BMC Bioinformatics

RESEARCH

Open Access

Graph-based machine learning model for weight prediction in protein–protein networks

Hajer Akid[1*], Kirsley Chennen[1], Gabriel Frey[1], Julie Thompson[1], Mounir Ben Ayed[2] and Nicolas Lachiche[1]

*Correspondence:
akid.hajer@gmail.com

[1] ICube, University of Strasbourg, 67412 Illkirch Cedex, France
[2] REGIM-Lab, University of Sfax, 3038 Sfax, Tunisia

Abstract

Proteins interact with each other in complex ways to perform significant biological functions. These interactions, known as protein–protein interactions (PPIs), can be depicted as a graph where proteins are nodes and their interactions are edges. The development of high-throughput experimental technologies allows for the generation of numerous data which permits increasing the sophistication of PPI models. However, despite significant progress, current PPI networks remain incomplete. Discovering missing interactions through experimental techniques can be costly, time-consuming, and challenging. Therefore, computational approaches have emerged as valuable tools for predicting missing interactions. In PPI networks, a graph is usually used to model the interactions between proteins. An edge between two proteins indicates a known interaction, while the absence of an edge means the interaction is not known or missed. However, this binary representation overlooks the reliability of known interactions when predicting new ones. To address this challenge, we propose a novel approach for link prediction in weighted protein–protein networks, where interaction weights denote confidence scores. By leveraging data from the yeast *Saccharomyces cerevisiae* obtained from the STRING database, we introduce a new model that combines similarity-based algorithms and aggregated confidence score weights for accurate link prediction purposes. Our model significantly improves prediction accuracy, surpassing traditional approaches in terms of Mean Absolute Error, Mean Relative Absolute Error, and Root Mean Square Error. Our proposed approach holds the potential for improved accuracy in predicting PPIs, which is crucial for better understanding the underlying biological processes.

Keywords: Protein–protein interactions, Weighted graphs, Machine learning, Link prediction

Background

Proteins are complex and essential molecules that play critical roles in various biological functions within living organisms, including DNA transcription and replication, hormone transport, signal transduction, and catalyzing biochemical reactions [1–3]. These molecules interact with each other to regulate and coordinate intricate biological

processes through protein–protein interactions (PPIs). Therefore, modeling PPI networks is vital for identifying proteins' roles in cellular functions [4, 5]. The study of PPIs also has important applications, for example in disease therapies and development of novel drugs [6]. Targeting specific PPIs enables the modulation of protein functions and the influence on pathways implicated in diseases [7, 8].

Despite significant progress in experimental techniques, detecting PPIs remains a challenging, time-consuming, and labor-intensive task. High-throughput experimental techniques such as yeast two-hybrid (Y2H), protein microarrays, and mass spectrometry (MS) have significantly enhanced our ability to detect PPIs and have enabled the collection of extensive PPI data [9]. However, these techniques come with their own set of limitations. One significant challenge is the occurrence of false positives, where nonspecific interactions are incorrectly identified as true interactions, and false negatives, where actual interactions are missed. These inaccuracies lead to incomplete and sometimes unreliable PPI datasets, making it difficult to assess the true nature and extent of protein interactions [10, 11].

Numerous experimental PPI datasets are available via public online databases, such as BioGRID (Biological General Repository for Interaction Datasets) [12], or STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) [13]. In STRING, data is collected from diverse sources and confidence scores are assigned to each interaction, assisting researchers in evaluating the reliability of the data. Each interaction in STRING is annotated based on scores from seven individual channels: experiments, database curation, text mining, coexpression, neighborhood, fusion, and cooccurrence. The combined score is calculated by integrating the probabilities from the different individual channels and adjusted for the probability of randomly observing an interaction [14].

These confidence scores, typically normalized to a range between 0 and 999, offer a probabilistic measure of the interaction's reliability, enabling researchers to explore and analyze complex networks of protein interactions effectively. The confidence scores in STRING are categorized based on their values, providing a tiered assessment of interaction reliability. Interactions with scores greater than 900 are considered to have the highest confidence level, indicating robust supporting evidence. Those with scores between 700 and 900 are classified as high confidence, while scores between 400 and 700 are categorized as medium confidence. Scores between 150 and 400 are considered low confidence, and scores less than 150 are very low confidence. As shown in Table 1, in the last three versions (11.0, 11.5, and 12.0) of STRING used to validate our work, only a few PPIs are considered to have high confidence levels. For example, as depicted in Fig. 1, the interaction between the protein pair 4932.

Table 1 Percentage of interactions for each confidence level for the three versions V11.0, V11.5 and V12.0 of the STRING database

Confidence level	% of PPI in V11.0	% of PPI in V11.5	% of PPI in V12.0
Highest (≥ 900)	7.5	6.3	3.7
High (700–900)	5.3	5.8	3.6
Medium (400–700)	16.0	16.4	12.5
Low (150–400)	71.2	71.5	80.2

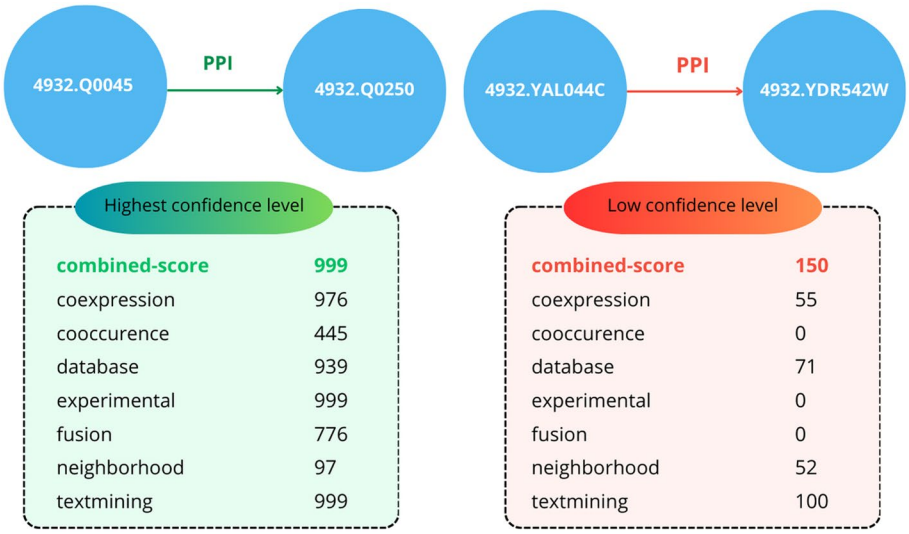

Fig. 1 Examples of two PPI confidence scores from STRING database (version 12.0)

