Arabian Journal for Science and Engineering (2026) 51:5837–5852
https://doi.org/10.1007/s13369-025-10498-y

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

CanGRNNA: Ensembling Message Passing and Temporal Dynamics with Attention for Prediction of Structural Cancer Protein–Protein Interactions

Rafiya Jan[1] · Ahsan Hussain[1] · Assif Assad[1] · Basharat Bhat[2]

Received: 12 August 2024 / Accepted: 13 July 2025 / Published online: 1 August 2025

Abstract
Protein–protein interactions (PPIs) are critical in various biological processes, such as signal transduction, immunological responses, and cellular pathways. Structural PPI prediction leverages computational models to predict interactions reliably using protein structures. Understanding comprehensive cancer PPIs is essential for uncovering the precise molecular processes that drive different diseases and advancing specific therapeutics. Nevertheless, precisely predicting cancer PPIs is still a challenge due to the intricate and dynamic nature of the proteins. The research presents a novel approach called CanGRNNA that integrates the graph structure and the temporal patterns of proteins using graph recurrent neural networks (GRNN) with attention for accurate prediction of cancer PPIs. Attention identifies the most pertinent residues and context of protein structures to enhance the precision of interaction predictions. The approach captures the structural knowledge of proteins with long-range dependencies by attention-enhanced GRNN to more accurately represent the complex interactions within protein complexes. The results indicate that the proposed approach outperforms existing state-of-the-art techniques. The approach represents a significant advancement in the computational cancer prediction of PPIs, providing a reliable method for researchers to detect possible interaction sites and better understand protein functions. The importance of the research lies in its potential to expedite the identification of novel pharmacological targets and therapeutic approaches, ultimately leading to progress in precision medicine and personalised healthcare.

Keywords Protein–protein interaction (PPI) · GRNN (graph recurrent neural network) · Cancer proteins · GNN (graph neural network)

1 Introduction

PPI is vital in almost all cellular activities by regulating crucial processes like cell cycle regulation, signal transmission, and metabolic activities. PPI is critical for understanding the cancer molecular basis and processes. PPIs are essential for comprehending the molecular processes of cancer. PPI prediction is a challenging task with significant consequences for biological investigation and healthcare.

Cancer proteins are pivotal in cancer's initiation, advancement, and spread. These proteins are classified as oncogenes, tumour suppressors, and proteins involved in cell cycle regulation. Oncogenes are typically genes that usually function, but when they undergo mutations or become overexpressed, they stimulate unregulated cell growth and survival. Tumour suppressors often restrain cell proliferation and enhance programmed cell death, thereby maintaining the integrity of the genome. Genetic mutations or functional impairments

✉ Rafiya Jan
rafiyajan2012@gmail.com; rafiya.jan@iust.ac.in

Ahsan Hussain
ahsan.hussain@islamicuniversity.edu.in

Assif Assad
assif.assad@islamicuniversity.edu.in

Basharat Bhat
basharat@skuastkashmir.ac.in

[1] Department of Computer Science and Engineering, Islamic University of Science and Technology, Awantipora, Pulwama, J&K 192122, India

[2] Centre for Artificial Intelligence and Machine Learning, Sher-e-Kashmir University of Agricultural Sciences and Technology Kashmir, Shalimar, Srinagar, J&K 190025, India

in these proteins can result in the breakdown of regular regulatory processes, enabling cells to proliferate and divide without restraint. Furthermore, proteins like BRCA1 play a crucial role in the repair of DNA. Mutations in these proteins can lead to genomic instability, which in turn promotes the progression of cancer. Predicting PPI in cancer is crucial for comprehending the intricate signalling networks that control cellular activities and how their disruption contributes to cancer development. By mapping these relationships, researchers pinpoint crucial nodes and pathways vital for cancer cells' survival and rapid growth. This knowledge uncovers biomarkers that are used for early diagnosis and prognostic assessment. Hence, research on cancer PPI not only deepens the comprehension of cancer biology but also propels the invention of precise therapeutics that boost patient outcomes. In the past, PPI identification mainly relied on experimental methods such as yeast two-hybrid screening, co-immunoprecipitation, and mass spectrometry. These procedures yield valuable data, but are expensive and require significant time, and the scope is generally limited. Due to the intricate and varied nature of the proteome, there is a need for more effective and adaptable methodologies, resulting in a transition towards computational methods. Computational prediction approaches, such as docking simulations and sequence-based algorithms, provide expedited and cost-efficient means to speculate probable interactions. Nevertheless, these approaches may occasionally exhibit imprecision due to their dependence on simplified models and restricted contemplation of the dynamic characteristics of proteins.

PPI prediction is a challenging task with significant consequences for biological investigation and healthcare. Conventional experimental techniques often prove slow and expensive, leading to the emergence of computational methods. Machine learning methods, particularly those utilising deep learning and graph-based representations, have demonstrated potential in improving the precision and effectiveness of PPI prediction. However, these approaches are premised on sequence and expression data and do not encompass the intricate, multi-dimensional and structural features for precise PPI prediction. Recent breakthroughs in deep learning have greatly improved the precision and effectiveness of PPI predictions. Deep learning algorithms, mainly graph neural networks (GNNs) [1], prove to be better for protein structures' intricate and multi-dimensional features and their interactions. GNNs represent proteins as graphs, with nodes representing amino acids and edges representing interactions between them. GNNs capture the proteins' local and global structural information. Researchers achieve great precision in predicting novel PPIs by training these networks on extensive datasets containing known protein structures and interactions. This method not only speeds up the identification of possible therapeutic targets but also offers an understanding of the molecular processes of cancer. It opens up opportunities for creating focused treatments that can disrupt crucial protein interactions and contribute to the advancement of cancer.

The paper presents a comprehensive approach highlighting the benefits of utilising structural data and showcases the effectiveness of employing GRNN with attention called Cancer CanGRNNA to predict cancer PPIs. The research presents a novel technique that combines the graph structure and sequential dependencies of proteins using a GRNN-based model. The model utilises recurrent neural networks (RNNs) in the graph architecture to effectively capture temporal patterns and long-term dependencies in protein sequences and enhance the accuracy of cancer PPIs' prediction. The research explores the application of GRNN with attention to predicting cancer-related PPIs by utilising the structural data of proteins. The paper presents an integrated strategy, highlights the benefits of employing structural data using the GRNN, and illustrates the effectiveness of CanGRNNA for PPI. This paper showcases the efficacy of GNNs in improving the precision of PPI predictions in cancer research, hence facilitating the development of more efficient therapeutic approaches.

The subsequent sections discuss literature survey, the process of curation of the cancer PPI dataset, and the CanGRNNA methodology. Following this, the evaluation of the proposed model, results, comparative analysis, applications, limitations, case study, and conclusion are explained in subsequent sections

2 Background

Proteins are extensive and intricate molecules that perform critical tasks in the body. Proteins are organic substances composed of carbon, hydrogen, nitrogen, oxygen, and sulphur atoms of one or more polypeptide chains. Amino acids, the building blocks of proteins, are connected through peptide bonds in a precise order, forming a polypeptide chain that folds into a distinct three-dimensional structure. The exact arrangement and conformation of the polypeptide chain are determined by the genetic code encoded in DNA. Protein synthesis, the process of protein formation, consists of two primary stages: transcription, which involves copying a segment of DNA into messenger RNA (mRNA), and translation, which involves ribosomes in the cytoplasm reading the mRNA to assemble amino acids in the correct sequence to create the protein. Proteins are indispensable for nearly all biological processes and essential for enormous cellular functions.

Figure 1 shows the systematic review method for cancer PPI prediction, specifying inclusion, exclusion criteria, and

