounts of training data, fine-tune parameters to adapt to different tasks, and robustly capture complex and non-linear relationships in PPIs. Hu et al. [52] proposed a deep learning model, DeepTrio, that used mask multiple parallel CNN for PPIs prediction. This model captured the multi-scale contextual information of protein sequences. Chen et al. [78] developed an end-to-end framework, PIPR, which employed a Siamese architecture based on a deep residual recurrent convolutional neural network to effectively capture the mutual influence of protein pairs. The model performed well on interaction type and binding affinity estimation prediction tasks and can be generalized to different PPIs prediction tasks without the need for predefined features. However, the bidirectional gated recurrent unit of this algorithm suffered from slow convergence speed and low learning efficiency.

5. NLP or PLM extract features In recent years, the computational methods of extracting protein sequence features based on natural language processing (NLP) and Protein language model (PLM) have become increasingly popular. These methods require training on a large number of protein sequences to extract informative features of protein sequences. They can not only sufficiently consider the semantic information in the entire sequence, such as the order of residues, but also mine massive potential information of unlabeled protein. For instance, Yang et al. [58] combined a doc2vec embedding method with RF to predict human-virus PPIs. The doc2vec embedding method captured the semantic information of residues in the whole sequence as much as possible, thus representing the protein sequence as rich feature vectors of low dimensionality. Tsukiyama et al. [59] reported LSTM-PHV, an approach that combined the long short-term memory (LSTM) model with word2vec to predict human-virus PPIs. The word2vec embedding method utilized the amino acid sequence context as a word to encode sequence features, which effectively improved the prediction performance. Madan et al. [79] developed an approach that used the ProtBERT [80] deep sequence embedding method and Siamese neural network to detect PPIs. These methods avoid the error of encoding protein sequences by manual features and can potentially capture more comprehensive protein sequence information.

2.2.2. Structure-based computational methods The protein structure is closely related to its function. It contains essential protein biological information, which can be used to guide the prediction of PPIs. Protein interaction studies using structure information are more reliable and accurate at the atomic level.

1. Structural similarity measurement Protein complexes with known three-dimensional structures provide the best context, containing reliable information about protein interactions. Song et al. [54] proposed an end-to-end method, TAGPPI, which utilized protein structural information for PPIs prediction. This method extracted multi-dimensional features by employing convolutional structure on amino acid sequences and graph learning method on contact maps constructed from AlphaFold [61], so it obtained much spatial structure information. Baranwal et al. [53] proposed Struct2Graph, a structural approach was predicting PPIs solely from protein 3D structures. Struct2Graph used a graph convolutional network-based representation of a protein globule rather than descriptors like solvent accessible surface area. Doolittle et al. [81] used protein structural similarity to identify the interactions between HIV-1 and human, based on the assumption that human proteins with highly similar structures may have similar interacting pathogen partners. This approach was applicable to any host-pathogen system with known protein structures.

2. Domain information The domain is a special amino acid sequence with conserved protein function, which is the structural and functional unit of protein. The domain has a relatively independent and stable 3D-dimensional conformation, and its function is relatively conservative. Abnormality of the domain may lead to dysfunction and even induce disease [82, 83]. Behind PPIs are often physical interactions between protein domains to perform significant functions. Therefore, predicting the interaction between proteins can be transformed into predicting whether domains interact. Deng et al. [84] applied the Maximum Likelihood Estimation method, which estimated the interactions probabilities of domain pairs, to validate domain–domain interactions consistent with the observed PPIs. This probabilistic model was robust in handling experimental errors and allowed for incorporating various PPIs data, such as from different organisms. However, the model assumed the independence of the domain-domain interaction, which was not completely consistent with the fact. Moreover, the PFAM domains are not necessarily subunits with a special structure essential to the protein interaction. Ma et al. [85] presented a computational method based on interolog and the

domain-based method to predict blast fungus-rice PPIs. Then SVM and the enrichment of pathogenic proteins were further used to identify potential PPIs.

3. Docking system Traditional protein-protein docking methods play a vital role in sampling the conformational space of protein complexes. High-precision docking methods have been developed and continuously improved, including HDOCK [86], ZDOCK [87], LzerD [88], HADDOCK [89], ClusPro [90], and LightDock [91]. Mirabello et al. [92] proposed a fully automated pipeline, InterPred, to predict PPIs using structural modeling combined with massive structural comparisons and molecular docking. A vital component of the method was that the RF classifier integrated several structural features to distinguish PPIs. The time-consuming steps of structural template searching and docking decreased the efficiency of InterPred.

The methods based on protein structure similarity and molecular docking simulation are limited by the number of protein structures analyzed experimentally. While AlphaFold can obtain massive protein structures with atomic accuracy, it is limited by the length of the protein sequence. Most domain–based methods use correlation algorithms and maximum likelihood estimation to obtain the interaction probability of a single domain. Nevertheless, they only consider the quantitative relationship of a single domain, ignoring the internal domain properties. Therefore, the structure-based approach needs to be further improved.

2.2.3. Genomic-based computational methods The arrangement of genes in the genome is regular, and genes with similar functions tend to be arranged more closely. Therefore, studying the location of protein-coding genes in the genome can help infer functional similarity and interaction between proteins. Existing computational methods mainly involve the following three categories: gene fusion event, gene neighborhood and phylogenetic profile.

1. Gene fusion event Gene fusion occurs when two independent proteins in one species fuse into a protein component or a polypeptide chain in another species. A gene fusion event between two proteins during the evolution of a species is thought to be an interaction between these two proteins. Marcotte et al. [93] developed a computational approach to identifying PPIs from complete genome sequences based on observing gene fusion events. Enright et al. [94] proposed a computational method detecting gene fusion event complete genomes from sequence comparison. Gene fusion-based methods can

only predict functional associations between proteins or direct interactions that have fused but cannot predict proteins without fusion events through genome sequencing analysis.

2. Gene neighborhood Gene neighborhood speculates that there will be functional interactions between the gene products of an operon [95, 96]. Therefore, the interactions between proteins can be predicted based on the adjacent conservatism of genes in different biological genomes. However, gene neighborhood-based methods only work for microorganisms with simple structures in early evolution. It is not applicable to eukaryotes with complex structures, so it cannot be widely used.