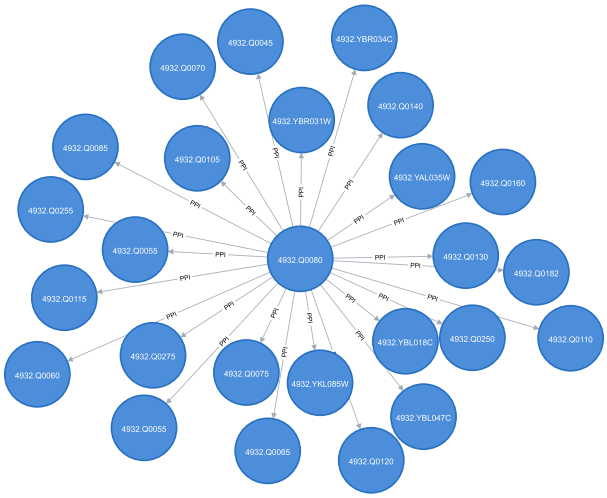

Fig. 2 Subset of the PPI Network from STRING Database Version 12.0 showing the known interactions with protein 4932.Q0050

Q0045 and 4932.Q0250 is assigned the highest confidence score of 999 in version 12.0 of STRING. Conversely, while a PPI exists between the protein pair 4932.YAL044C and 4932.YDR542W, its confidence score is low, at 150.

In addition to experimental techniques, researchers have increasingly adopted and evolved computational methods to study PPIs. These computational approaches aim to reduce the number of candidate protein pairs requiring experimental validation by leveraging existing PPI datasets to identify potential novel interactions. Given the graph-like structure of PPI networks, most computational methods utilize a graph-based representation, where nodes represent proteins and edges denote interactions between them as shown in Fig. 2. In a graph-like structure, the connections between proteins are typically represented as unweighted and binary, indicating the presence

(1) or absence (0) of an interaction. Detecting new interactions is thus framed as a link prediction task, which aims to predict unknown connections between proteins within the PPI network. Computational approaches for predicting PPIs can be broadly categorized into two main types: similarity-based methods (network-based features) [3, 15–17] and machine learning models, which encompass both traditional and deep learning techniques [18–21]. Among machine learning models, graph representation learning approaches are particularly well-suited for predicting interactions and recognizing complex structures in PPI networks, as they effectively capture the underlying topological and relational patterns within these networks [22–26].

Similarity-based methods for PPI prediction

Similarity-based methods constitute a class of computational techniques that leverage various similarity indices, initially pioneered in social network analysis, to infer missing connections. These methods primarily rely on existing distinct paths of length 2 between two nodes to predict the missing direct link between them. Within the realm of protein–protein interaction (PPI) prediction, researchers have adapted and tested several similarity indices to assess the likelihood of interactions between proteins. Notable examples include common neighbors [27], Adamic Adar [28], and preferential attachment [29]. These indices are designed to compute a likelihood measure for each pair of unconnected proteins, capturing their topological or structural similarities within the PPI network. Subsequently, these computed measures are sorted in descending order to prioritize the most promising candidate interactions. A comprehensive exploration of these similarity indices is provided in [17]. In this section, we specifically focus on the similarity indices employed in our research, drawing from established methods widely documented in the literature.

Common Neighbors (CN)

CN is one of the most prevalent similarity calculation indices in link prediction domains due to its simplicity and effectiveness in certain areas such as scientific collaboration graphs [29] and social networks [30]. The idea behind CN is that the probability of two nodes being connected in the future is influenced by the number of their common neighbors, meaning two nodes are highly likely to form a link if they share more neighbors. This measure is defined as follows (1):

$$CN(X, Y) = |N(X) \cap N(Y)| \tag{1}$$

where $N(X)$ is the set of nodes adjacent to node X, and $N(Y)$ is the set of nodes adjacent to node Y.

Jaccard coefficient (CJ)

Unlike the unnormalized CN index, CJ not only considers the number of common neighbors but also normalizes it by considering the total set of shared and unshared neighbors [31]. The formulation of this measure is as follows (2):

$$CJ(X, Y) = \frac{CN(X, Y)}{TN(X, Y)} \tag{2}$$

where $TN(X, Y)$ is the total number of neighbors of X and Y, defined as follows (3):

$$TN(X,Y) = |N(X) \cup N(Y)| \tag{3}$$

Adamic adar (AA)

The AA index, proposed by [28], gives more weight to relatively fewer common neighbors. In the AA formula, the shared neighbors of two nodes are penalized by the logarithm of their degrees as follows (4):

$$AA(X,Y) = \sum_{u \in N(X) \cap N(Y)} \frac{1}{\log |N(u)|} \tag{4}$$

Preferential attachment (PA)

It has been shown by [29] that new nodes joining the network are highly likely to be connected to an existing node with higher degrees rather than to a node with lower degrees. Thus, the PA index was proposed (5):

$$PA(X,Y) = |N(X)| * |N(Y)| \tag{5}$$

Resource allocation (RA)

An RA index, very similar to AA, was proposed by [15]. Some studies show that the performances of RA and AA are very close when the average degree of the network is low. However, PA outperforms RA when the average degree of the network is high [32, 33]. The RA measure is defined as follows (6):

$$RA(X,Y) = \sum_{u \in N(X) \cap N(Y)} \frac{1}{|N(u)|} \tag{6}$$

Length-3 paths (L3)

The L3 similarity measure was introduced by [34] to predict missing links in PPI networks. In this study, the authors demonstrate that, unlike its performance in social networks, the Jaccard coefficient fails in the case of detecting new PPIs. Proteins interact not because they are similar to each other, but if one of them is similar to the neighboring proteins of the other. The formula for the L3 measure is as follows (7):

$$P(X,Y) = \sum_{U,V} \frac{a_{XU} * a_{UV} * a_{VY}}{\sqrt{|N(U)||N(V)|}} \tag{7}$$

where $a_{XU} = 1$ if proteins X and U interact, and zero otherwise.