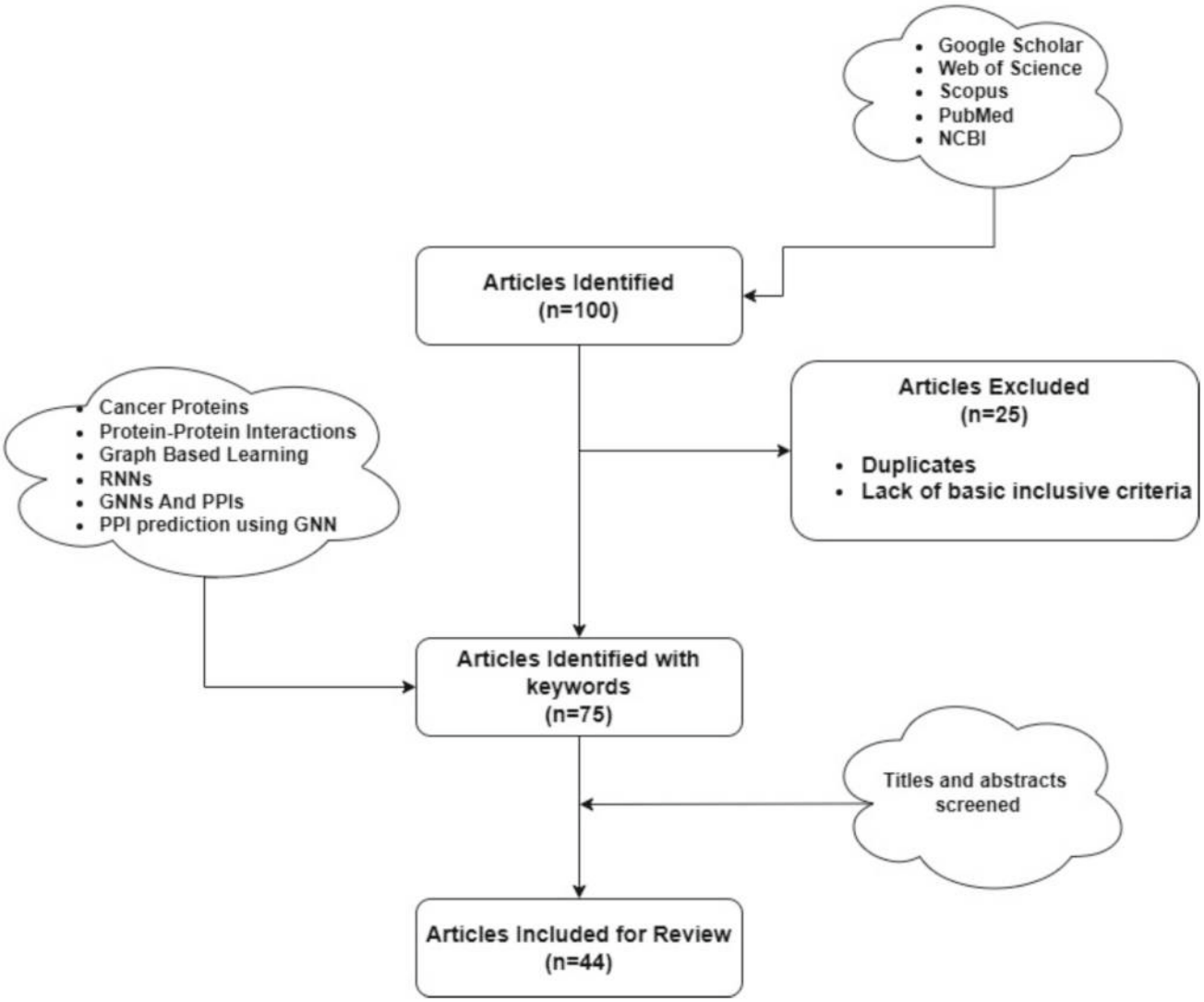

Fig. 1 Systematic review process for cancer PPI prediction

screening steps for including 44 research articles in review process after excluding irrelevant papers.

The research emphasises on cancer PPIs as the modelling substrate and represents a novel and underexplored approach, as most prior studies in this domain tend to generalise across all protein interactions without focusing on disease-specific mechanisms. The research seeks to improve the predictive models' biological significance and translational capacity by concentrating on cancer PPIs. Figure 2 provides an overview of the literature review process for cancer PPI prediction including data sources, approaches, evaluation metrics, and comparative analysis.

The further subsections include some of the state-of-the-art technologies that have a profound impact on the prediction of the PPIs in the below subsections:

2.1 Computational and Experimental Methods

Computational and experimental methods are employed to predict the PPIs. Understanding these interactions comprehensively is crucial for unravelling the intricate signalling networks in cancer cells, pinpointing possible biomarkers, and formulating targeted therapeutics. By predicting PPIs, scientists systematically analyse the complex protein associations contributing to cancer-causing processes. It enables to investigate innovative therapeutic approaches aimed at disrupting these interactions. The acquisition of PPI network data offers a detailed insight into intricate cellular processes in biological systems [2, 3] and is crucial in the identification of drugs and the development of therapies [4, 5]. Over the past few decades, there have been ongoing experimental efforts to identify PPIs on a wide scale in model organisms [6]. PPI's computational algorithms utilise the amino

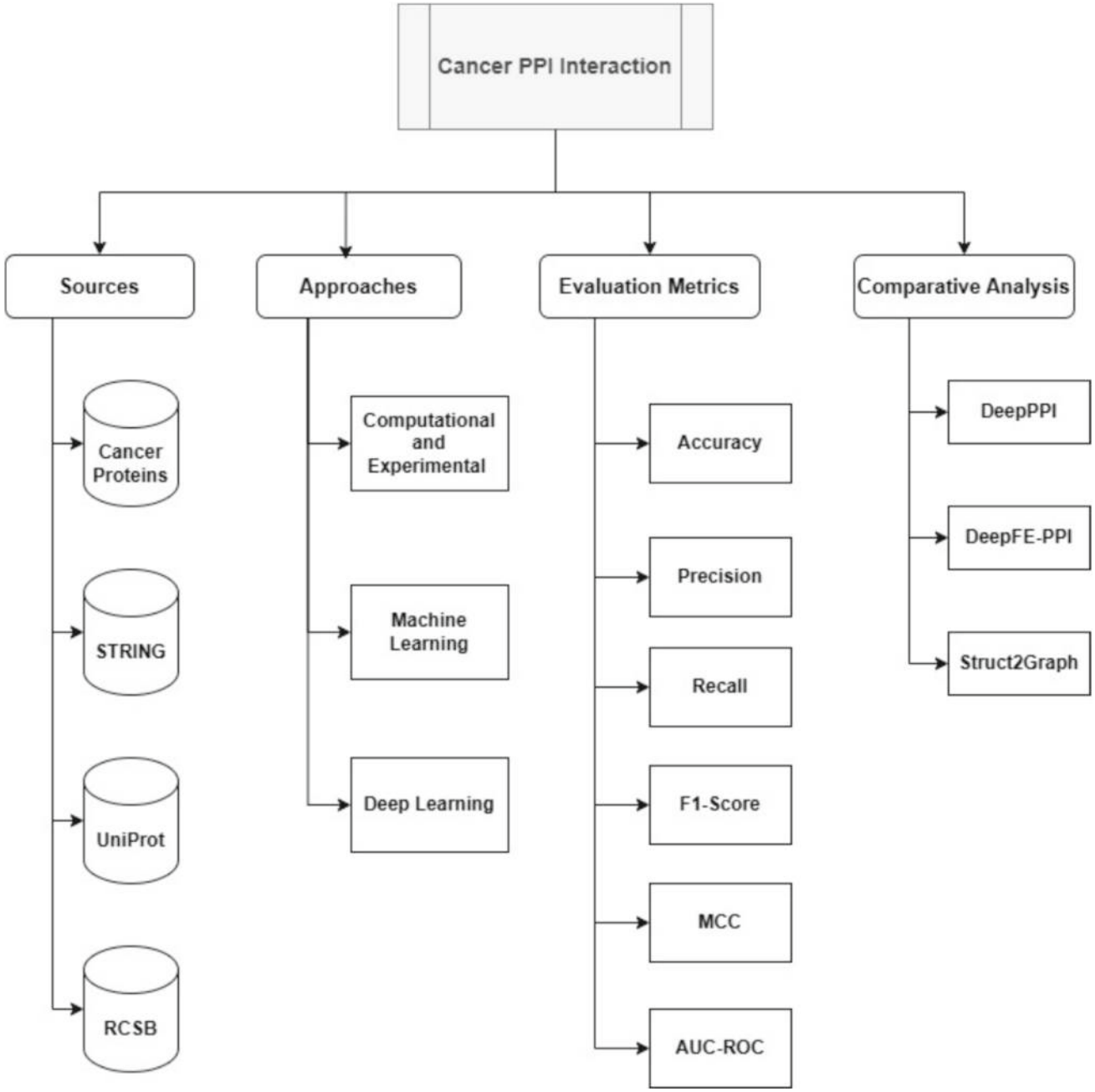

Fig. 2 Summary of literature survey: cancer PPI prediction

acid sequences of proteins to analyse and identify the interactions. The computational methods employ the statistical characteristics of proteins. Validating the PPI empirically in wet labs requires significant financial resources, time, and staffing. Identifying PPI sites allows for the development of new medications, the exploration of the uncharted roles of proteins, and the prediction of adverse effects of drugs. While commonly considering the benchmark, the experimental identification of PPIs is contingent upon precise experimental conditions, resulting in frequently restricted coverage. Due to their time-consuming and labour-intensive nature, experimental methods are supplemented by in silico approaches, such as machine learning (ML)-based methods, which have gained popularity in providing testable hypotheses [7, 8]. In the past decades, computational approaches have investigated biological processes and facilitated progress in developing novel pharmaceuticals and therapies.

2.2 ML Methods

For several decades, ML methods are used for PPI analysis [9]. ML approaches use hidden aspects of existing PPIs to predict new interactions. These models frequently rely on similarity criteria, which presume that proteins with a shared interaction partner should have comparable characteristics. These variables often represent the physicochemical characteristics of amino acid sequences, structural similarity, evolutionary identity, PPI network partners, or topological traits [10–12]. Among them, protein sequence features are the most widely studied and favoured. Amino acid sequences primarily determine the actions of proteins, as they serve as the main structural components of proteins. Given the abundance of protein sequences, numerous research endeavours employed protein sequence features to forecast PPIs. The ML approaches are categorised into three types: supervised learning (decision trees, Naive Bayes, artificial neural networks (ANNs), and support vector machines (SVMs)), unsupervised learning (including K-means clustering), and reinforcement learning. Martin et al. [13] created a distinctive molecular descriptor to encode protein sequences to predict PPIs using a SVM classifier. Shen et al. [14] introduced the conjoint triad (CT) descriptors as a concise way to characterise the composition of amino acid sequences. Due to the inadequacy of CT descriptors in capturing the long-range interactions of residues, which are crucial for describing PPIs, [15] devised an auto-covariance encoding technique to account for the impacts of neighbouring residues. Additional coding schemes based on component frequency, such as composition-transition distribution and composition of k-spaced amino acid pairs (CKSAAP), are also extensively utilised in predicting protein–protein interactions [16, 17]. PIPE2 tool calculates the similarity in polypeptide sequences between query proteins and known PPIs to assess if two proteins interact [18].

While sequence-based methods show more effectiveness, then encodings of protein sequences that interact with each other alone are required to capture all critical information related to PPI. Evolutionary profiles of sequences and structures offer further characteristics that go beyond sequence composition, enabling more robust prediction of PPIs. PSSMs (position-specific scoring matrices) represent the interacting proteins [19], whereas [20] employed evolutionary profiles, which demonstrate improved prediction performance and resilience. Therefore, computational methods are being utilised to determine PPIs accurately.