3. Phylogenetic profile If two genes have identical or similar phylogenetic profiles, it can be inferred that they are functionally related and likely to have functional interactions. Pellegriniet et al. [97] constructed phylogenetic profiles of three proteins through fully sequenced bacterial genomes, which showed that phylogenetic profiles could cluster functionally related proteins. However, the limitation of the phylogenetic profile-based methods is that it is impossible to determine whether the function-related proteins are in "physical" direct contact. Its accuracy depends on the number of sequenced genomes and the reliability of the phylogenetic profile construction.

2.2.4. Network-based computational methods Different PPIs databases store different protein interactions and related information. Still, most data mainly focuses on protein interactions and records relevant information, such as obtained experimental methods, interaction types, etc. The proteins in the PPIs database can be connected and mapped into one or several PPI networks. A PPI network is a simple undirected graph, where the nodes are proteins, and the edges indicate an interaction between the two proteins. As the coverage of PPI networks has increased, many network-based computational models have been developed to predict missing PPIs.

1. Local network information Local network information refers to the proximity information between a target protein and its nearest neighbor protein. Li et al. [98] designed a new method, LAC, to identify essential proteins by deeply considering the relationship between the target protein and their neighbors. Therefore, the local network information contains rich network features of the protein in the PPI network. Studies have shown that if one of the two proteins is similar to the interacting partner of the other, the

two proteins have a higher probability of interacting. L3 [99] defined a degree standardized score that depends on a network path of length three, and its performance was significantly superior to all existing link prediction methods. However, L3 is not suitable for predicting PPIs between far-apart proteins that do not have a common neighbor.

2. Global network information In the PPI network, the global network information refers to the topological information considering the whole protein interaction network. Global network information can provide more comprehensive evidence for verifying PPIs. For example, Lei et al. [100] considered global network structure and developed a novel network topology-based algorithm to reduce the noise present in PPI networks. Two proteins sharing some higher-order topological similarities, as measured by a new random-walk-based algorithm, are likely to interact.

2.2.5. Other computational methods

1. Multiple information combinations Different information sources have advantages and disadvantages, and they can complement each other to characterize proteins better. Zhang et al. [101] proposed the PrePPI model, which integrated protein structural information and other biological functional evidence to detect PPIs. Wang et al. [102] presented DeepViral, a deep learning-based method jointly learning from protein sequences, phenotype functions and taxonomy features to predict potential protein interactions between viruses and human hosts. The scarcity of training interaction data between viruses and hosts other than humans limited the generalization performance of the model. Lei et al. [103] combined manifold embedding with multiple information integration to identify PPIs, which embedded PPI networks into low-dimensional metric space based on manifold learning theory. Liu et al. [104] utilized graph convolutional networks (GCNs) to learn protein position information in the PPI networks graph and capture the graph structure information. This method combined sequence information with position information to represent proteins, which improved the prediction performance of PPIs.

2. Literature information mining The literature database covers much information about protein interactions, such as subcellular localization and biological function. Therefore, some computational methods for PPIs prediction are based on literature mining. Huang et al. [105] used dynamic programming and matching algorithms to detect PPIs from literature, with a

recall rate of 80.00% and a precision rate of 80.50%. Hao et al. [106] presented an approach to identify PPIs from literature, which can automatically discover and optimize English expression related to protein interactions. Jang et al. [107] proposed a PubMed-abstract-based PPIs validation method that can automatically query and extract much interaction information for two given proteins. The research found that 67.37% of interactions in the DIP were from PubMed abstracts, and 87.37% were from PubMed full text.

3. Protein-protein non-interaction

3.1. Negative samples construct

Previous studies have proved that dataset quality significantly impacts the robustness of the computational models [108]. Positive samples can be easily collected from PPIs databases, where many experimentally verified interacting protein pairs are stored. These samples are verified to have high reliability by biological experiments. A high-quality positive dataset can be obtained by filtering the interacting protein pairs depending on the sequence similarity and distribution. Since few experiments have established PPNIs databases of non-interacting protein pairs, finding reliable samples of non-interacting protein pairs is more of a challenge. The difficulty in constructing a high-quality negative dataset is that there is no "gold standard" for planning negative samples in reality. Therefore, negative samples must be collected with care, or it may adversely affect the predicted accuracy. Existing computational models rely on the following negative sample construction methods.

3.1.1. Protein pairing

1. Random Pairing The proteins in the positive dataset were randomly paired to generate candidate non-interacting protein pairs [46, 109], and then the candidate pairs in the positive dataset were removed to obtain the final negative samples. Many authors use this simple approach to construct a negative dataset [50, 78, 110, 111]. For example, Chen et al. [46] constructed a class of negative datasets to develop KUPS database using uniform random pairs, which is generally considered less biased than selection methods based on molecular processes. However, the false negative probability of the negative samples constituted by this mechanism is relatively high, which leads to problems such as the prediction approach will learn the pattern of missing values. The estimates show that the number of interacting protein pairs in the randomly pairing dataset can be unacceptably high when considering the specific biological context [47].

2. Subcellular Location Non-interacting protein pairs were generated by pairing proteins with different subcellular localization and deleted protein pairs appearing in the positive dataset [109]. Previous studies have proven that proteins from different subcellular locations are unlikely to interact, so most interaction prediction studies use this common method to define negative samples for training models [112]. The annotations of protein subcellular location are available from the UniProt [15] and Swiss-Prot [16] databases. The limited distribution of negative samples generated by this method makes the PPIs prediction task easier, resulting in a biased estimation of accuracy [10].

3.1.2. Protein similarity

1. Reversed order One of the interacting protein pairs reverses its order of amino acids and forms a non-interacting protein pair with the remaining protein. Many researchers have constructed negative samples through this method [109]. It has been proven that if one sequence of an interacting protein pair is shuffled, the possibility of the interaction between the remaining protein and the new protein can be considered negligible [113]. However, artificially altered amino acid sequences do not exist in reality, so this method cannot reflect the natural protein interaction sequences.