For example, the two proteins 4932.*YDL*015*C* and 4932.*YHL*009*C* are not connected by a path of length 2 in STRING version 11.5. Consequently, the value of the CN measure is 0, and the proteins are considered non-relevant candidate protein pairs. However, in the same version, they are linked by 1259 paths of length 3. Figure 3 illustrates some of the paths of length 3 connecting these proteins. Upon examining version 12.0

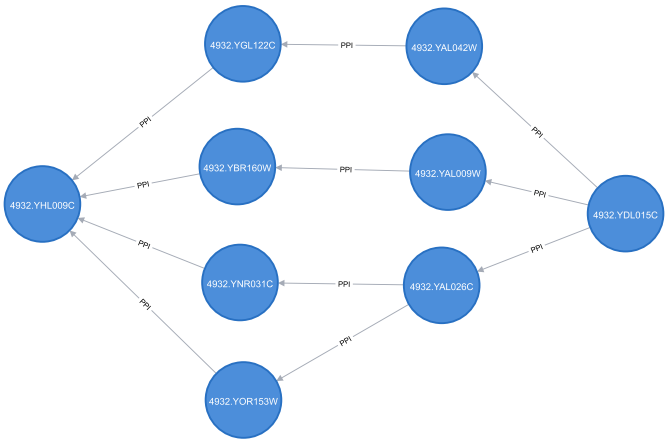

Fig. 3 Paths of length 3 between the proteins 4932.*YDL*015*C* and 4932.*YHL*009*C* in STRING v12.0

Table 2 Similarity measures for link prediction

Similarity measure	Formula	Path length	Weighted?
Common neighbors (CN)	$CN(X,Y) = \lvert N(X) \cap N(Y)\rvert$	2	No
Jaccard coefficient (CJ)	$CJ(X,Y) = \frac{CN(X,Y)}{TN(X,Y)}$	2	No
Adamic adar (AA)	$AA(X,Y) = \sum_{u \in N(X) \cap N(Y)} \frac{1}{\log \lvert N(u)\rvert}$	2	No
Preferential attachment (PA)	$PA(X,Y) = \lvert N(X)\rvert * \lvert N(Y)\rvert$	2	No
Resource allocation (RA)	$RA(X,Y) = \sum_{u \in N(X) \cap N(Y)} \frac{1}{\lvert N(u)\rvert}$	2	No
Length-3 paths (L3)	$P(X,Y) = \sum_{U,V} \frac{a_{XU} * a_{UV} * a_{VY}}{\sqrt{\lvert N(U)\rvert \lvert N(V)\rvert}}$	3	No

of STRING, a new direct link between these proteins has been added with the highest confidence level, indicated by a combined score of 856.

L3 has been tested on several databases such as HI-II-14 [35] and the PPIs of the *Saccharomyces cerevisiae* organism in STRING [36]. Since a similarity-based method provides a measure for a pair of proteins not connected by a direct link, this value was directly used for predictions [34], typically by considering the top k computed measures. Note that in this case, there is no construction of a model using a training set and PPI collected from STRING are used without considering their confidence scores. The performance evaluation of the L3 method showed promising results and better performance compared to topological measures based on paths of length 2. The performance of L3 has been evaluated up to paths of length 8, and the best results were detected for length 3.

Table 2 summarizes the various similarity measures we used in our work. Each measure is defined by its formula, the path length used, and whether or not it takes into account the weights of the interactions.

Some research works have also proposed approaches to predict missing links by combining multiple topology-based measures. These approaches utilize the values of topological measures calculated based on the information regarding the existence or absence of a path of length 2. These values are subsequently fed into a supervised learning algorithm to predict the existence of a potential missing link [37, 38].

Machine learning models

Several researchers have proposed various machine learning models for PPI prediction [39–41]. The models can be divided into three categories based on the type of features used: sequence-based, 3D structure-based, and hybrid methods [18].

Sequence-based methods analyze the amino acid sequences of the proteins to predict interactions. These methods utilize various computational and statistical techniques to infer interactions based on sequence information content. However, effectively extracting and combining these features remains a challenge. Various sequence-based models have been proposed using machine learning techniques, such as support vector machines (SVM) and random forest-based methods [42–44]. While these methods have demonstrated effectiveness, they require comprehensive and high-quality sequence data, which may not always be available, leading to incomplete or biased predictions.

3D structure-based prediction models rely on the three-dimensional structures of proteins to predict interactions. These models consider the precise conformational and sequence characteristics that distinguish various structures within a family. The use of docking simulation and modeling physical interactions provides more detailed and specific interaction information. Several machine learning models based on 3D structure have been proposed, such as the RF classifier proposed in [45], which integrates several structural features to distinguish correct from incorrect PPI models. Additionally, [46] proposed a graph convolutional network (GCN) and graph attention network (GAT) to predict interactions between proteins using structural information and sequence features. Although 3D structure-based methods offer more detailed insights into protein interactions, they are hindered by the limited number of known protein structures and the requirement of high computational resources [47].

Hybrid methods combine sequence-based and structure-based features with other functions to provide more comprehensive and accurate predictions [48]. However, they face challenges in effectively integrating disparate data types and addressing data quality and completeness issues.

While the existing proposed models have shown promising results, several limitations need to be considered. One of the main challenges with these approaches is feature extraction and the combination of these features, which requires detailed information about the proteins involved. This kind of information is often incomplete or unavailable, leading to inaccurate predictions. Moreover, most existing machine learning models for PPI prediction focus on the analysis of each pair of protein sequences without considering the accuracy of existing interactions.

In this work, we present an original and innovative approach designed to enhance current PPI datasets by identifying accurate PPIs. Although our methodology is grounded in the graph structure of PPI networks, its uniqueness lies in the incorporation of interaction weights into a PPI-weighted graph (PPI-WG) when predicting missing links. This approach not only predicts missing interactions but also evaluates their reliability. Our contributions are multifaceted: (i) proposing a novel representation of PPI networks that integrates interaction weights; (ii) combining topological measures with aggregation-based representations; and (iii) developing a machine learning model capable of making precise predictions.

Method

To achieve our objectives, we first proposed a novel representation of PPI networks that represents the interactions between two nodes by aggregating all the combined scores from STRING in paths of length 3 between these nodes. The combined scores from STRING provide a probabilistic measure of the interaction's reliability, which is crucial for our predictive model. Using the combined scores from STRING, our model leverages the confidence scores derived from all the individual channels, including experiments, database curation, text mining, coexpression, neighborhood, fusion, and cooccurrence. We then developed a new model incorporating different aggregation-based representations and topological measures to make accurate predictions. This multi-stage process is illustrated in Fig. 4.

