Journal of Computational Biophysics and Chemistry

Research

Cite This: *J. Comput. Biophys. Chem.* **2026**, *25* (2), 315–332

Cancer Interactome: An *in-silico* Novel Approach for Elucidating Cancer Protein-Protein Interactions (CPPIs) Using Structural Graph Learning Augmented with Attention

Rafiya Jan[*,‡], Ahsan Hussain[*], Assif Assad[*] and Basharat Bhat[†]

[*]*Department of Computer Science and Engineering, Islamic University of Science and Technology, Awantipora, Jammu and Kashmir, India*

[†]*Center for Artificial Intelligence & Machine Learning, Sher-e-Kashmir University of Agricultural Sciences and Technology Kashmir, Shalimar, Jammu and Kashmir, India*

[‡]Corresponding author. E-mails: *rafiyajan2012@gmail.com; rafiya.jan@iust.ac.in*

ABSTRACT: Protein–protein interactions (PPIs) are intricate and vital components of cellular processes, coordinating the different biological processes. The accurate prediction of PPIs is significant for unraveling the functionality of known cancer proteins, comprehending the underlying oncogenesis and determining the possible therapeutics. Although cancer-specific PPI datasets are crucial, but are currently unavailable. The research addresses the gap by curating a novel structural Cancer PPI (CPPI) dataset encompassing significant cancer proteins to represent an extensive picture of Cancer Protein interactions. The paper introduces an innovative approach integrating graph-based methods, viz: GraphSAGE and GIN with attention, to identify and interpret intricate interactions for accurate prediction of CPPIs. Moreover, incorporating attention mechanisms allows the model to prioritize pertinent information while transmitting messages, improving interpretability and predictive capabilities. The performance of the proposed approach is systematically evaluated, compared, and cross-validated with the existing baseline models. The results demonstrate the enhanced remarkable comparative outcomes and highlight the potential of attention-augmented graph-based learning for providing vital insights into the complex world of CPPIs by capturing the structural features and the sites that regulate CPPIs.

KEYWORDS: Amino acids (AAs); protein–protein interaction (PPI); graph neural network (GNN); GraphSAGE (graph sample and aggregate).

1. INTRODUCTION

Proteins, the versatile macromolecules are essential to almost all the biological processes. The proteins comprise 20 fundamental Amino Acids (AAs), and the position of these AAs is essential in protein sequence and is responsible for building the different millions of proteins in the cellular systems. Proteins serve multiple purposes in and outside of a cell, contributing to the complexity of life. For any biological process, proteins associate with other proteins through physical contact, biochemical reaction or signaling pathways known as Protein–Protein Interaction (PPI).[1] The proteins interact with other proteins through surface amino acids, forming larger protein complexes. Through PPI, proteins perform different functions and activities in cellular systems. Thus, PPIs are the key pillars in the cellular systems. PPI aids in modeling pathways for determining molecular processes and therapeutics.

PPIs unravel the functionality of known proteins, thus predicting PPIs is essential in computational biology for drug discovery and biological processes. In the past few decades, biological networks has significantly advanced the understanding of the associations between molecules.[2] The noncovalent bond between

Received: 25 November 2024
Accepted: 7 April 2025
Published: 3 June 2025

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

the R-side chains of AA sequences is responsible for protein interaction and foldings. The PPI information helps to determine the functions and pathways that aids in drug discovery and therapeutics. Therefore, highly accurate computation prediction approaches premised on exact and factual information can be an effective PPI model to complement laboratory experiments.

In the ever-evolving biomedical research landscape, unraveling the complexities of diseases like cancer is a constant challenge in the dynamic field of biomedical research. The inclination for this research stems from the existing datasets depicting the interactions that do not represent a comprehensive and consistent picture across different types of cancer and an understanding of the critical role of PPIs in the complex machinery of biological activities. Thus there is an imperative need for a standardized curated Cancer PPI (CPPI) dataset. The availability of curated CPPI dataset paves the way for understanding the role of cancers in human health and the development of more precise and efficient treatment. The high-quality curated CPPI dataset is preeminent for the learning approach to Artificial Intelligence (AI).