2. Similarity and degree Zhang et al. [114] have proposed that for an interacting protein pair (protein i and protein j) that experiments have verified, the larger the sequence dissimilar between protein i and protein k, the lower the probability that protein j interacted with protein k. A new method to construct negative samples, NIP-SS, was proposed based on the above assumptions. The steps were: Firstly, the sequence similarities between any two proteins in the positive dataset were counted and sorted in ascending order. Then, select the top-m protein pairs with low similarities from the sorted dataset as non-interacting protein pairs. Considering that the negative samples obtained in this way may be concentrated in a few proteins, it was not conducive to computational model construction and evaluation. Therefore, the negative samples were further adjusted by maintaining the similarity of the protein degree distribution in the positive dataset, while the generated non-interacting pairs appearing in the positive dataset were eliminated. Selecting negative samples with low sequence similarity to positive samples obviously reduced the probability of false negative. Still, it also made the task much easier than it is, resulting in over-optimistic estimates of accuracy. Eid et al. [57] introduced a method based on sequence

dissimilarity to generate non-interacting protein pairs for virus-host PPIs prediction, which reduced the noise in the negative dataset. This method was based on the hypothesis that if the sequences of two viral proteins are similar, a human protein interacting with one of them would have a high probability of pairing with the other virus as a positive sample. Conversely, negative examples can be found. All-versus-all global alignment bit-scores of the viral proteins were first calculated by the dissimilarity-random-sampling algorithm and normalized. Then the impossible negative cases were excluded according to a dissimilarity threshold, and the remaining negative interaction was randomly sampled.

3. Semantic similarity Chen et al. [46] defined functionally dissimilar pairs as non-interacting protein pairs in the KUPS database. KUPS first calculated the semantic similarities [115, 116] between any two proteins and ranked them. Based on the assumption that the most dissimilar annotation pairs have the lowest similarity score, potential non-interacting protein pairs are selected.

3.1.3. PPI network-based

1. Random walk Previous studies have proved that those interacting proteins are likely to share similar functions. So for the target protein in the PPI network, the neighborhood (or direct interacting) protein with the shortest path of 1 is more likely to share similar functions than the neighborhood protein with the shortest path of 2. Because the latter cannot interact directly with the target protein, the interaction is more likely to be mediated by another protein. In other words, the probability that two proteins share similar functions gradually decreases as the shortest path increases. Therefore, considering the small probability that two proteins share similar functions at a sufficiently large shortest path, they are regarded as a non-interacting protein pair. Zhang et al. [114] proposed NIP-RW based on random walk in PPI network to distinguish high-confidence non-interacting protein pairs.

2. Low degree Studies on the viral protein pathway showed that viruses are more likely to target higher-degree human proteins than lower-degree human proteins [117]. Dey et al. [118] have collected negative samples generated from virus protein and human proteins with a low degree in the human PPI network [34]. First, the degree of each human protein in the HPRD database was calculated and sorted. Then, one virus protein and one low-degree human are used to form non-interacting protein pairs. However, this negative sample construction method is only suitable for predicting virus-host PPIs.

3.1.4. Experiments evidence

1. Two-hybrid experiments Trabuco et al. [47] utilized a dataset from two-hybrid experiments to infer non-interacting protein pairs that were not observed experimentally. Negative samples in Struct2Graph [53] were retrieved from Trabuco et al. and selected pairs that were not involved in any interaction of STRING and IntAct Databases. However, this method was only suitable for two-hybrid experiments to detect protein pairs, not other experimental platforms.

2. Literature evidence Smialowski et al. [44] constructed the Negatome database 1.0 containing non-interacting protein pairs. The process was based on a simple keyword search in the PubMed database, especially noting a large amount of non-interaction information in many figures and tables. Blohm et al. [45] proposed the second version of this database, Negatome Database 2.0. They used an advanced text-mining process to guide the manual annotation process, which focused on the entire corpus of PubMed abstracts and PMC full-text articles. This method has shifted from the time-consuming analysis of figures and tables to a more high-throughput automated approach.

3. Protein complex-based KUPS database [46] defined non-interacting protein pairs not by spatial distance but by cellular component annotations without overlap. The Negatome database [44, 45] showed that some protein pairs in protein complexes with more than three chains do not interact directly with each other while they are in the immediate vicinity of the protein complex environment. So a new set of non-interacting protein pairs can be obtained by selecting complex chain pairs with a C_β-C_β (C_α-C_α for glycine) distance larger than 8Å.

4. Non-interacting domain pairs Chen et al. [46] collected a class of negative datasets based on non-interacting domain pairs defined in the Negatome database. Nevertheless, this approach assumed that non-interacting protein pairs have been obtained.

3.1.5. Database for negative samples High-quality negative samples are critical to capture protein-protein interacting and non-interacting information for protein tasks. This section describes two important PPNIs databases.

1. Negatome [44, 45] The non-interacting protein pairs in the Negatome database are collected from manual literature curation and analysis of 3D protein complexes. Many studies currently use real non-interacting protein pairs in the Negatome database to construct negative samples. Bryant et al. [119] applied AlphaFold2 and optimized multi-sequence alignment to predict heterodimeric protein complexes using negative samples from the Negatome database. Das et al. [120] used a negative dataset from the Negatome database to study the knowledge of protein-protein interface properties combined with SVM and delineate native-like protein complexes from non-native protein complexes. Therefore, real non-interacting protein pairs in the Negatome database can help to generalize the training model and improve the accuracy of task results. However, it also has some disadvantages: (i) The number of negative samples in the Negatome database is small and needs to be expanded. (ii) The Negatome database mainly collects non-interacting protein pairs that do not interact in a direct physical manner. (iii) This database relies on results from experiments, but the overlap between different experimental datasets is weak. And the fact that two proteins have not been reported to interact experimentally does not mean that they do not actually interact in cells. The Negatome database is freely available through the website http://mips.helmholtz-muenchen.de/proj/ppi/negatome.