Aggregation-based representations for weighted PPI networks (WPPNs)

The majority of similarity measures proposed in the literature are defined for unweighted graphs and typically for paths of length 2. To meet the emerging needs of making predictions from a weighted PPI and considering paths of length 3, we propose a new representation based on aggregation. Our approach leverages aggregation functions, namely MIN, AVG, and MAX, to synthesize the values carried by intermediate paths of length 3 between each pair of proteins. This process involves two key steps and the selection of one of the 3 aggregation functions at each step. Firstly, the aggregation of values for each intermediate path entails summarizing all values into a single score. For instance, the MIN function can be used to extract the minimum value on each intermediate path of length 3, resulting in a list comprising the minimum scores carried by all intermediate paths. Secondly, the values obtained from the first aggregation are further summarized. For example, the AVG function can be used to compute the average of all minimum values obtained, yielding a single score denoted as α. The formulation of the AVGMIN measure is as follows:

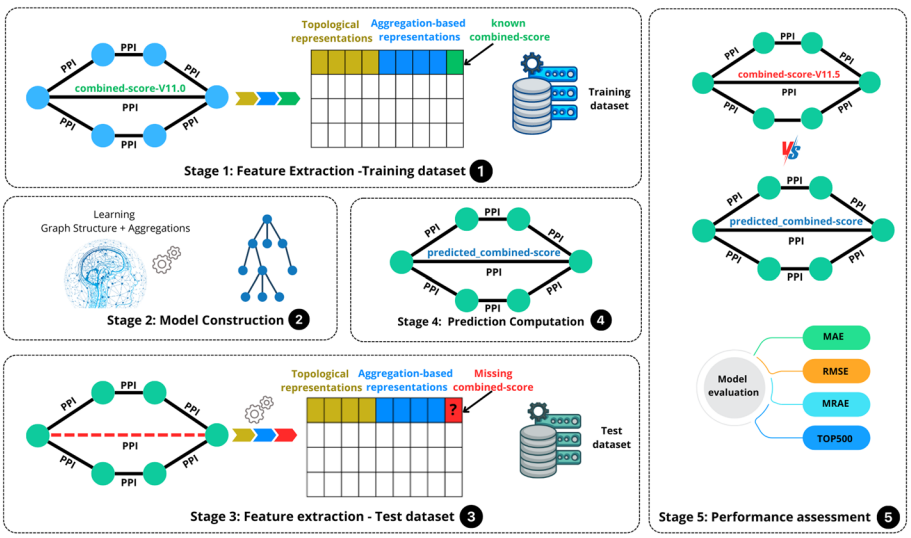

Fig. 4 Stages of the proposed approach for predicting new protein–protein interactions

$$AVGMIN(X,Y) = AVG_{\forall \text{path } p \text{ of length 3 between } X \text{ and } Y}(MIN_{\forall \text{arc } a \text{ of path } p}(score(a))) \quad (8)$$

Our aggregation-based representation approach thus allows us to define nine possible combinations (representations) that can be provided to a learning algorithm to predict missing interactions and their values.

Graph-based machine learning model incorporating different representations

We propose employing the various representations obtained through aggregations, along with different topological similarity measures, as input to a learning algorithm. Our method is a five-stage process that enables the learning algorithm to integrate information from the diverse descriptors, as follows:

Stage 1: feature extraction: training dataset

In the first stage, we used a subset of data selected from the STRING version 11.0 database. Protein pairs selected for analysis were those connected by paths of length 3 and possessing direct interactions simultaneously. For these protein pairs, we computed six topological measures: CN, AA, PA, RA, TN, L3. Additionally, nine aggregation-based representations of interaction weights were calculated for these pairs: MIN-MIN, MIN-AVG, MIN-MAX, AVG-MIN, AVG-AVG, AVG-MAX, MAX-MIN, MAX-AVG, and MAX-MAX. For each selected protein pair, these fifteen features, along with the known combined-scores, were fused into a single vector to represent the topological structure and weight of the interaction. The vectors of all selected pairs were then used as a training dataset.

Stage 2: model construction

Using the training dataset, a graph-based regression tree model was constructed to capture the intricate relationships between the extracted features and the known combined-scores of protein–protein interactions (PPIs). This choice of a regression tree was motivated by its ability to handle non-linear relationships, making it particularly well-suited for modeling the complex patterns inherent in biological networks. The regression tree model is designed to split the dataset recursively based on feature values, creating a series of decision nodes that lead to predictions of interaction scores. The aim of this study is not merely to identify the best algorithm but to demonstrate the value of combining topological information with aggregation-based representations in enhancing the accuracy of PPI predictions. By leveraging this mixed approach, the model is capable of discerning subtle patterns and relationships within the data that may be overlooked by traditional methods. This comprehensive representation of interaction features enhances the model's predictive power and provides deeper insights into the underlying biological processes. Furthermore, the model was rigorously validated through cross-validation techniques to ensure its robustness and generalizability. This validation process helps to mitigate overfitting and confirms that the model's predictions are reliable when applied to new, unseen data.

Stage 3: feature extraction: test dataset

After constructing the model using the training dataset, we employed another subset of the STRING 11.0 database to create a test dataset. For this purpose, we specifically selected pairs of proteins that were connected by paths of length 3 but did not have direct interactions. This selection criterion ensured that the pairs were not yet verified in STRING 11.0, making them potential candidates for novel PPIs. This hypothesis was to be tested using our model, which is capable of predicting missing PPIs and assessing their accuracies. Topological and aggregation-based representations were combined into a feature vector for each protein pair, forming the input for the trained regression tree model.

Stage 4: prediction computation

In this stage, we applied the trained graph-based regression tree model to the test dataset to predict the combined-scores for protein pairs that lacked direct interactions. The test dataset, prepared in the previous stage, consisted of protein pairs connected by paths of length 3, with calculated topological and aggregation-based features. Each feature vector in the test dataset was processed by the regression tree model to output a predicted combined-score, which ranges from 0 to 1000 and represents the likelihood and strength of a potential interaction between the protein pairs.

Stage 5: performance assessment

Since our predictions were made using STRING 11.0, where the selected protein pairs were not connected, we used the subsequent version, STRING 11.5, to verify whether our predicted interactions were added in this version and to assess how close our predicted scores were to the current scores in the database. This validation is challenging due to the dynamic nature of PPI networks, as some of our potential candidates may be highly relevant but not yet included in the database. To analyze the impact of this dynamic aspect in more depth, we also utilized the latest version, STRING 12.0, to determine the rate at which PPI predictions that were not included in STRING 11.5 were subsequently added in STRING 12.0. This comprehensive analysis allows us to understand better the temporal evolution of PPI networks and the potential lag in database updates for significant protein interactions. We compared the predicted scores with actual combined-scores from STRING V11.5 versions. Performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Relative Absolute Error (MRAE), and the Top500 predictions were used to assess the model's accuracy and robustness.