2.3 Deep Learning (DL) Methods

DL methods evaluate the functional consequences of sequence alterations. Deep mutational scans utilise different protein function evaluations, ranging from hundreds of thousands to even millions. These scans provide insights into the structural limitations imposed by a protein's inherent features and function. DL techniques are utilised to forecast PPIs [21–23]. Multilayer perceptrons and convolutional neural networks precisely forecast PPIs by including protein sequence attributes [24–26]. Natural language processing techniques are also employed to efficiently transform amino acid sequences into high-dimensional vectors for predicting PPIs [27, 28]. DL models are collaboratively utilised to exploit their respective advantages more effectively. The approach based on convolutional and RNNS to forecast PPIs allows to capture the locally important features and sequence characteristics from the primary protein sequences [29]. However, current prediction methods primarily rely on sequence information and need to consider the critical role of protein structural properties. Proteins carry out their tasks by adopting 3D structures, which enable them to interact with other molecules in the 3D environment. To achieve this objective, an effective method is devised for predicting PPI by initially looking for a complex template that matches the query protein using sequence and structural alignment and then a Bayesian classifier is used to predict the probability of contact [12, 30]. The main challenge in incorporating protein structural features into PPI predictions has been the limited availability of accurate, large-scale protein structures. However, the recent advancement of AlphaFold [31] made it possible to predict protein monomer structures from protein sequences with accuracy similar to experimental methods. This development provides a way to consider protein structures in predicting PPIs. Protein 3D structures pose a more significant challenge in feature extraction than linear sequences, primarily because of their intricate topologies. One common approach to this problem is to transform protein structures into residue networks or graphs, where residues are treated as nodes and residue connections are treated as edges.

2.4 Graph Neural Network (GNN)

The graph convolutional network (GCN) is a prevalent deep learning model that captures the structural relationships inside graph-structured data. GCN has found extensive application in protein bioinformatics for tasks such as predicting protein binding interfaces, annotating protein functions, and discovering drugs [32]. A GCN model is employed to extract characteristics from protein pocket and ligand graph representations, and impressive results are acquired on widely used virtual screening benchmark datasets, demonstrating the model's competitiveness [33]. The combination of GCN and a natural language model predicts protein activities based on computationally derived structures [34]. In a recent study, a GCN-based model forecasts the probable residues to interact with other proteins [35].

Struct2Graph [36], a graph attention network, reliably predicts the PPIs based on 3D structural data and provides insights regarding the crucial residues involved in these interactions. [37] employs GCN and graph attention networks (GAT) to effectively predict PPIs based on protein sequence and structure data and demonstrate better performance. (SGPPI) [38], a PPI prediction model called structure and graph-based predictions of protein interactions (SGPPI) using a GCN, examine the overall structural properties of proteins and the specific structural attributes of patches located at putative protein interaction interfaces to understand the structural patterns of PPIs. In addition, SGPPI included the evolutionary profiles in the structural representation of PPIs to enhance their performance. CollaPPI [39] is a collaborative learning framework that improves information sharing at protein and task levels. CollaPPI demonstrates improved performance compared to current approaches on PPI benchmarks and exhibits robust generalisation in supplementary tasks.

In this research, a PPI prediction model called CanGRNNA using a graph-based RNN model with attention is developed. The model utilises RNNs in the graph architecture to effectively capture temporal patterns and long-term dependencies in protein sequences. The research presents a novel technique that combines the graph structure and sequential dependencies of cancer proteins using a GRNN-based model. Moreover, attention is incorporated in CanGRNNA in the structural representation of PPIs to enhance the performance. CanGRNNA considers the overall structural features of proteins to comprehend both the local and global structural patterns of proteins.

3 Methodology

The research presents a novel method called CanGRNNA that ensembles the graph structure and sequential dependencies of proteins using a GRNN-based method with attention. Graph GRNN integrates the strengths of graph-based and recurrent models to effectively process the fundamental graph structures and sequential interactions of proteins. GRNNs are suitable for predicting PPIs as they effectively comprehend the complex graph structure of protein networks and the sequential characteristics of protein sequences. The model utilises RNNs in the graph architecture to effectively capture temporal patterns and long-term dependencies in protein structure and enhance predictions' accuracy. Attention enables the ability to dynamically assign varying priority levels to distinct features and interactions, resulting in more accurate predictions that are easier to interpret. Attention enhances the model's ability to identify intricate patterns that conventional aggregation approaches in GNNs may overlook by emphasising the most influential interactions. It leads to improved efficiency and a better understanding of the factors that influence PPIs. The overall methodology of CanGRNNA is shown in Fig. 3.

3.1 Cancer PPI Dataset

The cancer PPI data [40] is acquired from the publicly available STRING database [41]. The research employs a consistent and biologically meaningful selection technique to remove low-confidence interactions and standardise cancer protein identifiers. The cancer PPI dataset acquired from String database is mapped using the UniProt [42] database by mapping each protein with the protein accession ID. Each UniProt accession ID is mapped to the RCSB [43] database to obtain the pdb protein structure.

A single UniProt ID may correlate to several PDB entries due to experimental conditions and changes in protein conformation. As a result, numerous cross-references to PDB files are present, indicating the availability of diverse structures for the cancer proteins. The problem of proteins with multiple PDB entries corresponding to a single UniProt ID is addressed by selecting the PDB entry with the highest resolution for each UniProt ID, as it corresponds to the most accurate structural representation. When several high-resolution structures are available, the one that thoroughly covers the UniProt sequence is selected. Not all proteins are completely crystallised in every PDB file; therefore, arbitrarily selecting a PDB file may lead to omitting information regarding the protein's binding site. PDB files are organised according to the length of chains and high resolution, and thus, obtaining the most comprehensive structural target protein is important. Each protein's crystal structure's chain length and resolution are sourced from the RCSB database. The curated balanced PPI dataset comprises positive and negative interactions with 4434 interactions. The cancer PPI dataset curated in this research comprises protein pairings that are annotated with interaction labels "+" and "−" where "+" indicates positive interactions and "−" indicates negative interactions. Organelle-specific proteins do not interact with each other, which forms the basis for curating negative interactions. Each protein is denoted by its adjacency matrix and a sequence of fingerprint attributes. The process of curation of cancer PPI dataset is shown in Fig. 4

The curated dataset provides substantial benefits in coverage and quality compared to numerous existing datasets and includes experimentally validated enhancing the reliability of the training and evaluation processes, leading to more precise cancer PPI predictions.

3.2 CanGRNNA

The research presents a novel approach called CanGRNNA that integrates the graph structure, and the temporal pat-