2. KUPS [46] The KUPS database contains datasets of non-interacting protein pairs using four negative sample construction methods to alleviate the biased estimation problems, including uniform random pairs, functionally dissimilar pairs, spatially separated pairs and non-interacting domains. In addition, KUPS created two benchmark datasets: one with balanced interacting protein pairs and non-interacting protein pairs and the other with unbalanced interacting protein pairs and non-interacting protein pairs. These negative sample construction methods still have scope for improvements, such as considering some physical and chemical properties, including solvent accessibility and hydrophobicity. In addition, the established method can be extended to proteome and other organisms. The KUPS is freely available through the website: http://www.ittc.ku.edu/chenlab/.

3.2. Unbalanced dataset

To obtain a high-quality dataset, positive samples directly collected from PPIs databases must be preprocessed. It mainly includes the following four points: (i) Length. Protein pairs were generally removed if the sequence

length of one protein was less than 50 amino acids [69, 72]. Different experiments have different requirements for protein length. A few experiments require the protein length to be a minimum of 30 and a maximum of 5000 amino acids [58], while others even require a protein length of 30–1000 [59] or 150–1000 [52]. Protein sequences can be retrieved from UniProt/Swiss-Prot. (ii) Nonredundant. To avoid the classifier will possibly be biased to homologous sequence pairs, protein pairs were deleted if two proteins share a sequence identity greater than or equal to 30% [58] (or 40% [69, 72]) by cluster analysis of the CD-HIT [121] program. (iii) MS experiments. Protein pairs were collected if the experimental interactions were from two or more high-throughput MS experiments [60]. Each PPIs database uses different publications and ontologies to report protein interactions. Consequently, PPIs reported by each database are different, with only up to a 75% consistency between all PPIs databases [122]. (iv) Interactions of inter-species with a certain confidence. Protein pairs with an MI score of below 0.3 were deleted. MI score is a confidence score for protein interactions, which can be obtained from IntAct and VirHostNet. Different methods for obtaining negative samples detail the specific conditions for collecting non-interacting protein pairs and then randomly selected from the collected negative samples according to the required number of samples.

Protein-protein interaction and non-interaction prediction is a binary problem in which positive samples are collected from PPI databases, while negative samples are almost all other protein pairs. The number of negative samples collected is much higher than that of positive samples, which is vast and extremely unbalanced data. For intra-species PPIs prediction, balanced datasets are usually used in previous studies. A common practice is forming a balanced dataset by randomly sampling the same number of negative samples from the original dataset as the positive ones. For example, Shen et al. [48] conducted training and testing on a balanced dataset of 16443 positive samples and 16443 negative samples for PPIs prediction. Guo et al. [49] utilized a balanced dataset of 5545 positive samples and 5545 negative samples to identify PPIs. Zhao et al. [51] trained the prediction model using a balanced dataset comprising an equal number of interacting and non-interacting protein pairs. Most of these approaches have achieved satisfactory prediction performance on balanced datasets. However, the construction of this dataset differs from that of the cell environment. Therefore, such predictive performance may not be achieved under natural conditions. Recently, researchers have begun to conduct simultaneous experiments on unbalanced datasets. For instance, Baranwal et al. [53] used two kinds of datasets: (i) Balanced dataset: Close to 1:1 (4698 positive samples and 5036

negative samples); (ii) Unbalanced dataset: Design different ratios of positive to negative pairs, including 1:2 (2518 positive samples and 5036 negative samples), 1:3 (1679 positive samples and 5036 negative samples), and 1:10 (504 positive samples and 5036 negative samples). Unbalanced datasets are usually used for inter-species PPIs prediction, and the ratio of positive samples and negative samples is mostly 1:10. Further, the selected samples were divided into a training set (80%) and an independent test set (20%), which were used for model training and performance evaluation respectively. Several training sets and independent testing sets can be randomly constructed to reduce sampling bias caused by sample division.

3.3. Applications of real non-interacting pairs

The artificially constructed negative samples can not reflect the actual prediction situation. More and more models have begun testing on the pairs of real non-interacting proteins to obtain a prediction model with solid generalization ability. Zhang et al. [114] constructed six independent datasets containing only interacting protein pairs (five datasets) and non-interacting protein pairs (one dataset). The negative samples dataset was Mammalian collected from Negatome 2.0 (1937 non-interacting pairs). The results showed that the accuracy of the Mammalian dataset using the NIP-SS and NIP-RW strategies (the proposed negative examples generation methods) was 3.36 and 3.98 times that of pairing proteins with different subcellular localization, respectively. It is inferred that negative samples generated by subcellular localization may have a bias in the predictive performance during model training and testing. Zhao et al. [51] collected two types of datasets: real dataset and constructed dataset. The real dataset consisted of real interacting protein pairs and real non-interacting protein pairs. The positive samples were downloaded from the public DIP database and the negative samples were derived from the Negatome Database 2.0. Finally, 2434 protein pairs for *H. sapiens* and 694 protein pairs for *M. musculus* were obtained, with the same number of positive and negative samples. For PPNIs, the proposed method obtained accuracies of 86.23% for *H. sapiens* and 85.34% for *M. musculus* real datasets. It also performed well on three non-interaction networks consisting of non-interacting protein pairs, including one-core network, multiple-core network and crossing network. To test whether the lack of core protein non-interaction information leads to low accuracy, they added existing non-interaction knowledge for 10%, 30%, and 40% of core proteins in the multiple-core network. The accuracy obtained could be increased from 80.49% to 85.37%, 92.68% and 96.34%, respectively. Patrick Bryant

et al. [119] applied Alphafold2 and optimized multi-sequence alignment to predict heterodimeric protein complexes. The non-interacting proteins used were the negative samples collected in the Negatome database. Real non-interacting protein pairs can be used to evaluate PPIs/PPNIs prediction methods. Especially the non-interacting PDB pairs can help predict the 3D structure of protein complexes and improve the training of interaction and non-interaction prediction algorithms.