The PPI is a vital process as it illustrates how the alteration of any amino acid of the protein changes the interaction between two proteins, their characteristics, the interface, and signaling process. Therefore, it is imperative to have automated computational approaches to predict PPIs better. The mathematical graph representation of the PPIs facilitates analyzing the protein networks and hypothesizing some unknown functions of proteins. The graph architecture captures the spatial geometry of the proteins and can be integrated with neural networks to eliminate dimensional problems. The PPI prediction task should consider the 3D structural aspects, spatial residues and domains of proteins. The task of determining the 3D structure of proteins was a hard problem till Alphafold2.[3] The problem of PPI prediction in multichain protein complexes is still a buzz.[4] Thus, predicting PPI should encompass both 3D structural representation and sequence information.

The Graph Neural Network (GNN) has the inherent ability to analyze and learn from the structured data from PPI. In a PPI network, the edges represent the protein interactions, and nodes represent the AAs or residues, thus correlates perfectly with the spatial graphical structure of GNN. By aggregating the information from neighboring nodes and incorporating it into a node's representation, GNNs capture local and global context. GNNs excel at learning expressive nodes and graph-level representations by incorporating sequence, PPI sites, and structural features. Thus, GNN has become an indispensable tool for unraveling the complex network of PPI and thus excels at learning expressive nodes.

The research presents a novel CPPI dataset. The well-curated dataset is crucial for analysis and the formation of the research base. It represents an extensive picture by encompassing the various types of cancer proteins and an understanding of the critical role of PPIs in the complex machinery of biological activities. The availability of a curated CPPI dataset paves the way for understanding the role of cancers in human health and the development of more precise and efficient treatment. The research proposes an innovative architecture combining graph-based methods viz: Graph SAmple and aggreGatE (GraphSAGE) and Graph Isomorphism Network (GIN) with attention to identify and interpret intricate connections for predicting CPPIs. The research aims to provide answers to the following crucial questions:

(1) How the graph-based methods viz: GraphSAGE and GIN effectively determine the critical nodes and edges in the protein interaction graph?
(2) Why is the attention mechanism incorporated in the proposed GNN-based models for predicting CPPIs?
(3) In terms of predictive performance, how does the GraphSAGE and GIN model attention compare to existing state-of-the-art methods?
(4) How are the attention weights visualized or interpreted biologically?

The primary objective is to contribute substantially to the domain of CPPI prediction by creating an advanced computational model that integrates attention mechanisms with the resilient functionalities of GNN to improve the precision of CPPI forecasts and acquire a clear depiction of the complex interconnections.

The subsequent sections are structured to present an overview of CPPI, emphasizing its importance within the scientific community, followed by a comprehensive analysis of the review of the research in Sec. 2. Section 3 outlines the approaches with their respective algorithms, followed by Sec. 4, which presents the empirical findings of the approaches and summarizes the main results, and discussion on the broader significance of the acquired results. Section 5 presents the comparison the novel Cancer dataset with an existing dataset, employing both the innovative

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

techniques GraphSAGE and GIN augmented with attention and an existing method. Section 6 provides a concise overview of CPPIs and suggests potential avenues for future research to enhance the understanding of PPI dynamics.

2. BACKGROUND

Genes are the fundamental part that creates proteins for different biological processes. The genes in DNA encode the proteins that lead to millions of possible protein structures. Proteins are linear polymers built of monomer units called AAs. AAs are the building blocks of proteins. There exist only 20 basic amino acids, whose sequence and association form the different unique proteins. Each amino acid comprises an amino group ($-NH_3$), a carboxyl group (–COO) attached to central α-carbon, and an R-group. The R-group for each amino acid differs in polarity, structure, and charge. It allows amino acids to bond with each other according to the chemical characteristics of the side chains.[5]

The PPIs are the vital processes as it illustrates how the alteration in the interaction between two proteins changes the characteristics of proteins, the interface, and signaling between the two interacting proteins. Thus, identifying PPI is crucial to interpret the function of proteins, detect diseases and design new drugs.[6] The *in-vivo* and *in-vitro* detection methods are employed extensively in these recent decades for PPI.[7] The compilations of deep experiments are challenging as both these methods identify huge PPI networks but are tedious and labor-intensive. Various *in-silico* methods are proposed for PPI in the last two decades.