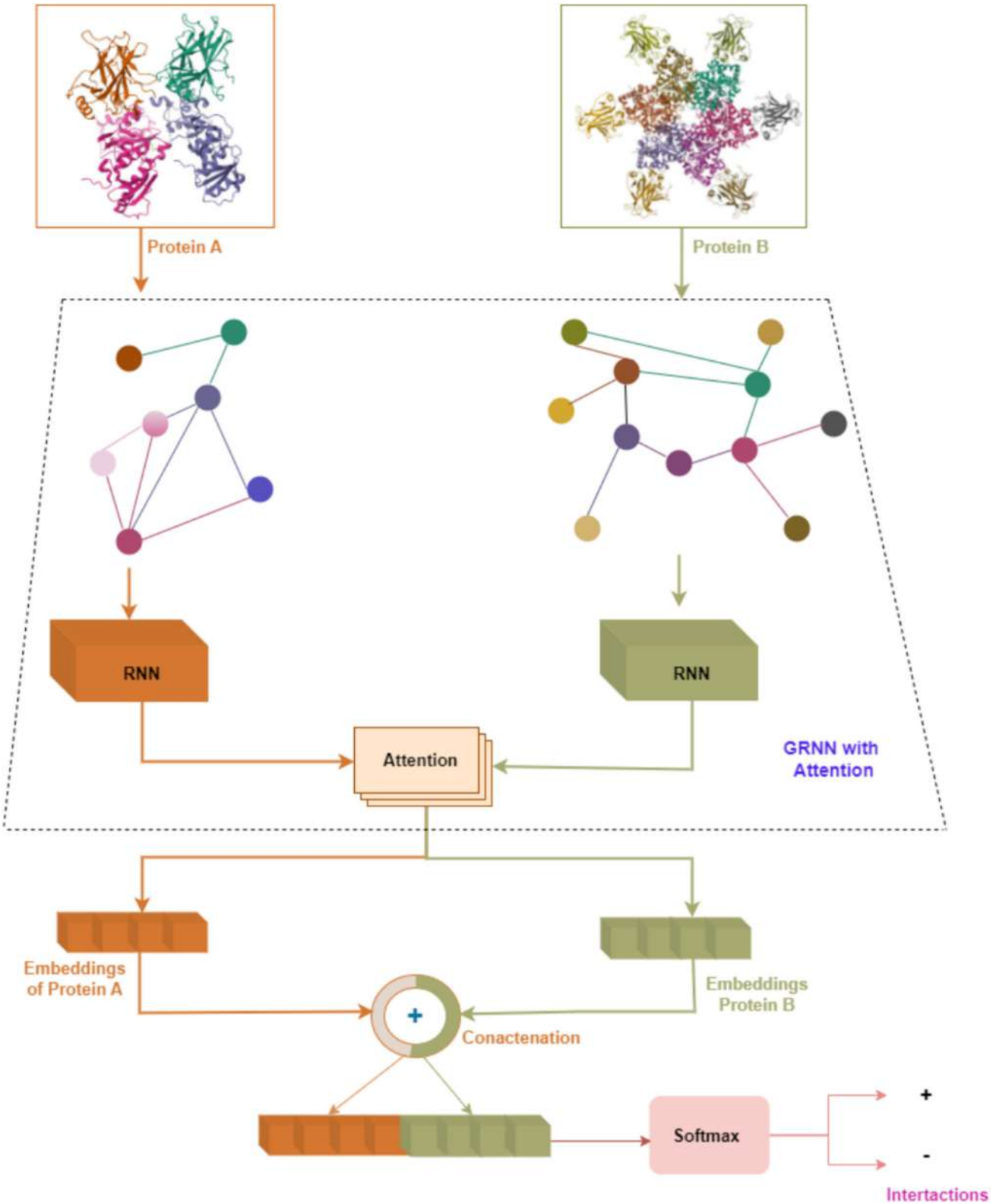

Fig. 3 Overall methodology of CanGRNNA

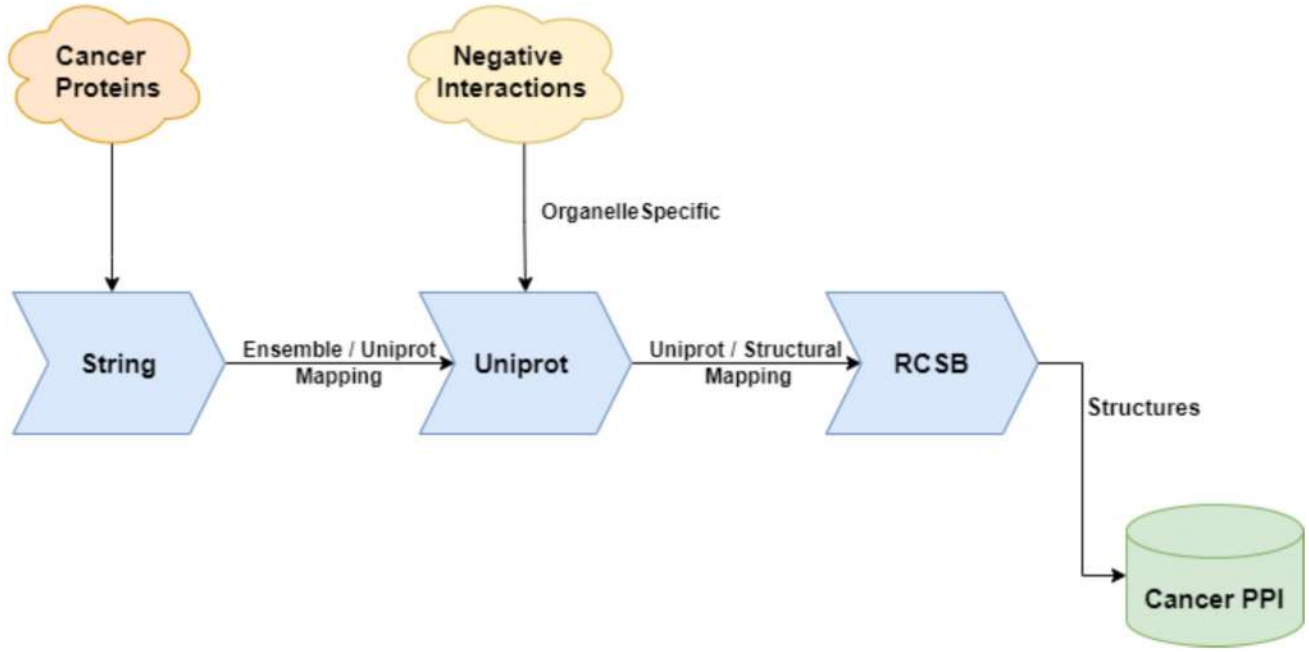

Fig. 4 Curation of cancer PPI dataset

terns of structural proteins using RNNs with attention for prediction of cancer PPIs. RNNs identify patterns in data sequences and possess connections that create directed cycles, enabling them to retain a 'memory' of past inputs. RNNs capture temporal dynamics and long-range relationships in sequences and thus are highly suitable for tasks requiring context or sequential dependencies. RNNs offer an efficient way to model these interdependencies and improve PPI prediction accurately. The GRNN uses graph-based and sequential aspects to create a comprehensive representation of proteins structural and sequential features, improving PPI predictions' accuracy. The CanGRNNA model utilises 3D structural data of proteins obtained from PDB files as the principal input for training and assessment. It analyses residue-level information to create residue-level spatial graphs, with each node representing an amino acid residue and edges indicating spatial proximity based on a distance threshold. The model generates categorical fingerprints for each residue using a Weisfeiler–Lehman-like graph kernel, which captures local subgraph structures by aggregating neighbourhood information. Besides structural context, the model preserves sequence-level information by translating residue types into one-letter amino acid codes via a predetermined mapping. The codes are integrated in conjunction with the structural graph data. CanGRNNA integrates 3D spatial and sequential representations to get a comprehensive contextual encoding at the residue level, which is crucial for precisely modelling cancer PPIs. The model comprises several layers that sequentially interpret the data, enabling it to acquire knowledge and identify significant characteristics from intricate cancer PPI dataset. The embedding layer (E_l) is the first step in the CanGRNNA model, converting protein fingerprints into vector representations. A protein fingerprint is a residue-level representation of a protein, wherein each residue is allocated a categorical identifier based on structural or sequence-derived characteristics. The identifiers are mapped using a predetermined fingerprint dictionary and embedded into dense vectors through a Embedding layer. The resultant vectors function as node features in a GNN facilitating efficient learning of interaction patterns. The fingerprinting approach enables the GNN to aggregate the information throughout the protein graph efficiently and facilitates the subsequent mutual-attention process for simulating protein interactions. The conversion is essential as it transforms discrete protein aspects into a continuous space, enabling the model to carry out the operations and learn patterns effectively. The embeddings encapsulate crucial data of each protein, serving as a basis for the subsequent layers to further develop. The output of the embedding (E_l) for protein (P_A) with fingerprint F_{P_A} is represented as

$$(E_l) = \text{Embedding}(F_{P_A})$$

GNNs play a crucial role in the GRNN architecture by allowing the model to leverage the structural connections across proteins. The residues are regarded as nodes rather than whole proteins. Each protein is depicted as a graph in which each node represents a residue, and the features of the nodes are generated from fingerprints at the residue level, which are embedded into dense vectors. The connections between

nodes are represented by adjacency matrices, which are precomputed and loaded during the training and testing phases. It enables the GNN to represent local residue–residue interactions within the protein structure The GNN updates the hidden state of each node by aggregating information from its adjacent nodes, thus effectively capturing the graph's topology.

$$h_{v_p}^{(k)} = \sigma\left(W^{(k)} h_v^{(k-1)} + \sum_{u_p \in N(v_p)} W^{(k)} h_{u_p}^{(k-1)} + b^{(k)}\right) \quad (1)$$

where $h_{v_p}^{(k)}$ is the hidden state of node v_p at layer k, $N(v_p)$ represents the neighbours of v_p, $W^{(k)}$ and $b^{(k)}$ are layer-specific weights and biases, and σ is an activation function. The iterative approach allows the GNN to acquire complex and hierarchical representations of the graph's structure.

CanGRNN incorporates RNNs to address the sequential properties of protein data effectively. Proteins consist of sequences of amino acids, and their specific arrangement and context significantly influence their functions and interactions. RNNs excel at capturing the sequential relationships in data, making them an excellent choice for modelling protein sequences.

The hidden state h_t at time step t in an RNN is represented as:

$$h_t = \sigma\left(W_h h_{t-1} + W_x x_t + b\right) \quad (2)$$

where h_t is the hidden state at time step t, h_{t-1} is the hidden state from the previous time step, x_t is the input at time step t, W_h and W_x are weight matrices, b is the bias term, and σ is the activation function. The RNN's capacity to retain information from past inputs enables it to effectively represent extended relationships and sequential patterns in protein sequences. The output O_t at time step t is given by:

$$y_t = W_y h_t + c \quad (3)$$

where W_y is the weight matrix for the output and c is the output bias. This approach allows CanGRNN to efficiently capture the sequential information essential for PPI prediction.

Including the attention mechanism in the GRNN enhances its capabilities by enabling the model to concentrate on the most pertinent aspects of the input data. Within the realm of PPI prediction, attention mechanisms enhance the model's ability to assess the significance of various protein characteristics and their interconnections, thereby enhancing both the comprehensibility and effectiveness of the model. Attention enables the ability to dynamically assign varying priority levels to distinct features and interactions, resulting in more accurate predictions that are easier to interpret. Attention enhances the model's ability to identify intricate patterns that conventional aggregation approaches in GNNs may overlook by emphasising the most influential interactions. It leads to enhanced efficiency and a more comprehensive comprehension of the factors that influence PPIs.