Experiments and results

To conduct our experiments, we used a computer equipped with 32 GB of RAM and 8 TB of disk storage, running Ubuntu 18.04.01 LTS. We utilized Neo4j (version 4.4.2) to import and store the STRING PPIs for *Saccharomyces cerevisiae*. To compare our predictions based on a subset of STRING 11.0 with the interactions in STRING 11.5, we created separate databases for each version. The predictions we made were added as new weighted links in STRING 11.5 and then compared to the actual interactions

in that version. All queries were written and executed in Neo4j using the Cypher query language. To create the training and test datasets, we used cypher to compute topological and aggregation-based representations. Using a graph database facilitated the identification of all paths of length 3 between selected protein pairs in both the training and test datasets. Table 3 summarizes the size of the training and test sets, detailing the number of paths of length 3 between proteins connected directly (training) and those not connected directly (test). It also provides the number of subscores from the individual channels. As shown in Table 3, most data in the training and test sets come from the coexpression, experimental, and text mining channels.

To compare the information provided by aggregation-based representations and regression model tree learning with various traditional topological measures, we used a metric that calculates the precision of the top 500 predictions (top500) [49]. In [34], precision is evaluated as the ratio of the number of links *Lp* in the Top500 to the number of links added in STRING 11.5. Therefore, if *Lp* interactions among the top500 were added in version STRING 11.5, precision is calculated as the proportion of these *Lp* predictions among 500.

In addition, to evaluate the ability of our model to predict accurate PPIs, we used three other metrics: Mean Absolute Error (*MAE*), Mean Relative Absolute Error (*MRAE*), and Root Mean Square Error (*RMSE*). The *MAE* and *RMSE* metrics calculate the average deviation between the predicted score and its actual value. The *MRAE* metric measures the average ratio of the absolute error between the expected score and the predicted score to the value of the expected score. *MRAE* is generally expressed as a percentage. For these last three metrics, we consider the regression tree model since other similarity measures do not compute a combined-score directly comparable to the score added in the STRING V11.5. Additionally, we conduct an ablation study to evaluate the contribution of our aggregation-based representations in predicting missing values. We compare the *MAE*, *MRAE*, and *RMSE* values of model trees constructed with all measures (ALL Features), with all measures except L3 (Without L3), with all measures except our aggregation representations (Without Aggregations), and with all measures except L3 and our aggregation representations (Without Aggregations and L3). The objective of this study is to determine whether the use of our aggregation representations improves the learning performance. Let A be the actual value of the missing score in the PPI graph and E be the predicted score using the data extracted from the graph.

Table 3 Training and test dataset sizes across different channels

Channel	Training dataset	Test dataset
Combined	919,914	234,172
Coexpression	552,315	118,733
Cooccurrence	3539	29
Database	43,856	78,317
Experimental	394,274	162,880
Fusion	3916	545
Neighborhood	119,195	11,489
Text mining	693,904	177,847

The formulas for the different metrics are as follows:

$$Precision = \frac{Lp}{500} \tag{9}$$

$$MAE(A,E) = \frac{1}{n}\sum_{i=1}^{n}|A_i - E_i| \tag{10}$$

$$MRAE(A,E) = \sum_{i=1}^{n}\frac{|A_i - E_i|}{E_i} \tag{11}$$

$$RMSE(A,E) = \sqrt{\frac{\sum_{i=1}^{n}(A_i - E_i)^2}{n}} \tag{12}$$

One of the significant challenges in evaluating our model is the variability in scores across different versions of the STRING database. Scores between versions, such as STRING 11.0 and 11.5, can change due to the addition of new interactions or modifications in the computation of certain subscores, such as those derived from text mining. Consequently, even the raw data exhibits discrepancies between versions. These changes introduce inherent errors, making it difficult to make highly accurate predictions of combined-scores for novel predicted PPIs. This issue is illustrated in the heatmap shown in Fig. 5, highlighting the differences in scores between STRING 11.0 and 11.5.

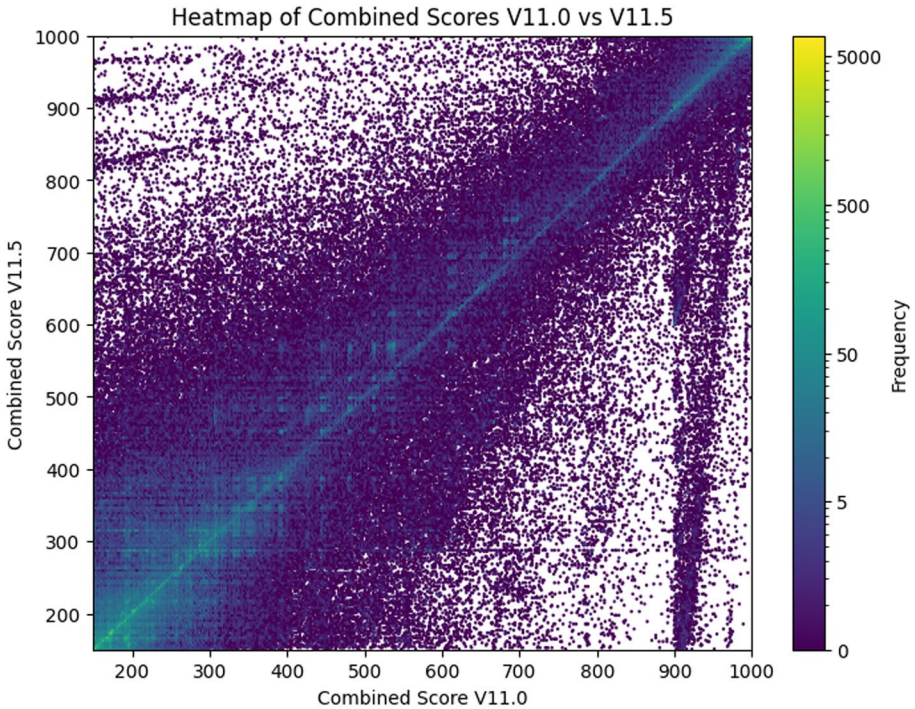

Fig. 5 Heatmap of the correlation between the combined score of String v11.0 and String v11.5