4. Challenge and future work

PPIs and PPNIs are critical to understanding the mechanisms of most biological processes. Since laboratory-based approaches have the disadvantages of being time-consuming and labor-intensive, various computational models have been developed to supplement them. This paper summarizes the effective prediction models of PPIs and PPNIs, focusing on data analysis and feature representation. However, these models still have many challenges and future work.

1. Data quality At present, many computational methods for predicting the interaction between proteins have achieved very high accuracy. But is such high accuracy reliable? Continue to test/doubt it. To improve the quality of the dataset, we can proceed from three aspects: (i) Negative sample construction. The common negative sample construction methods have the problems of uneven distribution, which causes the deviation of prediction results. New effective negative sample construction methods need to be further developed. (ii) Data preprocessing. Both the positive samples collected from the PPIs databases and the negative samples constructed by the existing sampling methods are redundant and imperfect. The screening criteria should be strictly controlled for filtering. (iii) The ratio of positive and negative samples. In computational experiments, the training and testing of the model should be carried out on different proportions of balanced and unbalanced datasets. In addition, multiple datasets can be set for the computational evaluation, and the average value is taken as the final prediction result.

2. Species Most of the existing computational methods focus on *S. cerevisiae* and *H. sapiens*, which involve a small range of species. So it is difficult to ensure that the high prediction performance can be maintained on other species datasets. At the same time, the intra-species PPIs prediction is mostly a eukaryotic dataset, which is difficult to extend to bacteria and

other research problems. Most of the hosts for inter-species PPIs prediction are human, which makes the generalization performance of models to other species not high.

3. Computational method In the past, with the emergence of the induced fit theory, the primary sequence was the most accessible data to obtain protein information, and sequence-based methods have become the most widely used method. But other data types, such as protein structure, are not easy to obtain on a large scale. Now advances in artificial intelligence have made it much more efficient at capturing large amounts of diverse protein data. Different protein information can complement each other to better characterize proteins. The computational methods have gradually changed from relying on single protein information to combining multiple protein information for prediction. Although integrating these different data types may provide additional evidence for PPIs predictions, their relationship remains to be studied in depth. Effectively integrating various data types with machine learning technology is crucial to predict PPIs successfully.

4. Genetic, dynamic and disordered PPIs PPIs can be classified as physical interaction and genetic interaction. However, readily available PPIs and PPNI databases have only shown whether there are physical interactions but are not sure free of any genetic interactions. In addition, PPIs are dynamic in cells, but the scientific quest to capture PPIs dynamics under physiological and disease conditions are limited [14]. Few computational models have been developed to predict dynamic PPIs, which should be further improved in future work. Intrinsically disordered protein (IDP) is a kind of protein that lacks a stable three-dimensional structure in its natural state. Natural disordered proteins cover a series of proteins in various states, from completely without fixed ordered structures to partial structures. Research emphasizes that the role of disordered protein interaction as a critical coordinator of gene expression and other complex biological functions is underestimated [123]. These findings will also help better understand other potential diseases, such as cancer and viral infection. Therefore, PPIs of disordered proteins are also an important research issue.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (22XNH158).

References

[1] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki, *A comprehensive two-hybrid analysis to explore the yeast protein interactome*. Proceedings of the National Academy of Sciences of the United States of America, **98**, 4569–4574 (2001).

[2] F. Pazos and A. Valencia, *In silico two-hybrid system for the selection of physically interacting protein pairs*. Proteins, **47**, 219–227 (2002).

[3] A. I. Nesvizhskii, A. Keller, E. Kolker, R. Aebersold, *A statistical model for identifying proteins by tandem mass spectrometry*. Analytical Chemistry, **75**, 4646–4658 (2003).

[4] Y. Ho et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry*. Nature, **415**, 180–183 (2002).

[5] H. Zhu et al., *Global analysis of protein activities using proteome chips*. Science, **293**, 2101–2105 (2001).

[6] O. Puig 1, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm and B. Séraphin, *The tandem affinity purification (TAP) method: a general procedure of protein complex purification*. Methods, **24**, 218–229 (2001).

[7] A. C. Gavin et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes*. Nature, **415**, 141–147 (2002).

[8] P. Uetz et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae*. Nature, **403**, 623–627 (2000).

[9] C. von Mering et al., *Comparative assessment of large-scale data sets of protein-protein interactions*. Nature, **417**, 399–403 (2002).

[10] A. Ben-Hur and W. S. Noble., *Choosing negative examples for the prediction of protein-protein interactions*. BMC Bioinformatics, **7 Suppl 1**, S2 (2006).

[11] X. Hu, C. Feng, T. Ling and M. Chen, *Deep learning frameworks for protein-protein interaction prediction*. Computational and Structural Biotechnology Journal, **20**, 3223–3233 (2022).

[12] F. Soleymani, E. Paquet, H. Viktor, W. Michalowski and D. Spinello, *Protein-protein interaction prediction with deep learning: A comprehensive review*. Computational and Structural Biotechnology Journal, **20**, 5316–5341 (2022).

[13] A. Chakraborty et al., *Determining protein–protein interaction using support vector machine: a review.* IEEE Access, **9**, 12473–12490 (2021).

[14] L. Hu, X. Wang, Y. A. Huang, P. Hu and Z. H. You, *A survey on computational models for predicting protein-protein interactions.* Briefings in Bioinformatics, **22**, bbab036 (2021).

[15] UniProt Consortium, *UniProt: a worldwide hub of protein knowledge.* Nucleic Acids Research, **47**, D506–D515 (2019).

[16] B. Boeckmann et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Research, **31**, 365–370 (2003).

[17] D. G. George, W. C. Barker, H. W. Mewes, F. Pfeiffer and A. Tsugita, *The PIR-international protein sequence database.* Nucleic Acids Research, **24**, 17–20 (1996).

[18] E. W. Sayers et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Research, **49**, D10–D17 (2021).

[19] J. S. Garavelli, Z. Hou, N. Pattabiraman and R. M. Stephens, *The RESID Database of protein structure modifications and the NRL-3D Sequence-Structure Database.* Nucleic Acids Research, **29**, 199–201 (2001).