The equation that combines the GRNN with an attention mechanism called CanGRNNA for predicting PPI is formulated as follows:

$$h_{v_p}^{(k)} = \sigma\left(W^{(k)} {h_{v_p}}^{(k-1)} + \sum_{u_p \in N(v_p)} \alpha_{u_p v_p}^{(k)} W^{(k)} {h_{u_p}}^{(k-1)} + b^{(k)}\right) \quad (4)$$

where the attention coefficient $\alpha_{u_p v_p}^{(k)}$ is calculated as:

$$\alpha_{u_p v_p}^{(k)} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top [W^{(k)} {h_{v_p}}^{(k-1)} \,||\, W^{(k)} {h_{u_P}}^{(k-1)}]\right)\right)}{\sum_{k \in N(v_p)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top [W^{(k)} {h_{v_p}}^{(l-1)} \,||\, W^{(l)} h_k^{(k-1)}]\right)\right)} \quad (5)$$

Here, $h_v^{(k)}$ is the hidden state of node v_p at layer k, $N(v_p)$ represents the neighbours of v_p, $W^{(k)}$ and $b^{(k)}$ are layer-specific weights and biases, and σ is an activation function. The equation represents the node update rule of CanGRNNA, incorporating the attention mechanism. The attention mechanism preferentially highlights critical protein residues and their adjacent structural context within cancer PPI networks. Each node in the graph signifies a protein, with its initial feature vector originating from residue-level embeddings encapsulating biochemical characteristics and spatial aspects of amino acids. To emphasise the most functionally significant residues, a node-level attention layer is integrated that calculates attention scores utilising both residue embeddings and three-dimensional structural information, including C-alpha distances and sequence proximity. The attention score between nodes i and j is calculated using a compatibility function of their hidden states, subsequently using normalisation for interpretability. It enables the model to allocate weights to interaction pairs with greater r structural significance. Additionally, to maintain structural context, edge features are incorporated to that denote physical closeness and the probability of domain–domain interactions, which influence the message-passing phase. The attention mechanism dynamically prioritises essential residues and structural motifs linked to cancer activities, enhancing prediction accuracy and interpretability.

The CanGRNNA model is designed to elucidate intricate relationships within cancer-specific PPI networks. CanGRNNA fundamentally enhances the conventional GRNN framework by integrating recurrent memory and graph-based contextualised attention mechanisms. The model employs

curated cancer PPI dataset comprising proteins that are differentially expressed, mutated, or implicated in the cancer. The model acquires interaction patterns and structural dependencies specific to the cancer through training on this data, improving its predictive accuracy and biological significance. Significantly, CanGRNNA utilises 3D structural pdb data files to construct intricate residue-level or domain-level graphs. These structural graphs represent atomic or residue closeness, enabling the model to comprehend the spatial configuration of cancer proteins. Integrating a 3D structure facilitates more accurate edge delineations in the graph and allows the model to identify conformational characteristics or binding sites essential for cancer. Consequently, CanGRNNA utilising PDB-based graphs enhance focused drug development, biomarker identification, and pathway-level insights pertinent to the pathology under investigation. The architecture initiates at the residue level, which is succeeded by graph building, and then multi-head attention module enables the model to focus concurrently on multiple interaction attributes, enhancing the model's robustness and interpretability. In contrast to traditional GRNNs that utilise uniform message transmission throughout the graph, CanGRNNA implements node-level attention weights that emphasise interaction signals pertinent to carcinogenic pathways. The recurrent layers in CanGRNNA are designed to maintain long-range dependencies in highly coupled PPI subgraphs, especially those abundant in cancer hallmarks.

4 Results and Discussion

The CanGRNNA model is trained using cross-entropy loss and optimised using the Adam optimiser. The dataset is split into training and test sets using five-fold cross-validation via the KFold class, incorporating shuffling and a fixed random seed to guarantee repeatability and better performance. The dataset is split into five mutually exclusive subsets, with each fold utilised once as the test set and the other four for training. The splitting is performed at the level of protein pairings instead of individual proteins, guaranteeing that no identical interaction pairs are present in both the training and test sets. To mitigate the risk of data leaking, it is confirmed that identical protein pairs do not appear across folds and that adjacency matrices and fingerprint encodings are rigorously separated for training and testing inside each fold. Upon meticulously re-evaluating the data pipeline, we affirm it is affirmed that no leakage exists between the training and test sets. The assessment measures, characterised by elevated recall and precision, accurately represent the model's performance under appropriate cross-validation conditions. Standard performance measures are calculated, including accuracy, precision, recall, F1-score, and ROC-AUC. The evaluation metrics are calculated using true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{F1-score} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{9}$$

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{10}$$

The model surpasses conventional techniques and achieves the highest level of performance in PPI prediction tests.

The bar graph Fig. 5 depicts the occurrences of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) during 30 epochs. The statistical analysis of the graph demonstrates that the model exhibits remarkable performance across all attributes of the confusion matrix. The performance ensures that the model consistently maintains high accuracy, precision, recall, and specificity during training, thereby achieving a balance between sensitivity and specificity. The consistent stability of these metrics over the epochs indicates that the model is not affected by overfitting or underfitting and maintains a consistently high level of performance during the whole training period.

Figure 6 depicts the performance metrics of a model during 30 epochs, with a specific emphasis on accuracy, recall, specificity, and precision. The research demonstrates that the model's accuracy begins at 0.94 and quickly increases by epoch 5, remaining stable with slight variations afterwards. It indicates that the model attains a significant degree of overall accuracy, recall, specificity, and precision at an early stage of the training process and maintains it consistently and accurately, recognises all positive cases from this point on, detects the majority of negative instances with little deviation and suggests a strong level of dependability in the model's optimistic prediction.

The Matthews correlation coefficient (MCC) and F1-score demonstrate swift convergence, attaining elevated values within the second epoch as shown in Fig. 7. The rapid enhancement reflects the model's efficacy in acquiring PPI patterns from the dataset. During subsequent periods, both measures remain consistent, with the MCC changing slightly. This indicates that the model is resilient and does not suffer from overfitting or underfitting. The constantly high MCC value indicates a robust correlation between the predicted and actual interactions, while the high F1-score demonstrates the model's excellent balance between precision and recall. The performance indicators illustrate the model's