The process of validating the PPI experimentally in wet labs demands high cost, time and labor. Thus, computational techniques are being employed to determine PPI effectively. Currently, computational approaches are widely employed to explore biological functions and promote the advancement in designing new drugs and therapeutics. Machine learning algorithms are employed for PPI for the last few decades.[8] The machine learning models like Random forest,[9] the SVM, and its derivatives are employed to reduce dimensions of proteins.[10] SVM employ protein features like 3D structures, domain information sequence, and patterns to find the optimal hyperplanes that separate the proteins with different labels with a maximum margin.[11] Decision tree recursively partitions the sample space based on sequence,[12] 3D structures[13] and domain features.[14] Deep Learning (DL) techniques are used to assess the functional impact of sequence changes. DL methods have become one of the revolutionary tools for studying and predicting PPI. These methods are primarily premised on learning the protein representations[15,16] from structural information of protein sequences and learning the PPI by link predictions.[17]

The different DL networks that are mainly employed for the structure prediction of PPIs are listed in Table 1.

Recent improvements in protein structure prediction use information of residue pair co-evolution related to protein sequences to extract information on residue pair sites and distances to determine 3D protein structure predictions accurately.[19] The network recognizes direct interactions, allowing correct predictions to be produced even for sequences with few or no associated sequences.

Autoencoders are also used to solve the various biological challenges like protein sequence engineering.[22] Mainly, CNNs and RNNs tend to capture key sequence residues with high generalization. 3D CNNs[16,18] extracts the 3D structural features of proteins. Thus, identifying residues' spatial, sequential structure is important for PPI but is prone to quantization errors[26] due to limited resolution and huge computations.

Graphs represent complex relationships, encompassing chemical structures, social networks, and knowledge graphs. The PPI network can be represented by undirected graphs. EduCross[27] an adversarial bipartite hypergraph learning framework improves cross-modal retrieval (CMR) in educational resources. The method uses hypergraph learning to represent complex relationships and integrates framelet-based deep understanding to extract nuanced characteristics from multimodal slides. It enhances retrieval accuracy relative to current methodologies and shows enhanced efficacy for

Table 1. State-of-the-art techniques.

Technology	Advantages
Convolutional Neural Networks[16,18,19]	Learns the patterns from the protein structure. Can process Variable length protein Sequences.
Recurrent Neural Networks[20,21]	Can process Variable length protein Sequences. To find the sequences.
Auto encoders[22]	Useful for generating new residues. Low dimensional representation for visualizing the data.
Graph Neural Network[23–25]	Variable graph sizes. Learns patterns by graph connectivity Employs Relevant interactions

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

multimodal data. HAQJSK[28] presents an innovative approach to improve graph classification, particularly for un-attributed graphs. The methodology employs a hierarchical alignment strategy, generating Hierarchical Transitive Aligned Adjacency and Density Matrices within the Continuous-Time Quantum Walk (CTQW) framework to convert graphs into fixed-size representations. HAQJSK uses the Quantum Jensen–Shannon Divergence (QJSD) to assess graph similarity, encompassing local and global structures that provide precise and dependable similarity assessments.

Graph PPIs are designed for predictions of PPI sites.[29] In the few years, many variants of GCN are effectively employed like geometric knowledge in PPIs.[30] It integrates the information of the amino acid sequence with the position of proteins. The GCN-based model with attention called Struct2Graph[23] is employed for PPI prediction from 3D protein structure information. PEGFAN[31] presents Haar-type graph framelets characterized by permutation equivariance, facilitating resilient multi-scale feature extraction addressing the complexities associated with heterophilous graphs in GNNs. PEGFAN demonstrates enhanced performance on heterophilous datasets by utilizing these framelets.