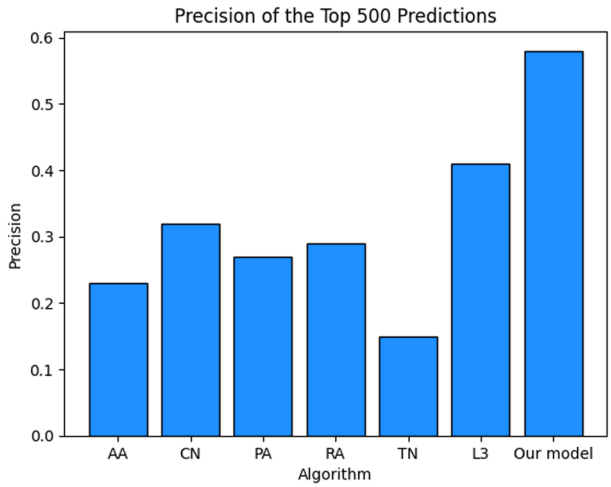

Fig. 6 Precision of the Top500 Predictions across different representations

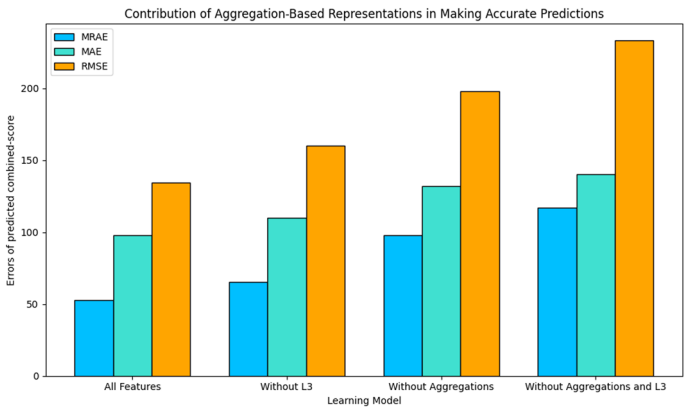

Fig. 7 Ablation study: contribution of aggregation-based representations in making accurate predictions

Regarding the Top500 precision, our results, illustrated in Fig. 6, show that our regression model tree outperforms all traditional similarity measures. Additionally, the L3 measure performs better than topological measures based on paths of length 2. Using these two representations, in addition to topological measures, could contribute to predicting missing PPI links with minimal error.

In order to evaluate the contribution of our aggregation-based representations in predicting accurate missing PPIs, we conducted an ablation study. The contribution of the nine aggregation functions was examined, with AVG-MIN identified as the most impactful. However, since all functions contribute to the overall performance, the evaluation of the complete set of aggregation-based measures was conducted to ensure a comprehensive analysis. Figure 7 shows that removing L3 increases *MAE* by 15.3% and *RMSE* by 25.2%. However, removing the aggregation representations increases *MAE* by 32.9% and *RMSE* by 52.1%. Moreover, removing L3 and the various aggregation representations increases *MAE* by 64.2% and *RMSE* by 85.4%. Additionally, the results of the *MRAE* metric presented in Fig. 7 show that the difference between the predicted value and the actual value increases significantly when

Table 4 MRAE of predicted scores accros STRING channels

Channel	Without aggregations and L3	Without aggregations	Without L3	ALL features
Combined-score	117.01	98.00	65.19	52.88
Coexpression	88.30	61.98	44.29	33.12
Cooccurence	89.14	69.30	57.11	48.10
Database	132.76	119.21	89.51	77.31
Experimental	114.56	89.14	63.99	54.16
Fusion	44.91	34.99	24.83	12.55
Neighborhood	79.10	54.6	33.29	23.44
Text mining	137.10	111.21	99.29	86.49

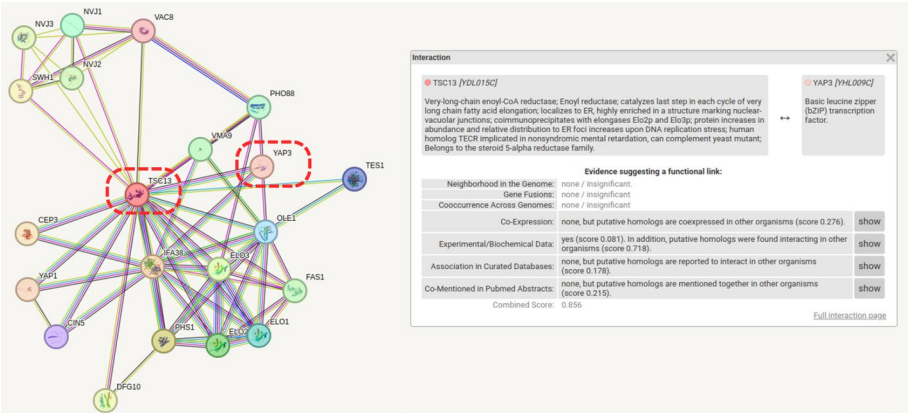

Fig. 8 Example of a predicted PPI using our regression tree model, added in STRING v12.0

removing our aggregation representations. Thus, although the contribution of the aggregation representations to learning is more significant than that of L3, using L3 improves the performance of our model.

The contribution of using our aggregation-based measures in learning was also confirmed for individual channels. The results of our experiments illustrated in Table 4 show that using all measures (ALL Features) yields the closest score to the value added in version 11.5 of STRING.

Similar results were found across all individual channels. This is illustrated in Table 4, which shows the errors we found when predicting missing links based only on individual channels. It is important to note that the value of errors for a channel also depends on the distribution of individual scores. For example, the coexpression scores are usually low in STRING, so the predicted values are also low, which led to low errors.

As an example of a newly predicted PPI, the two proteins 4932.*YDL*015*C* (TSC13) and 4932.*YHL*009*C* (YAP3) were not directly connected in STRING versions 11.0 and 11.5. Despite this, our regression tree model identified them as relevant candidate proteins. Although this pair was not included in STRING version 11.5, an interaction between them has been added in the latest version, STRING 12.0, as illustrated in Fig. 8 extracted from STRING. This example demonstrates the efficacy of our

predictive model: it correctly identified a potential interaction that was subsequently validated in a later version of STRING. It also underscores the dynamic nature of PPI networks, where new interactions can be discovered and incorporated over time. Therefore, predictions made using earlier versions of STRING, such as 11.0, may not appear immediately in subsequent versions like 11.5, but can emerge in later updates, reflecting the continuous advancements in our understanding of PPIs.