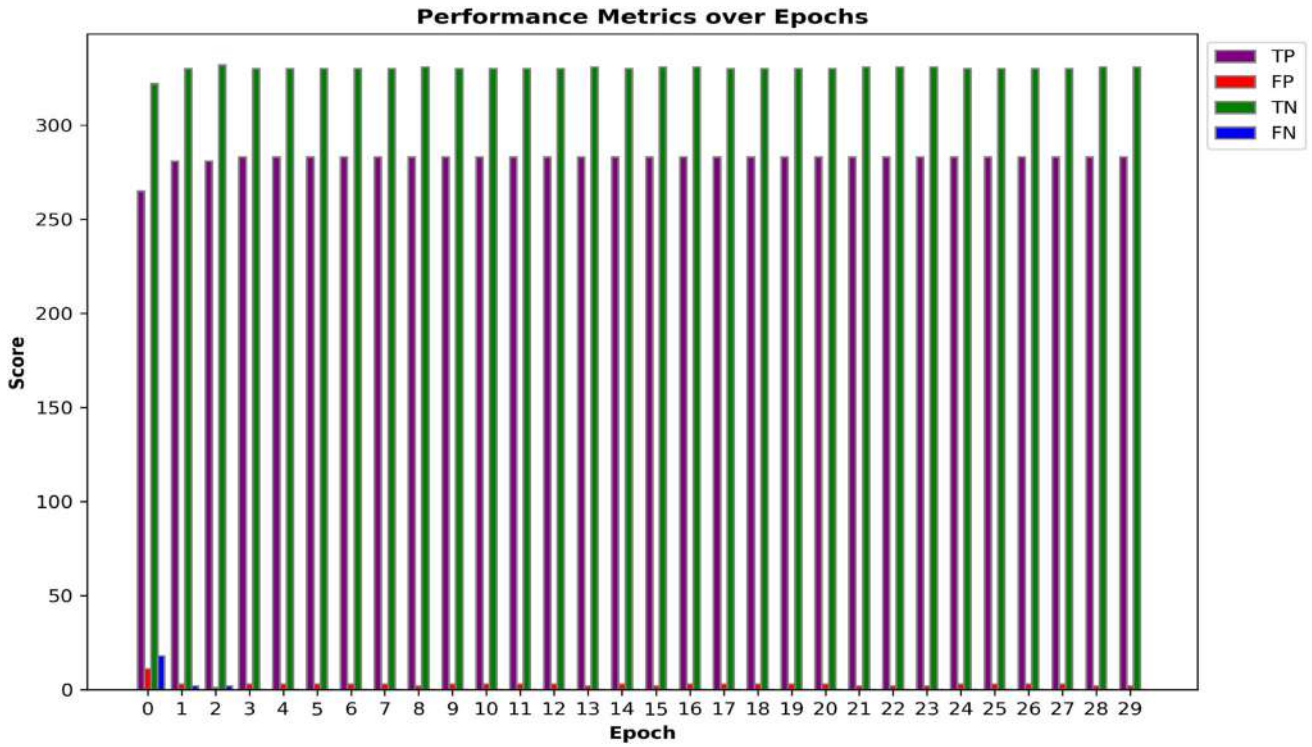

Fig. 5 Evaluation of actual and predicted values over 30 epochs

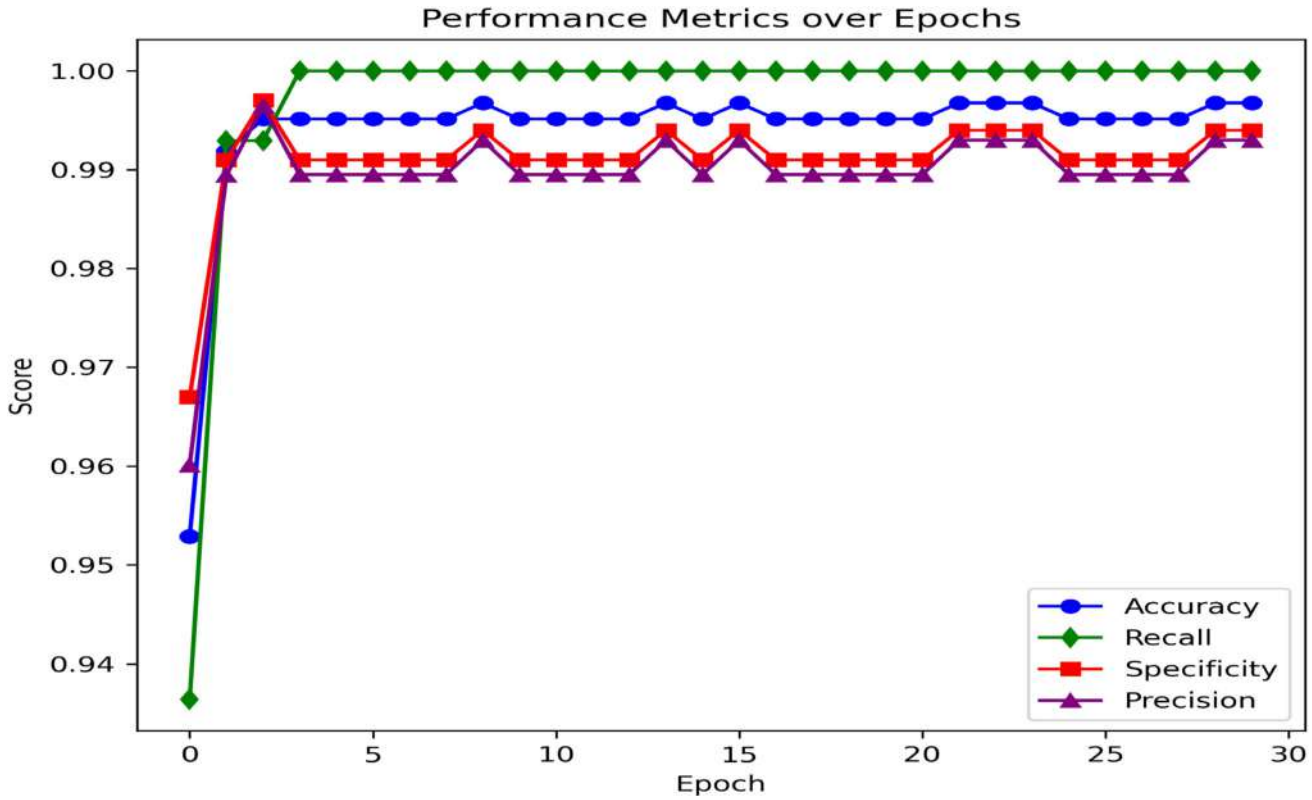

Fig. 6 Evaluation metrics

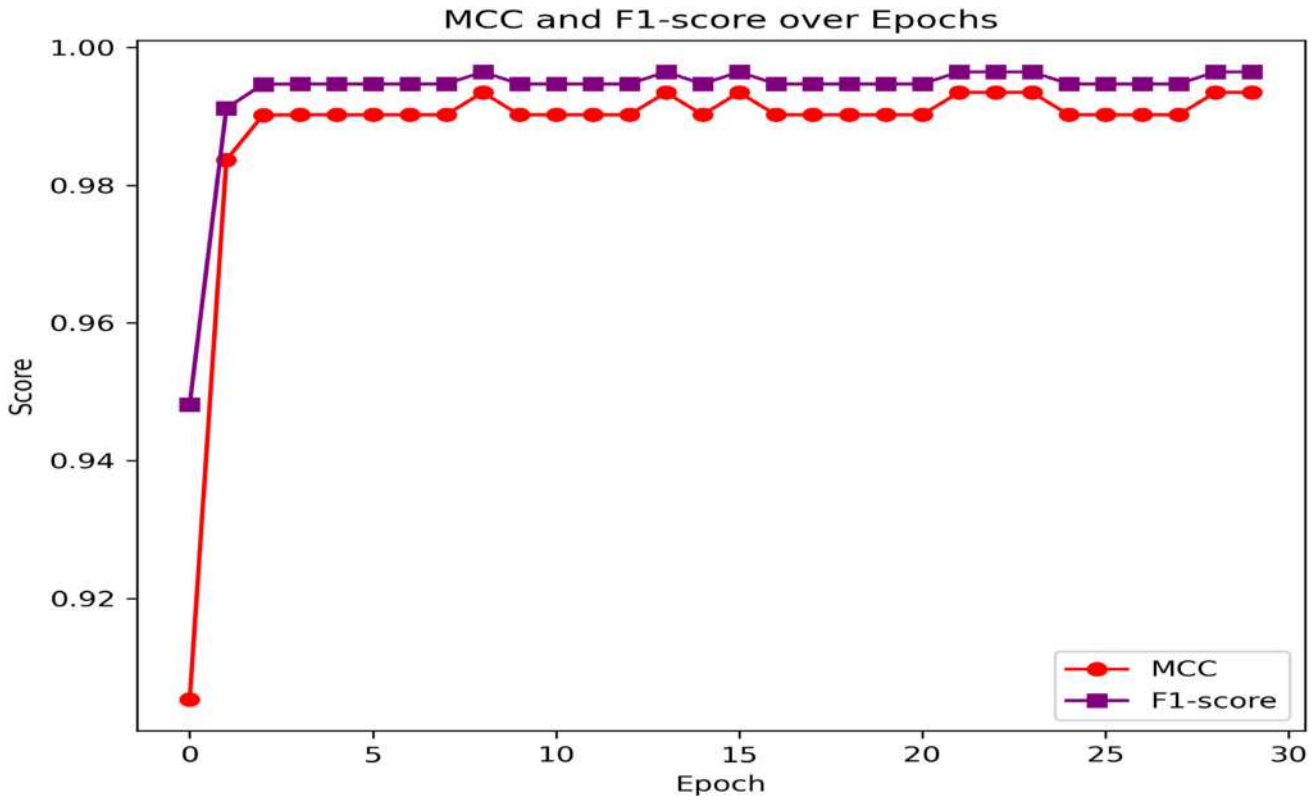

Fig. 7 Evaluation of MCC and F1-scores

exceptional precision and dependability in forecasting protein interactions. The model's efficiency and robustness are demonstrated by its rapid convergence and continuous stability, making it a powerful tool for PPI investigations. In summary, the research highlights the model's ability to offer accurate and dependable predictions, essential for the comprehension of cancer protein interactions.

5 Comparative Analysis

The comparison of the CanGRNNA with the existing state-of-the-art techniques (DeepPPI [24], DeepFE-PPI [44], Struct2Graph [36]) across four performance metrics: MCC, F1-score, Accuracy, and Precision is shown in Fig. 8.

The figure illustrates that the CanGRNNA model exhibits the highest MCC, indicating a superior correlation between predicted and actual classifications. The MCC values indicate that all models have a high level of performance, with CanGRNNA being the most reliable in terms of prediction accuracy. The F1-score, which considers precision and recall, is highest for CanGRNNA, suggesting it has the best balance of precision and recall. The high F1-scores across all models suggest effective handling of accurate positive and false favourable rates, with CanGRNNA outperforming the others. All models exhibit high precision, but CanGRNNA stands out for correctly identifying positive cases with minimal error. The statistical analysis reveals the effectiveness of CanGRNNA in predictive modelling for cancer PPIs, suggesting its suitability for applications requiring high accuracy and precision.

Compared to traditional GNN architectures, CanGRNNA implements an advanced node embedding technique utilising Weisfeiler-Lehman-like graph kernels to represent residues according to their type and spatial neighbourhoods. This substructure-aware fingerprinting markedly enhances its capacity to distinguish minor structural differences essential for PPI prediction. Moreover, by incorporating sequence and structural context, CanGRNNA mitigates the constraints of sequence-only models and attains enhanced reliability, especially for experimentally validated interactions. CanGRNNA demonstrates enhanced performance across various metrics, including accuracy, precision, recall, and MCC. Its capacity to function with high-quality 3D data provides it a unique advantage over models trained exclusively on sequence or interaction graphs from biological networks. CanGRNNA surpasses other sequence or topology-based models and facilitates more accurate and interpretable PPI predictions crucial for applications such as drug development and targeted cancer therapy. In cancer PPI prediction, CanGRNNA demonstrates competitive and frequently superior performance relative to traditional deep learning models,