The existing state-of-the-art algorithms employed for Protein Engineering are expensive and tedious. Most of these usually employ local, sequence, and global features. Therefore, it is imperative to have automated computational approaches to predict better PPIs. The GNN have the inherent ability to analyze and learn from the structured data from PPI. In a PPI network, the edges represent the protein interactions, and nodes represent the protein residues, thus correlates perfectly with the spatial graphical structure of GNN. GNNs excel at learning expressive nodes and graph-level representations by incorporating sequence, PPI sites, and structural features. By aggregating this information from neighboring nodes and incorporating it into a node's representation, GNNs capture local and global context, complex relationships and patterns in the data, which is particularly advantageous in bioinformatics, where understanding interactions between biological entities is crucial. GNN has become crucial for protein engineering.

The paper proposes the novel architecture that explores the pertinence of the variants of GNN: GraphSAGE and GIN using attention for determining PPI. The graph architecture captures the spatial geometry of the proteins and can be integrated with neural networks to eliminate dimensional problems.

3. METHODOLOGY

PPIs are crucial to comprehend disease mechanisms and cellular processes. The paper proposes a novel methodology by leveraging the capabilities of GNNs by employing two distinct models, GIN and GraphSAGE, each augmented with a mutual attention mechanism. The primary objective of the novel methodology is to improve the precision of cancer interaction predictions by capturing the intricate interactions within PPI networks. The proposed methods, represents a unique strategy to combine the strengths of local and global graph-based representations, providing a better comprehensive understanding of protein interactions for prediction of CPPIs.

The proposed methodology for the prediction of PPI comprises four modules: Curation of CPPI data, Protein Graph, GraphSAGE with Mutual Attention and GIN with Mutual Attention.

3.1. Curation of novel CPPI dataset

The accurate prediction of PPIs is significant for unraveling the functionality of known cancer proteins, comprehending the underlying oncogenesis and determining the possible therapeutics. Although cancer-specific PPI datasets are crucial, they are currently unavailable. The research addresses the gap by curating a novel structural CPPI dataset encompassing significant cancer proteins to represent an extensive picture of cancer. The overall process of data curation is illustrated in Fig. 1.

The first step of our methodology is a rigorous curation process that involves collecting positive cancer protein interactions of Homo sapiens from STRING.[32] Thorough filtering eliminates interactions with low confidence and standardizes protein IDs. The structural data is collected from a protein data bank and includes only the protein pairs connected to publicly accessible PDB files. The cancer proteins are meticulously explored due to their importance in modern medical and biotechnological research. Consequently, multiple cross-references to PDB files exist, reflecting the accessibility of various structures for these proteins. All proteins in the database are matched to their corresponding UniProt accession numbers, and from there, the PDB files in UniProt are connected to the proteins in the database.[33] Not all proteins are fully crystallized in every PDB file; thus, randomly picking a PDB file could result in missing details about the protein's binding site. PDB files are curated according to chain ID length. To ensure the most