[20] S. K. Burley et al., *Protein Data Bank (PDB): the single global macromolecular structure archive.* Methods in Molecular Biology (Clifton, N.J.), **1607**, 627–641 (2017).

[21] A. Andreeva et al., *SCOP database in 2004: refinements integrate structure and sequence family data.* Nucleic Acids Research, **32**, D226–D229 (2004).

[22] E. L. Sonnhammer, S. R. Eddy and R. Durbin, *Pfam: a comprehensive database of protein domain families based on seed alignments.* Proteins, **28**, 405–420 (1997).

[23] S. K. Ng, Z. Zhang, S. H. Tan and K. Lin, *InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes.* Nucleic Acids Research, **31**, 251–254 (2003).

[24] R. Mosca, A. Céol, A. Stein, R. Olivella and P. Aloy, *3did: a catalog of domain-based interactions of known three-dimensional structure.* Nucleic Acids Research, **42**, D374–D379 (2014).

[25] Gene Ontology Consortium, *Expansion of the Gene Ontology knowledgebase and resources*. Nucleic Acids Research, **45**, D331–D338 (2017).

[26] D. Binns et al., *QuickGO: a web-based tool for Gene Ontology searching*. Bioinformatics, **25**, 3045–3046 (2009).

[27] P. Pagel et al., *The MIPS mammalian protein-protein interaction database*. Bioinformatics, **21**, 832–834 (2005).

[28] J. Binkley et al., *The Candida Genome Database: the new homology information page highlights protein similarity and phylogeny*. Nucleic Acids Research, **42**, D711–D716 (2014).

[29] I. Xenarios et al., *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions*. Nucleic Acids Research, **30**, 303–305 (2002).

[30] S. Orchard et al., *The MIntAct project – IntAct as a common curation platform for 11 molecular interaction databases*. Nucleic Acids Research, **42**, D358–D363 (2014).

[31] G. D. Bader, D. Betel and C. W. Hogue, *BIND: the Biomolecular Interaction Network Database*. Nucleic Acids Research, **31**, 248–250 (2003).

[32] A. Chatr-aryamontri et al., *MINT: the Molecular INTeraction database*. Nucleic Acids Research, **35**, D572–D574 (2007).

[33] A. Chatr-Aryamontri et al., *The BioGRID interaction database: 2017 update*. Nucleic Acids Research, **45**, D369–D379 (2017).

[34] T. S. Keshava Prasad et al., *Human Protein Reference Database – 2009 update*. Nucleic Acids Research, **37**, D767–D772 (2009).

[35] D. Szklarczyk et al., *The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible*. Nucleic Acids Research, **45**, D362-d368 (2017).

[36] X. Yang et al., *HVIDB: a comprehensive database for human-virus protein-protein interactions*. Briefings in Bioinformatics, **22**, 832–844 (2021).

[37] A. Calderone, L. Licata and G. Cesareni, *VirusMentha: a new resource for virus-host protein interactions*. Nucleic Acids Research, **43**, D588–D592 (2015).

[38] G. Alanis-Lobato, M. A. Andrade-Navarro and M. H. Schaefer, *HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks*. Nucleic Acids Research, **45**, D408–D414 (2016).

[39] H. Gu, P. Zhu, Y. Jiao, Y. Meng and M. Chen, *PRIN: a predicted rice interactome network*. BMC Bioinformatics, **12**, 161 (2011).

[40] A. Sapkota et al., *DIPOS: database of interacting proteins in Oryza sativa*. Molecular BioSystems, **7**, 2615–2621 (2011).

[41] S. Durmuş Tekir et al., *PHISTO: pathogen-host interaction search tool*. Bioinformatics, **29**, 1357–1358 (2013).

[42] A. Chatr-aryamontri et al., *VirusMINT: a viral protein interaction database*. Nucleic Acids Research, **37**, D669–D673 (2009).

[43] T. Guirimand, S. Delmotte and V. Navratil, *VirHostNet 2.0: surfing on the web of virus/host molecular interactions data*. Nucleic Acids Research, **43**, D583–D587 (2015).

[44] P. Smialowski et al., *The Negatome database: a reference set of non-interacting protein pairs*. Nucleic Acids Research, **38**, D540–D544 (2010).

[45] P. Blohm et al., *Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis*. Nucleic Acids Research, **42**, D396–D400 (2014).

[46] X. W. Chen, J. C. Jeong and P. Dermyer., *KUPS: constructing datasets of interacting and non-interacting protein pairs with associated attributions*. Nucleic Acids Research, **39**, D750–D754 (2011).

[47] L. G. Trabuco, M. J. Betts and R. B. Russell, *Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments*. Methods, **58**, 343–348 (2012).

[48] J. Shen et al., *Predicting protein-protein interactions based only on sequences information*. Proceedings of the National Academy of Sciences of the United States of America, **104**, 4337–4341 (2007).

[49] Y. Guo, L. Yu, Z. Wen and M. Li, *Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences*. Nucleic Acids Research, **36**, 3025–3030 (2008).

[50] S. Hashemifar, B. Neyshabur, A. A. Khan and J. Xu, *Predicting protein-protein interactions through sequence-based deep learning*. Bioinformatics, **34**, i802-i810 (2018).

[51] N. Zhao, M. Zhuo, K. Tian and X. Gong, *Protein-protein interaction and non-interaction predictions using gene sequence natural vector*. Communications Biology, **5**, 652 (2022).

[52] X. Hu, C. Feng, Y. Zhou, A. Harrison and M. Chen, *DeepTrio: a ternary prediction system for protein-protein interaction using mask multiple parallel convolutional neural networks*. Bioinformatics, **38**, 694–702 (2021).

[53] M. Baranwal et al., *Struct2Graph: a graph attention network for structure based predictions of protein-protein interactions*. BMC Bioinformatics, **23**, 370 (2022).

[54] B. Song et al., *Learning spatial structures of proteins improves protein-protein interaction prediction*. Briefings in Bioinformatics, **23**, bbab558 (2022).

[55] S. Wuchty, *Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens*. PloS ONE, **6**, e26960 (2011).