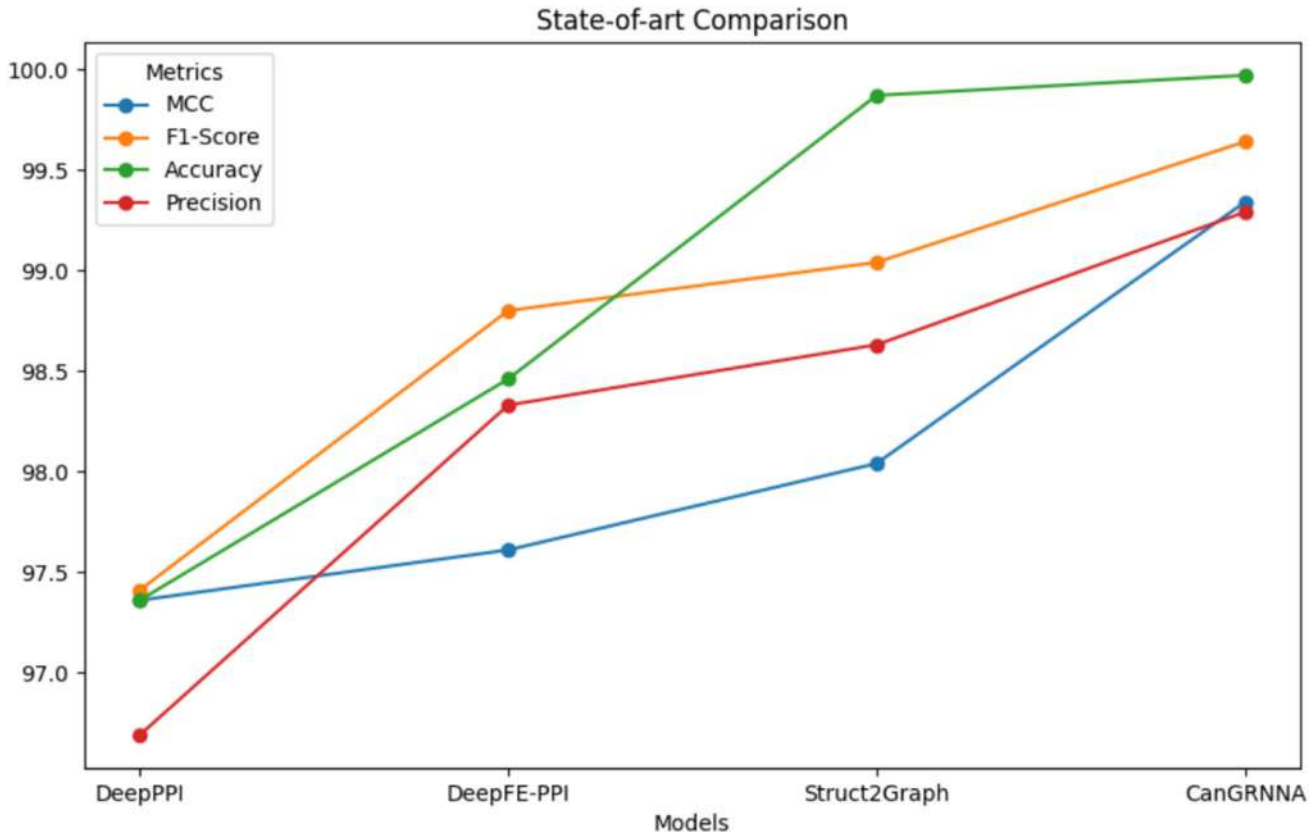

Fig. 8 Comparative analysis with existing methods

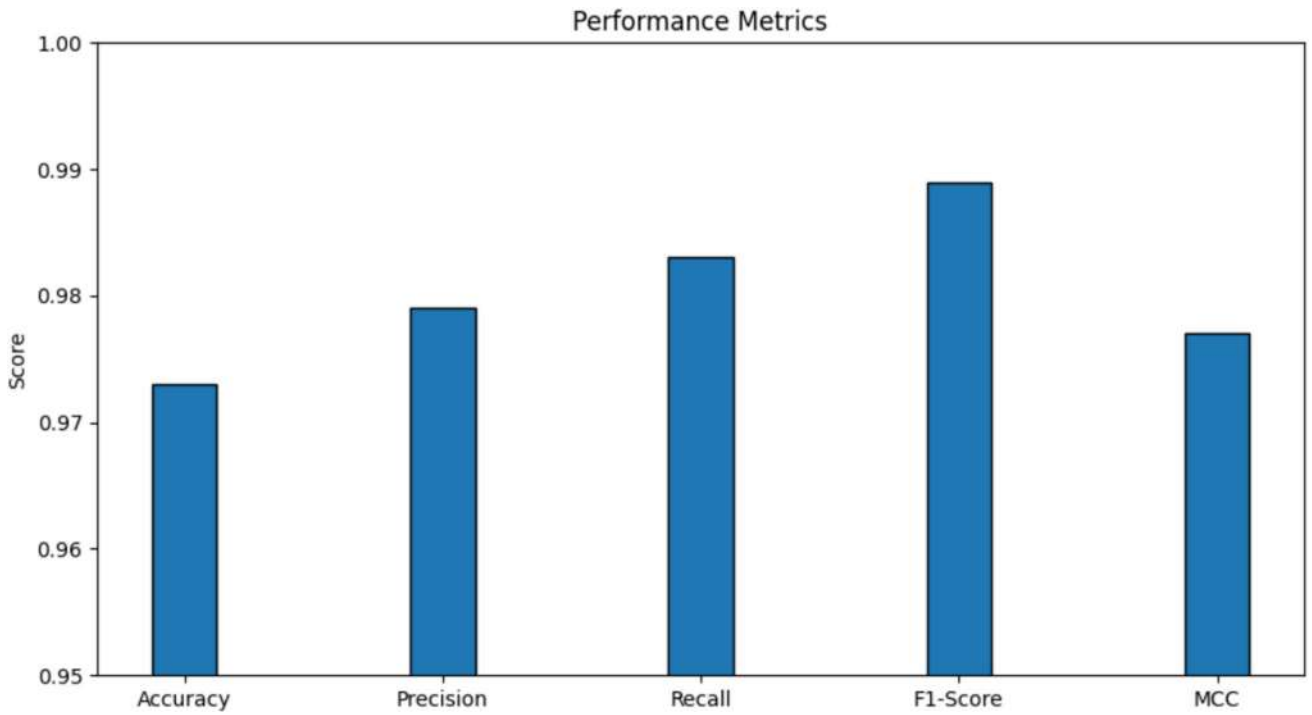

Fig. 9 Case study: evaluation metrics on breast cancer dataset

including CNNs, simple GNNs, and Transformer-based architectures. In contrast to conventional CNNs or sequence-based Transformers that depend mainly on linear amino acid sequences, CanGRNNA utilises residue-level spatial graphs derived from 3D structural data, effectively capturing local and global structural dependencies essential for precisely modelling interaction interfaces.

6 Applications of CanGRNNA

The model considerably enhances the comprehension of biological mechanisms and pathways in cancer by offering a system level perspective on protein interactions and their functional consequences. The model employs curated cancer PPI databases to delineate critical signalling networks frequently dysregulated in cancer. The approach utilises graph-based networks that integrate structural and sequence-based information to emphasise crucial interaction residues and highlight essential amino acid residues implicated in these interactions. The residue-level interpretability explains how particular mutations or structural alterations impair everyday cellular communication, promoting oncogenesis. Furthermore, if multi-omics data are incorporated, the model might prioritise the biomarkers and treatment targets by assessing proteins' centrality or impact within interaction networks. In that case, it enriches the biological context, enabling researchers to associate molecular changes with functional outcomes in signalling networks. The model bridges unprocessed biological data and molecular comprehension, facilitating hypothesis formulation and cancer research validation.

Cancer PPI prediction has significant pharmacological and therapeutic uses, especially within cancer research and treatment. Primarily, these anticipated PPIs can assist in identifying new therapeutic targets, particularly those related to oncogenic drivers or signalling hubs essential for tumour development and survival. The cancer PPIs provide significant insights into drug repurposing and the development of biomarkers, assisting in identifying interaction profiles that differentiate cancer subtypes or correlate with prognosis or treatment. Furthermore, these predictions are incorporated into precision oncology procedures, where patient-specific mutation profiles are aligned with the projected PPI network to evaluate the impact of disruptions on signalling cascades and facilitate individualised treatment strategies and the systematic development of targeted treatments.

7 Limitations of CanGRNNA

The CanGRNNA model, effective in the prediction of cancer PPIs, has few limitations. A significant limitation is its dependence on experimentally derived protein structures, which confines its applicability to proteins with accessible, high-quality structural data. The substantial segment of the proteome, especially inadequately researched or physically ambiguous proteins, is not incorporated in the research, constraining the model's comprehensiveness. The model predicts interactions using static structural snapshots, not capturing the dynamic nature of protein conformations and interactions that may fluctuate based on physiological circumstances, binding events, or post-translational modifications.

8 Case Study: Breast Cancer PPI Prediction

The case study demonstrates the practical application of CanGRNNA in biological fields, explicitly targeting the prediction of PPIs related to breast cancer. The aim is to assess the performance of how structurally informed CanGRNNA architecture, works in a real-world, disease-focused context. The high-confidence dataset of breast cancer-associated proteins, known to play roles in breast cancer pathways, such as BRCA1, BRCA2, TP53, HER2, PIK3CA, AKTI, PTEN, ESR1, and ESR1, with established interactions, is obtained from STRING, UniProt, and RCSB PDB. Only proteins with experimentally determined three-dimensional structures (PDB files) are used. Redundant and low-confidence entries are eliminated, yielding a meticulously curated dataset with extra interactions reserved for independent evaluation. Employing a GNN model grounded in the CanGRNNA architecture, the residue-level protein graphs are developed wherein each node symbolises a residue fingerprint, and edges are formed based on spatial proximity. The model utilises a graph attention method to dynamically ascertain residue-level significance, enabling it to concentrate on biologically relevant areas within the protein structure. The network underwent training with five-fold cross-validation, incorporating early stopping to mitigate overfitting. The structurally informed GNN shows better performance compared to baseline methods. It outperforms a CNN-based model and a standard GNN lacking structural features, achieving an enhanced Precision , Recall, F1-score, and an MCC as shown in Fig. 9.

The results underscore the benefit of incorporating 3D spatial information and disease-specific filtering when modelling PPIs. Notably, the model accurately recovers several held-out PPIs involving known breast cancer regulators, supporting its potential utility in uncovering novel interactions relevant to cancer biology. The case study illustrates that GNNs augmented with structural biology and cancer-specific context can be powerful tools for advancing our understanding of protein interaction landscapes in complex diseases like breast cancer. Such models hold promise for interaction

prediction and guiding applications such as drug target identification and personalised medicine.

9 Conclusion

The research presents a novel approach called CanGRNNA that integrates the graph structure and the temporal patterns of structural proteins using RNNs with attention for prediction of cancer PPIs. The paper presents an integrated strategy, highlights the benefits of employing structural data using the GRNNA, and illustrates the effectiveness of CanGRNNA for PPI. The paper showcases the efficacy of CanGRNNA in improving the precision of PPI predictions in cancer research, hence facilitating the development of more efficient therapeutic approaches. The approach tackles the complexity and dynamic nature of protein interactions by incorporating structural information of proteins and capturing long-range interdependence.