Discussion

The experiments conducted in this work confirm the hypothesis that using paths of length 3 for predicting missing links in PPI networks is more effective than using paths of length 2. In terms of Top500 precision, the use of the L3 measure allows for the prediction of missing PPIs with greater accuracy compared to all traditional topological measures based on paths of length 2. Furthermore, incorporating paths of length 3 and their confidence scores into our aggregation-based learning approach achieves a precision that surpasses even the L3 measure alone. This demonstrates that leveraging the confidence scores carried by paths of length 3 significantly enhances prediction precision. Regarding the error between the predicted scores and the actual scores added in subsequent versions, our results indicate that our graph-based machine learning model can predict scores for missing interactions with values relatively close to those added in version V11.5. Given the dynamic nature of the STRING database, where scores can change between versions due to new data or revised computations, our approach may be affected by false positives in version V11.0 that are corrected in version V11.5. Despite this, we have demonstrated the robustness of our model. Additionally, the results from our ablation study reveal that all representations used in the model are informative. However, the presence of the L3 measure and our aggregation-based representations significantly contribute to minimizing prediction error. The study underscores the importance of using these advanced representations to capture the complex relationships within PPI networks accurately.

Conclusion

In this work, we have developed a novel graph-based machine learning model that leverages paths of length 3 and aggregation-based representations to predict missing protein–protein interactions (PPIs) with high precision. Our results demonstrate that using paths of length 3, particularly the L3 measure, significantly enhances prediction accuracy compared to traditional topological measures based on paths of length 2. Furthermore, incorporating confidence scores from these paths into our model has shown to improve precision even further. Our model's ability to predict scores for missing interactions with values closely matching those added in subsequent STRING database versions underscores its robustness and reliability. Despite the inherent challenges posed by the dynamic nature of the STRING database, such as changes in scores between versions, our approach has proven to be effective and resilient. The findings from our ablation study highlight the importance of both topological measures and aggregation-based representations in capturing the complex relationships within PPI networks. The significant contribution of the L3 measure and aggregation representations to minimizing prediction error further validates our approach. While our study presents promising results,

several avenues for future research can be explored to further enhance and validate our model:

- Apply the model to PPI datasets from other organisms to assess its generalizability and robustness across different biological contexts.
- Incorporate other types of biological data, such as gene expression profiles, protein 3D structures and functional annotations, to enrich the feature set and potentially improve prediction accuracy.
- Develop a hybrid approach that leverages the graph structure of PPIs and their associated confidence scores, while also incorporating additional data about the proteins themselves, such as sequence information, post-translational modifications, and interaction domains. This combined approach could more effectively identify relevant missing PPIs by utilizing both network-based and protein-specific information.
- The performance of our method could still be further improved. For instance, employing a distributed computing framework, similar to the one presented in [50], could significantly enhance the scalability and efficiency of our approach when dealing with larger datasets.

In conclusion, our graph-based machine learning model represents a significant step forward in predicting accurate PPIs. By integrating diverse data sources and leveraging advanced representations, our approach offers a powerful tool for expanding and refining PPI datasets, ultimately contributing to a deeper understanding of the intricate web of protein interactions.

Acknowledgements
Not applicable.

Author contributions
K.C. and J.T. contributed their expertise in the field of biology, providing explanations for many of the concepts presented in this work and validating the results obtained. H.A., G.F., and N.L. proposed the novel graph-based approach for predicting new protein–protein interactions and conducted all the experiments described in this study. The manuscript was primarily written by H.A., with all authors reading, improving, and approving the final version.

Funding
Not applicable.

Availability of data and materials
The yeast Saccharomyces cerevisiae data was obtained from STRING database and can be downloaded from https://string-db.org/. The code for the method WPPN is available in a Git repository: https://git.unistra.fr/icube_sdc/WPPNs

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 31 July 2024 Accepted: 31 October 2024
Published online: 08 November 2024

References

1. Yang F, Fan K, Song D, Lin H. Graph-based prediction of protein-protein interactions with attributed signed graph embedding. BMC Bioinform. 2020;21:1–16.