[56] M. Kshirsagar, J. Carbonell and J. Klein-Seetharaman., *Multitask learning for host-pathogen protein interactions*. Bioinformatics, **29**, i217–i226 (2013).

[57] F. E. Eid, M. ElHefnawi and L. S. Heath, *DeNovo: virus-host sequence-based protein-protein interaction prediction*. Bioinformatics, **32**, 1144–1150 (2016).

[58] X. Yang, S. Yang, Q. Li, S. Wuchty and Z. Zhang., *Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method*. Computational and Structural Biotechnology Journal, **18**, 153–161 (2020).

[59] S. Tsukiyama, M. M. Hasan, S. Fujii and H. Kurata, *LSTM-PHV: prediction of human-virus protein-protein interactions by LSTM with word2vec*. Briefings in Bioinformatics, **22**, bbab228 (2021).

[60] X. Yang, S. Yang, X. Lian, S. Wuchty and Z. Zhang, *Transfer learning via multi-scale convolutional neural layers for human-virus protein-protein interaction prediction*. Nature, **37**, 4771–4778 (2021).

[61] J. Jumper et al., *Highly accurate protein structure prediction with AlphaFold*. Nature, **596**, 583–589 (2021).

[62] F.-L. Lai and F. Gao, *Auto-Kla: a novel web server to discriminate lysine lactylation sites using automated machine learning*. Briefings in Bioinformatics, bbad070 (2023).

[63] J. R. Bock and D. A. Gough., *Predicting protein–protein interactions from primary structure*. Bioinformatics, **17**, 455–460 (2001).

[64] L. Yang, J. F. Xia and J. Gui, *Prediction of protein-protein interactions from protein sequence using local descriptors*. Protein and Peptide Letters, **17**, 1085–1090 (2010).

[65] Y. Zhou, Y. S. Zhou, F. He, J. Song and Z. Zhang, *Can simple codon pair usage predict protein-protein interaction?*. Molecular BioSystems, **8**, 1396–1404 (2012).

[66] H. S. Najafabadi and R. Salavati., *Sequence-based prediction of protein-protein interactions by means of codon usage*. Genome Biology, **9**, R87 (2008).

[67] C. Chen, Q. Zhang, Q. Ma and B. Yu, *LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion*. Chemometrics and Intelligent Laboratory Systems, **191**, 54–64 (2019).

[68] J. Wang, L. Zhang, L. Jia, Y. Ren and G. Yu, *Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences*. International Journal of Molecular Sciences, **18**, 2373 (2017).

[69] X. Du et al., *DeepPPI: boosting prediction of protein-protein interactions with deep neural networks*. Journal of Chemical Information and Modeling, **57**, 1499–1510 (2017).

[70] I. Ahmed, P. Witbooi and A. Christoffels, *Prediction of human-Bacillus anthracis protein-protein interactions using multi-layer neural network*. Bioinformatics, **34**, 4159–4164 (2018).

[71] L. Zhang, G. Yu, D. Xia and J. Wang, *Protein–protein interactions prediction based on ensemble deep neural networks*. Neurocomputing, **324**, 10–19 (2019).

[72] C. Chen et al., *Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier*. Computers in Biology and Medicine, **123**, 103899 (2020).

[73] L. Hu and K. C. Chan., *Extracting coevolutionary features from protein sequences for predicting protein-protein interactions*. IEEE/ACM

Transactions on Computational Biology and Bioinformatics, **14**, 155–166 (2017).

[74] C. Yin and S. S. Yau., *A coevolution analysis for identifying protein-protein interactions by Fourier transform*. PloS ONE, **12**, e0174862 (2017).

[75] D. T. Jones, *Protein secondary structure prediction based on position-specific scoring matrices*. Journal of Molecular Biology, **292**, 195–202 (1999).

[76] X. W. Chen and J. C. Jeong, *Sequence-based prediction of protein interaction sites with an integrative method*. Bioinformatics, **25**, 585–591 (2009).

[77] D. T. Jones and J. J. Ward., *Prediction of disordered regions in proteins from position specific score matrices*. Proteins, **53 Suppl 6**, 573–578 (2003).

[78] M. Chen et al., *Multifaceted protein-protein interaction prediction based on Siamese residual RCNN*. Bioinformatics, **35**, i305–i314 (2019).

[79] S. Madan, V. Demina, M. Stapf, O. Ernst and H. Fröhlich., *Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings*. Patterns, **3**, 100551 (2022).

[80] A. Elnaggar et al., *ProtTrans: toward understanding the language of life through self-supervised learning*. IEEE Transactions on Pattern Analysis and Machine Intelligence, **44**, 7112–7127 (2022).

[81] J. M. Doolittle and S. M. Gomez, *Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens*. Virology Journal, **7**, 82 (2010).

[82] W. Zhang, M. P. Coba and F. Sun, *Inference of domain-disease associations from domain-protein, protein-disease and disease-disease relationships*. BMC Systems Biology, **10 Suppl 1**, 4 (2016).

[83] G. M. Qin, R. Y. Li and X. M. Zhao., *Identifying disease associated miRNAs based on protein domains*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, **13**, 1027–1035 (2016).

[84] M. Deng, S. Mehta, F. Sun and T. Chen., *Inferring domain-domain interactions from protein-protein interactions*. Genome Research, **12**, 1540–1548 (2002).

[85] S. Ma et al., *Prediction of protein-protein interactions between fungus (Magnaporthe grisea) and rice (Oryza sativa L.)*. Briefings in Bioinformatics, **20**, 448–456 (2019).

[86] Y. Yan, H. Tao, J. He and S. Y. Huang, *The HDOCK server for integrated protein-protein docking*. Nature Protocols, **15**, 1829–1852 (2020).

[87] B. G. Pierce et al., *ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers*. Bioinformatics, **30**, 1771–1773 (2014). MR3064822

[88] C. Christoffer et al., *LZerD webserver for pairwise and multiple protein-protein docking*. Nucleic Acids Research, **49**, W359–W365 (2021).

[89] G. C. P. van Zundert et al., *The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes*. Journal of Molecular Biology, **428**, 720–725 (2016).