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

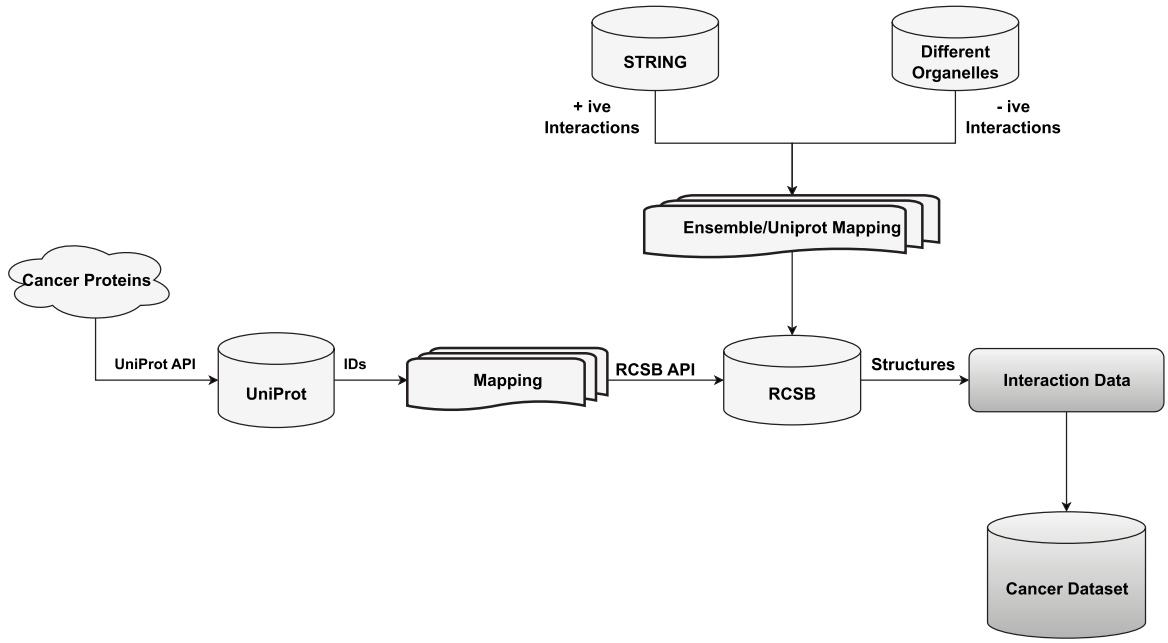


Fig. 1. Curation of CPPI dataset.

significant resolution, the target protein's most complete structure information is acquired. Each protein's crystal structure's chain length and resolution are obtained from the RCSB database.[34]

Cells are the fundamental anatomical and functional components of living organisms. Organelles, the "specialized cell structures" in the cell, perform particular functions. These organelles work together to maintain the cell's structure, function, and overall health. Diverse proteins are expressed within cellular organelles, each organelle harboring a specific set of proteins tailored to its unique function.[35] Interestingly, these organelle-specific proteins do not actively interact with each other, forming the basis for the construction of negative datasets. The curated dataset is a balanced set with an equal number of positive and negative pairs that consists of 2387 unique different proteins and 4434 interactions.

The curated CPPI dataset acquired from RCSB, STRING and UniProt databases provides significant and comprehensive coverage of CPPIs, especially for experimentally validated interactions. STRING accounts for roughly 71%[36] of the hits for experimentally validated PPIs. The quality of the dataset, including the STRING database for interaction information and 3D structural data from the RCSB PDB, is enhanced by rigorous curation processes emphasizing high-confidence experimental data. The curated dataset prioritizes experimentally validated data, enhancing confidence in the reported relationships. The extensive novel CPPI dataset identifies the novel cancer protein interactions, that are crucial to understand the complex biological systems and cancer therapeutics. The dataset's superior quality and comprehensive scope facilitate enhanced training and validation of the prediction model. The curated dataset from STRING, UniProt and RCSB excels in coverage and quality compared to other PPI databases. The experimentally validated interactions improves the reliability and provides the insights for enhancing the comprehension of molecular biology and refining treatment approaches.

The existing machine and deep learning techniques are employed to analyze these negative datasets to identify distinct features that characterize the lack of interaction among proteins. In this context, negative datasets consist of structures of proteins that do not interact within the cellular environment. The resulting novel curated PPI network ensures a high-quality dataset for subsequent analysis, paving the novel way for identification of cancer interaction patterns.

3.2. Protein graph

Graphs represent the PPI as protein entities which is defined by the interactions of the proteins with each other. *Protein graph construction*: The molecular protein graph (G_{P_A}) represents the spatial geometry of

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

proteins where $G_{P_A}=(V_A,E_A)$, where V_A denotes set of vertices and E_A set of edges. $v_a \in V_A$ is the AA or residue, and E_{ab} is the edge between them. Each node in the graph possesses different properties to be captured from the 3D structures and their respective sequence. Each residue possesses some features that feature vectors represent. The node features for PPI Graph G_{P_A} is given by $X^A_{v_a}, \{X^A_{v_a}, \forall v_a \in V_A\}$ and the neighboring function N_A as $(N_A): v_a \rightarrow 2^{V_A}$. The methodology relies on representing the PPIs as a graph, where the mutual attention mechanism of two distinct models, GraphSAGE and GIN, is crucial.