The potential enhancement to improve cancer PPI prediction GNNs is integrating multi-omics and structural interaction features. Integrating this enables the model to apprehend condition-specific interactions and more accurately represent the intricacies of cancer signalling. The further proposed research is to integrate molecular dynamics simulations to address PPIs' conformational flexibility. Incorporating explainable AI techniques might assist in identifying particular residues for interactions, aiding experimental validation and biological interpretation, and confirming predicted PPIs through wet lab studies or high-throughput screening to ensure that computational predictions provide biologically useful findings. These prospective initiatives jointly enhance graph-based cancer PPIs and their public relevance.

References

1. Rout, T.; Mohapatra, A.; Kar, M.: A systematic review of graph-based explorations of PPI networks: methods, resources, and best practices. Netw. Model. Anal. Health Inform. Bioinform. **13**(1), 29 (2024)
2. Bludau, I.; Aebersold, R.: Proteomic and interactomic insights into the molecular basis of cell functional diversity. Nat. Rev. Mol. Cell Biol. **21**(6), 327–340 (2020)
3. Keskin, O.; Tuncbag, N.; Gursoy, A.: Predicting protein-protein interactions from the molecular to the proteome level. Chem. Rev. **116**(8), 4884–4909 (2016)
4. Jubb, H.; Higueruelo, A.P.; Winter, A.; Blundell, T.L.: Structural biology and drug discovery for protein–protein interactions. Trends Pharmacol. Sci. **33**(5), 241–248 (2012)
5. Scott, D.E.; Bayly, A.R.; Abell, C.; Skidmore, J.: Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. Nat. Rev. Drug Discov. **15**(8), 533–550 (2016)
6. Von Mering, C.; Krause, R.; Snel, B.; Cornell, M.; Oliver, S.G.; Fields, S.; Bork, P.: Comparative assessment of large-scale data sets of protein–protein interactions. Nature **417**(6887), 399–403 (2002)
7. Hu, L.; Wang, X.; Huang, Y.-A.; Hu, P.; You, Z.-H.: A survey on computational models for predicting protein–protein interactions. Brief. Bioinform. **22**(5), 036 (2021)
8. Skrabanek, L.; Saini, H.K.; Bader, G.D.; Enright, A.J.: Computational prediction of protein–protein interactions. Mol. Biotechnol. **38**, 1–17 (2008)
9. Sarkar, D.; Saha, S.: Machine-learning techniques for the prediction of protein–protein interactions. J. Biosci. **44**(4), 104 (2019)
10. Bitbol, A.-F.; Dwyer, R.S.; Colwell, L.J.; Wingreen, N.S.: Inferring interaction partners from protein sequences. Proc. Natl. Acad. Sci. **113**(43), 12180–12185 (2016)
11. Zhang, Q.C.; Petrey, D.; Deng, L.; Qiang, L.; Shi, Y.; Thu, C.A.; Bisikirska, B.; Lefebvre, C.; Accili, D.; Hunter, T., et al.: Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature **490**(7421), 556–560 (2012)
12. Kovács, I.A.; 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. **10**(1), 1240 (2019)
13. Martin, S.; Roe, D.; Faulon, J.-L.: Predicting protein-protein interactions using signature products. Bioinformatics **21**(2), 218–226 (2005)
14. 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. **104**(11), 4337–4341 (2007)
15. 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. **36**(9), 3025–3030 (2008)
16. Lian, X.; Yang, S.; Li, H.; Fu, C.; Zhang, Z.: Machine-learning-based predictor of human-bacteria protein-protein interactions by incorporating comprehensive host-network properties. J. Proteome Res. **18**(5), 2195–2205 (2019)
17. Xiao, N.; Cao, D.-S.; Zhu, M.-F.; Xu, Q.-S.: protr/protrweb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics **31**(11), 1857–1859 (2015)
18. Pitre, S.; Hooshyar, M.; Schoenrock, A.; Samanfar, B.; Jessulat, M.; Green, J.R.; Dehne, F.; Golshani, A.: Short co-occurring polypeptide regions can predict global protein interaction maps. Sci. Rep. **2**(1), 239 (2012)
19. Zahiri, J.; Yaghoubi, O.; Mohammad-Noori, M.; Ebrahimpour, R.; Masoudi-Nejad, A.: Ppievo: Protein-protein interaction prediction from PSSM based evolutionary information. Genomics **102**(4), 237–242 (2013)
20. Hamp, T.; Rost, B.: Evolutionary profiles improve protein-protein interaction prediction from sequence. Bioinformatics **31**(12), 1945–1950 (2015)
21. Jothi, R.; Kann, M.G.; Przytycka, T.M.: Predicting protein-protein interaction by searching evolutionary tree automorphism space. Bioinformatics (Oxford, England) **21**(Suppl 1), 241 (2005)
22. Zhang, F.; Song, H.; Zeng, M.; Li, Y.; Kurgan, L.; Li, M.: Deepfunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. Proteomics **19**(12), 1900019 (2019)
23. Hu, X.; Feng, C.; Ling, T.; Chen, M.: Deep learning frameworks for protein–protein interaction prediction. Comput. Struct. Biotechnol. J. **20**, 3223–3233 (2022)
24. 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. **57**(6), 1499–1510 (2017)
25. Yang, X.; Yang, S.; Lian, X.; Wuchty, S.; Zhang, Z.: Transfer learning via multi-scale convolutional neural layers for human-virus protein–protein interaction prediction. Bioinformatics **37**(24), 4771–4778 (2021)

26. Chen, Z.; Zhao, P.; Li, C.; Li, F.; Xiang, D.; Chen, Y.-Z.; Akutsu, T.; Daly, R.J.; Webb, G.I.; Zhao, Q., et al.: iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. Nucleic Acids Res. **49**(10), 60–60 (2021)
27. Sledzieski, S.; Singh, R.; Cowen, L.; Berger, B.: D-script translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. Cell Syst. **12**(10), 969–982 (2021)
28. Sun, T.; Zhou, B.; Lai, L.; Pei, J.: Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinform. **18**, 1–8 (2017)
29. Chen, M.; Ju, C.J.-T.; Zhou, G.; Chen, X.; Zhang, T.; Chang, K.-W.; Zaniolo, C.; Wang, W.: Multifaceted protein-protein interaction prediction based on Siamese residual RCNN. Bioinformatics **35**(14), 305–314 (2019)
30. Zhang, Q.C.; Petrey, D.; Garzón, J.I.; Deng, L.; Honig, B.: Preppi: a structure-informed database of protein-protein interactions. Nucleic Acids Res. **41**(D1), 828–833 (2012)
31. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., et al.: Highly accurate protein structure prediction with alphafold. Nature **596**(7873), 583–589 (2021)
32. Sun, M.; Zhao, S.; Gilvary, C.; Elemento, O.; Zhou, J.; Wang, F.: Graph convolutional networks for computational drug development and discovery. Brief. Bioinform. **21**(3), 919–935 (2020)
33. Torng, W.; Altman, R.B.: Graph convolutional neural networks for predicting drug-target interactions. J. Chem. Inf. Model. **59**(10), 4131–4149 (2019)
34. Gligorijevic, V.; Renfrew, P.; Kosciolek, T.; Leman, J.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B.; Fisk, I.; Vlamakis, H., et al.: Structure-based protein function prediction using graph convolutional networks. Nat. Commun. **12**, 3168 (2021)
35. Yuan, Q.; Chen, J.; Zhao, H.; Zhou, Y.; Yang, Y.: Structure-aware protein-protein interaction site prediction using deep graph convolutional network. Bioinformatics **38**(1), 125–132 (2022)
36. Baranwal, M.; Magner, A.; Saldinger, J.; Turali-Emre, E.S.; Elvati, P.; Kozarekar, S.; VanEpps, J.S.; Kotov, N.A.; Violi, A.; Hero, A.O.: Struct2graph: a graph attention network for structure based predictions of protein-protein interactions. BMC Bioinform. **23**(1), 370 (2022)
37. Jha, K.; Saha, S.; Singh, H.: Prediction of protein–protein interaction using graph neural networks. Sci. Rep. **12**(1), 8360 (2022)
38. Huang, Y.; Wuchty, S.; Zhou, Y.; Zhang, Z.: SGPPI: structure-aware prediction of protein-protein interactions in rigorous conditions with graph convolutional network. Brief Bioinform. **24**(2), 020 (2023). https://doi.org/10.1093/bib/bbad020
39. Ma, W.; Bi, X.; Jiang, H.; Zhang, S.; Wei, Z.: Collappi: a collaborative learning framework for predicting protein-protein interactions. IEEE J. Biomed. Health Inform. **28**(5), 3167–3177 (2024). https://doi.org/10.1109/JBHI.2024.3375621
40. Jan, R.; Hussain, A.; Assad, A.; Bhat, B.: Cancer interactome: an in-silico novel approach for elucidating cancer protein–protein interactions (CPPIs) using structural graph learning augmented with attention. J. Comput. Biophys. Chem. 1–18 (2025)
41. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P., et al.: String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. **47**(D1), 607–613 (2019)
42. Bateman, A.: Uniprot: a universal hub of protein knowledge. In: Protein Science, vol. 28, pp. 32–32. Wiley, Hoboken (2019)
43. Burley, S.K.; Berman, H.M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J.M.; Dutta, S., et al.: Rcsb protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. **47**(D1), 464–474 (2019)
44. Yao, Y.; Du, X.; Diao, Y.; Zhu, H.: An integration of deep learning with feature embedding for protein–protein interaction prediction. PeerJ **7**, 7126 (2019)