2. Braun P, Gingras A-C. History of protein-protein interactions: from egg-white to complex networks. Proteomics. 2012;12(10):1478–98. https://doi.org/10.1002/pmic.201100563.
3. Keskin O, Tuncbag N, Gursoy A. Predicting protein-protein interactions from the molecular to the proteome level. Chem Rev. 2016;116(8):4884–909. https://doi.org/10.1021/acs.chemrev.5b00683.
4. Berggård T, Linse S, James P. Methods for the detection and analysis of protein-protein interactions. Proteomics. 2007;7(16):2833–42. https://doi.org/10.1002/pmic.200700131.
5. Nooren IM, Thornton JM. Diversity of protein-protein interactions. EMBO J. 2003;22(14):3486–92. https://doi.org/10.1093/emboj/cdg359.
6. Chang C-K, Lin S-M, Satange R, Lin S-C, Sun S-C, Wu H-Y, Kehn-Hall K, Hou M-H. Targeting protein-protein interaction interfaces in covid-19 drug discovery. Comput Struct Biotechnol J. 2021;19:2246–55. https://doi.org/10.1016/j.csbj.2021.04.003.
7. Bakail M, Ochsenbein F. Targeting protein-protein interactions, a wide open field for drug design. C R Chim. 2016;19(1–2):19–27. https://doi.org/10.1016/j.crci.2015.12.004.
8. Stumpf MP, Thorne T, De Silva E, Stewart R, An HJ, Lappe M, Wiuf C. Estimating the size of the human interactome. Proc Natl Acad Sci. 2008;105(19):6959–64. https://doi.org/10.1073/pnas.0708078105.
9. Felgueiras J, Silva JV, Fardilha M. Adding biological meaning to human protein-protein interactions identified by yeast two-hybrid screenings: a guide through bioinformatics tools. J Proteom. 2018;171:127–40. https://doi.org/10.1016/j.jprot.2017.05.012.
10. Chandrasekharan G, Unnikrishnan M. High throughput methods to study protein-protein interactions during host-pathogen interactions. Eur J Cell Biol. 2024;103(2): 151393. https://doi.org/10.1016/j.ejcb.2024.151393.
11. Lenz S, Sinn LR, O'Reilly FJ, Fischer L, Wegner F, Rappsilber J. Reliable identification of protein-protein interactions by crosslinking mass spectrometry. Nat Commun. 2021;12(1):3564.
12. Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, et al. The biogrid interaction database: 2019 update. Nucleic Acids Res. 2019;47(D1):529–41. https://doi.org/10.1093/nar/gky1079.
13. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, et al. The string database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51(D1):638–46. https://doi.org/10.1093/nar/gkac1000.
14. Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. String: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005;33:433–7. https://doi.org/10.1093/nar/gki005.
15. Zhou T, Lü L, Zhang Y-C. Predicting missing links via local information. Eur Phys J B. 2009;71:623–30.
16. Xian L, Wang Y. Advances in computational methods for protein-protein interaction prediction. Electronics. 2024;13(6):1059. https://doi.org/10.3390/electronics13061059.
17. Kumar A, Singh SS, Singh K, Biswas B. Link prediction techniques, applications, and performance: a survey. Phys A. 2020;553: 124289. https://doi.org/10.1016/j.physa.2020.124289.
18. Tang T, Zhang X, Liu Y, Peng H, Zheng B, Yin Y, Zeng X. Machine learning on protein-protein interaction prediction: models, challenges and trends. Brief Bioinform. 2023;24(2):076. https://doi.org/10.1093/bib/bbad076.
19. Sarkar D, Saha S. Machine-learning techniques for the prediction of protein-protein interactions. J Biosci. 2019;44(4):104.
20. Soleymani F, Paquet E, Viktor H, Michalowski W, Spinello D. Protein-protein interaction prediction with deep learning: a comprehensive review. Comput Struct Biotechnol J. 2022;20:5316–41. https://doi.org/10.1016/j.csbj.2022.08.070.
21. Liu L, Zhu X, Ma Y, Piao H, Yang Y, Hao X, Fu Y, Wang L, Peng J. Combining sequence and network information to enhance protein-protein interaction prediction. BMC Bioinform. 2020;21:1–13.
22. Zhang M, Li P, Xia Y, Wang K, Jin L. Revisiting graph neural networks for link prediction (2020)
23. Muzio G, O'Bray L, Borgwardt K. Biological network analysis with deep learning. Brief Bioinform. 2021;22(2):1515–30.
24. Kewalramani N, Emili A, Crovella M. State-of-the-art computational methods to predict protein-protein interactions with high accuracy and coverage. Proteomics. 2023;23(21–22):2200292. https://doi.org/10.1002/pmic.202200292.
25. Luo X, Wang L, Hu P, Hu L. Predicting protein-protein interactions using sequence and network information via variational graph autoencoder. IEEE/ACM Trans Comput Biol Bioinf. 2023;20(5):3182–94. https://doi.org/10.1109/TCBB.2023.3273567.
26. Yang Y, Su X, Zhao B, Li G, Hu P, Zhang J, Hu L. Fuzzy-based deep attributed graph clustering. IEEE Trans Fuzzy Syst. 2024;32(4):1951–64. https://doi.org/10.1109/TFUZZ.2023.3338565.
27. Yang J, Zhang X-D. Predicting missing links in complex networks based on common neighbors and distance. Sci Rep. 2016;6(1):1–10.
28. Adamic LA, Adar E. Friends and neighbors on the web. Social Networks. 2003;25(3):211–30. https://doi.org/10.1016/S0378-8733(03)00009-1.
29. Newman ME. Clustering and preferential attachment in growing networks. Phys Rev E. 2001;64(2): 025102. https://doi.org/10.1103/PhysRevE.64.025102.
30. Yao L, Wang L, Pan L, Yao K. Link prediction based on common-neighbors for dynamic social network. Proc Comput Sci. 2016;83:82–9. https://doi.org/10.1016/j.procs.2016.04.102.
31. Jaccard P. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. Bull Soc Vaudoise Sci Nat. 1901;37:241–72.
32. Liu H, Kou H, Yan C, Qi L. Link prediction in paper citation network to construct paper correlation graph. EURASIP J Wirel Commun Netw. 2019;2019(1):1–12. https://doi.org/10.1186/s13638-019-1561-7.
33. Wang P, Xu B, Wu Y, Zhou X. Link prediction in social networks: the state-of-the-art. Sci China Inf Sci. 2015;58(1):1–38. https://doi.org/10.48550/arXiv.1411.5118.
34. Kovács IA, Luck K, Spirohn K, Wang Y, Pollis C, Schlabach S, Bian W, Kim D-K, Kishore N, Hao T, et al. Network-based prediction of protein interactions. Nat Commun. 2019;10(1):1–8. https://doi.org/10.1038/s41467-019-09177-y.

35. Rolland T, et al. A proteome-scale map of the human interactome network. Cell. 2014;159(5):1212–26. https://doi.org/10.1016/j.cell.2014.10.050.
36. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, et al. The string database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49(D1):605–12. https://doi.org/10.1093/nar/gkab835.
37. Al Hasan M, Chaoji V, Salem S, Zaki M. Link prediction using supervised learning. In: SDM06: Workshop on Link Analysis, Counter-terrorism and Security, vol. 30, pp. 798–805 (2006). https://doi.org/10.1016/j.isci.2020.101626
38. Gao F, Musial K, Cooper C, Tsoka S. Link prediction methods and their accuracy for different social networks and metrics. Sci Programm. 2015. https://doi.org/10.1155/2015/172879.
39. Du X, Sun S, Hu C, Yao Y, Yan Y, Zhang Y. Deepppi: boosting prediction of protein-protein interactions with deep neural networks. J Chem Inf Model. 2017;57(6):1499–510. https://doi.org/10.1021/acs.jcim.7b00028.
40. Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinform. 2017;18:1–8.
41. Hashemifar S, Neyshabur B, Khan AA, Xu J. Predicting protein-protein interactions through sequence-based deep learning. Bioinformatics. 2018;34(17):802–10. https://doi.org/10.1093/bioinformatics/bty573.
42. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci. 2007;104(11):4337–41. https://doi.org/10.1073/pnas.0607879104.
43. 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(9):3025–30. https://doi.org/10.1093/nar/gkn159.
44. You Z-H, 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(5):0125811. https://doi.org/10.1371/journal.pone.0125811.
45. Mirabello C, Wallner B. Interpred: a pipeline to identify and model protein-protein interactions. Proteins Struct Funct Bioinf. 2017;85(6):1159–70.
46. Jha K, Saha S, Singh H. Prediction of protein-protein interaction using graph neural networks. Sci Rep. 2022;12(1):8360.
47. Maheshwari S, Brylinski M. Across-proteome modeling of dimer structures for the bottom-up assembly of protein-protein interaction networks. BMC Bioinform. 2017;18:1–14.
48. Jha K, Saha S. Amalgamation of 3d structure and sequence information for protein-protein interaction prediction. Sci Rep. 2020;10(1):19171.
49. Li S, Huang J, Zhang Z, Liu J, Huang T, Chen H. Similarity-based future common neighbors model for link prediction in complex networks. Sci Rep. 2018;8(1):1–11.
50. Hu L, Yang S, Luo X, Yuan H, Sedraoui K, Zhou M. A distributed framework for large-scale protein-protein interaction data analysis and prediction using mapreduce. IEEE/CAA J Autom Sin. 2022;9(1):160–72. https://doi.org/10.1109/JAS.2021.1004198.

Publisher's Note