3.3. GraphSAGE with mutual attention

GraphSAGE learns on graph-structured data and makes learning about graph representation easier by attempting to capture and encode the relationships of structural data. Scalability, inductive generalization, and task-specific application make it an effective tool for managing dynamic, large-scale, graph-structured data in PPIs. A novel GraphSAGE neural network augmented with mutual attention is presented to predict CPPIs accurately. The proposed approach integrates GraphSAGE layers with a mutual attention mechanism to extract crucial features and results in valuable insights into intricate CPPIs. It generates the low dimensional representations for proteins by acquiring the similar representation of the neighboring nodes based on the "context similarity assumption". Each protein node iteratively samples and aggregates information from its immediate neighbors. The focus on a diverse array of local contexts is significant for investigating the protein interactions, as the interactions often relies on the structures of neighboring proteins. It learns a function that generates the embeddings by sampling of the local neighbors. It includes context construction and aggregation to train the model to effectively acquire the embeddings for even the unknown cancer proteins. The overall methodology of GraphSAGE with attention is shown in Fig. 2.

The model's architecture includes GraphSAGE layers in which nodes iteratively sample and aggregate data from their neighbors, and mutual attention enhances the conventional GraphSAGE method, providing a more intricate comprehension of local relationships for thorough feature extraction. Incorporating a mutual attention mechanism into GraphSAGE improves its adaptability by enabling individual protein nodes to selectively focus on neighboring nodes that contain more informative information while aggregating, thereby enhancing the representation.

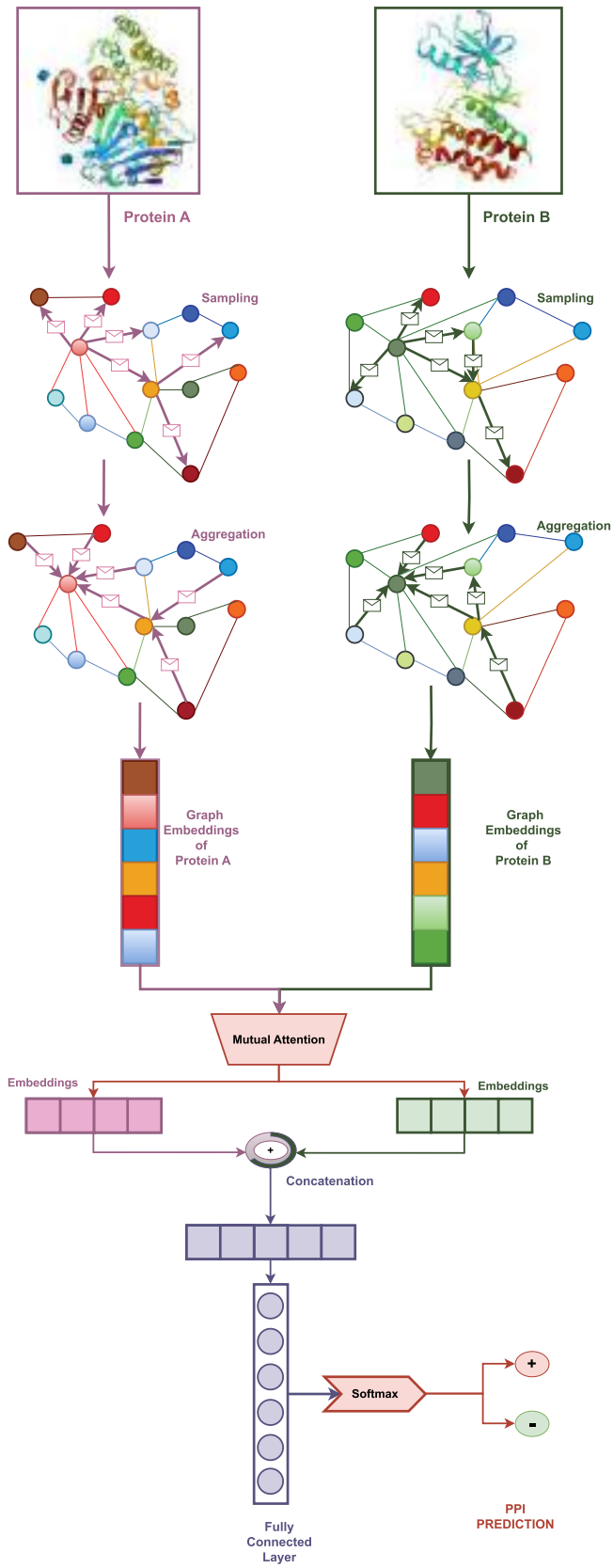


Fig. 2. (Color online) GraphSAGE with attention.