[90] D. Kozakov et al., *The ClusPro web server for protein-protein docking*. Nature Protocols, **12**, 255–278 (2017).

[91] B. Jiménez-García et al., *LightDock: a new multi-scale approach to protein-protein docking*. Bioinformatics, **34**, 49–55 (2018).

[92] C. Mirabello and B. Wallner, *InterPred: a pipeline to identify and model protein-protein interactions*. Proteins, **85**, 1159–1170 (2017).

[93] E. M. Marcotte et al., *Detecting protein function and protein-protein interactions from genome sequences*. Science, **285**, 751–753 (1999).

[94] A. J. Enright, I. Iliopoulos, N. C. Kyrpides and C. A. Ouzounis, *Protein interaction maps for complete genomes based on gene fusion events*. Nature, **402**, 86–90 (1999).

[95] J. Tamames, G. Casari, C. Ouzounis and A. Valencia, *Conserved clusters of functionally related genes in two bacterial genomes*. Journal of Molecular Evolution, **44**, 66–73 (1997).

[96] T. Dandekar, B. Snel, M. Huynen and P. Bork, *Conservation of gene order: a fingerprint of proteins that physically interact*. Trends in Biochemical Sciences, **23**, 324–328 (1998).

[97] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates, *Assigning protein functions by comparative genome analy-*

sis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences of the United States of America, **96**, 4285–4288 (1999).

[98] M. Li, J. Wang, X. Chen, H. Wang and Y. Pan, *A local average connectivity-based method for identifying essential proteins from the network level*. Computational Biology and Chemistry, **35**, 143–150 (2011). MR2848860

[99] I. A. Kovács et al., *Network-based prediction of protein interactions*. Nature, **10**, 1240 (2019).

[100] C. Lei and J. Ruan, *A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity*. Bioinformatics, **29**, 355–364 (2013).

[101] Q. C. Zhang et al., *Structure-based prediction of protein-protein interactions on a genome-wide scale*. Nature, **490**, 556–560 (2012).

[102] W. Liu-Wei et al., *DeepViral: prediction of novel virus-host interactions from protein sequences and infectious disease phenotypes*. Bioinformatics, **37**, 2722–2729 (2021).

[103] Y. K. Lei, Z. H. You, Z. Ji, L. Zhu and D. S. Huang, *Assessing and predicting protein interactions by combining manifold embedding with multiple information integration*. BMC Bioinformatics, **13 Suppl 7**, S3 (2012).

[104] L. Liu et al., *Combining sequence and network information to enhance protein-protein interaction prediction*. BMC Bioinformatics, **21**, 537 (2020).

[105] M. Huang et al., *Discovering patterns to extract protein-protein interactions from full texts*. Bioinformatics, **20**, 3604–3612 (2004).

[106] Y. Hao, X. Zhu, M. Huang and M. Li, *Discovering patterns to extract protein-protein interactions from the literature: Part II*. Bioinformatics, **21**, 3294–3300 (2005).

[107] H. Jang et al., *Finding the evidence for protein-protein interactions from PubMed abstracts*. Bioinformatics, **22**, e220–e226 (2006).

[108] L. Wei et al., *Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, **11**, 192–201 (2014).

[109] L. Wei et al., *Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier*. Artificial Intelligence in Medicine, **83**, 67–74 (2017).

[110] S. Sledzieski, R. Singh, L. Cowen and B. Berger, *D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions*. Cell Systems, **12**, 969-982.e966 (2021).

[111] R. K. Barman, S. Saha and S. Das, *Prediction of interactions between viral and host proteins using supervised machine learning methods*. PloS ONE, **9**, e112034 (2014).

[112] R. Jansen and M. Gerstein, *Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction*. Current Opinion in Microbiology, **7**, 535–545 (2004).

[113] D. Kandel, Y. Matias, R. Unger and P. Winkler, *Shuffling biological sequences*. Discrete Applied Mathematics, **71**, 171–185 (1996). MR1420298

[114] L. Zhang, G. Yu, M. Guo and J. Wang, *Predicting protein-protein interactions using high-quality non-interacting pairs*. BMC Bioinformatics, **19**, 525 (2018).

[115] P. W. Lord, R. D. Stevens, A. Brass and C. A. Goble, *Semantic similarity measures as tools for exploring the gene ontology*. in: Pacific Symposium on Biocomputing, 601–612 (2003).

[116] P. W. Lord, R. D. Stevens, A. Brass and C. A. Goble, *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation*. Bioinformatics, **19**, 1275–1283 (2003).

[117] G. Wu, X. Feng and L. Stein, *A human functional protein interaction network and its application to cancer data analysis*. Genome Biology, **11**, R53 (2010).

[118] L. Dey, S. Chakraborty and A. Mukhopadhyay, *Machine learning techniques for sequence-based prediction of viral-host interactions between SARS-CoV-2 and human proteins*. Biomedical Journal, **43**, 438–450 (2020).

[119] P. Bryant, G. Pozzati and A. Elofsson, *Improved prediction of protein-protein interactions using AlphaFold2*. Nature Communications, **13**, 1265 (2022).

[120] S. Das and S. Chakrabarti, *Classification and prediction of protein-protein interaction interface using machine learning algorithm*. Scientific Reports, **11**, 1761 (2021).

[121] W. Li and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences*. Bioinformatics, **22**, 1658–1659 (2006).

[122] B. Lehne and T. Schlitt, *Protein-protein interaction databases: keeping up with growing interactomes*. Human Genomics, **3**, 291–297 (2009).

[123] K. Cermakova et al., *A ubiquitous disordered protein interaction module orchestrates transcription elongation*. Science, **374**, 1113–1121 (2021).

Nan Zhao
Institute for Mathematical Sciences, Renmin University of China, Beijing, China
E-mail address: zhaonanchn@ruc.edu.cn

Xinqi Gong
Institute for Mathematical Sciences, Renmin University of China, Beijing, China
Beijing Academy of Artificial Intelligence, Beijing, China
E-mail address: xinqigong@ruc.edu.cn

Received January 2, 2023