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

The detailed algorithm of GraphSAGE with mutual attention for two proteins: *Protein* $A\{G_{P_A}=(V_A,E_A)\}$ with Node features feature $\{X_{v_a}^A,\forall v_a\in V_A\}$ and *Protein* $B\{G_{P_B}=(V_B,E_B)\}$ with Node features feature $\{X_{v_b}^B,\forall v_b\in V_B\}$ is presented as Algorithm 1.

Algorithm 1. GraphSAGE with attention for CPPI prediction.

Input: Two PPI Graphs G_{P_A} and G_{P_B} where,
$G_{P_A}=(V_A,E_A)$ with Node features feature $\{X_{v_a}^A,\forall v_a\in V_A\}$
and Graph $G_{P_B}=(V_B,E_B)$ with Node features $\{X_{v_b}^B,\forall v_b\in V_B\}$,
***Training Parameters*:** Weight matrices W^k with depth K,
Nonlinearity σ, Aggregator$_K$, $\forall k\in\{1,2,...,K\}$,

Neighboring function of Graph $G_{P_A}(N_A):v_a\to 2^{V_A}$,

Neighboring function of Graph $G_{P_B}(N_B):v_b\to 2^{V_B}$,

Initialize: Randomly initialize node embeddings $H_A^{(0)}$ and $H_B^{(0)}$;

$H_A^{(0)}\to X_{v_a}^A,\forall v_a\in V_A, H_B^{(0)}\to X_{v_b}^B,\forall v_b\in V_B$

for $k=1$ **to** K **do**

  for $v_a\in V_A$ **do**

    ***Neighbor Sampling*:** *Sample neighbors for each node*;

    ***Aggregation*:** *Aggregate neighbor information for each node*;

    $H_{N_A}^{(k)}=\text{AGGREGATE}_K\left(\{H_{u_a}^{(k-1)},\forall u_a\in\text{Neighbors}(v_a)\}\right)$

    Update: *Update node embeddings using a GraphSAGE layer*:

    $H_A^{(k)}=\sigma\left(W^{(k)}\cdot\text{CONCAT}H_{v_a}^{(k-1)},H_{N_A}\right)$,

  $H_{v_a}^{(k)}\leftarrow H_{v_a}^{(k)}/||H_{v_a}^{(k)}\ \forall v_a\in V_A$

  for $v_b\in V_B$ **do**

    ***Neighbor Sampling*:** *Sample neighbors for each node*;

    ***Aggregation*:** *Aggregate neighbor information for each node*;

    $H_{N_B}^{(k)}=\text{AGGREGATE}_K\left(\{H_{u_b}^{(k-1)},\forall u_b\in\text{Neighbors}(v_b)\}\right)$

    Update: *Update node embeddings using a GraphSAGE layer*;

    $H_B^{(k)}=\sigma\left(W^{(k)}\cdot\text{CONCAT}H_{v_b}^{(k-1)},H_{N_B}\right)$

  $H_{v_b}^{(k)}\leftarrow H_{v_b}^{(k)}/||H_{v_b}^{(k)}\forall v_b\in V_B$

Attention Mechanism:

$S_{A,B}$=**Similarity_Score** $(H_{A,v_a}^{(k)},H_{B,v_b}^{(k)})$,

$S_{B,A}$=**Similarity_Score** $(H_{B,v_a}^{(l)},H_{A,v_b}^{(k)})$

Attention Computation: *Compute attention weights using softmax*;

$$At_{A,B}=\frac{\exp(S_{A,B})}{\Sigma_{m\in V_B}\exp(S_{A,B,m})},\quad At_{B,A}=\frac{\exp(S_{B,A})}{\Sigma_{m\in V_A}\exp(S_{B,A,m})}$$

Mutual Attention: *Update node embeddings using mutual attention*;

$H_{A,v_a}^{(k)}=H_{A,v_a}^{(k)}+\sum_{v_b\in V_B}At_{A,B}\cdot H_{B,v_a}^{(k)}, H_{B,v_a}^{(k)}=H_{B,v}^{(k)}+\sum_{V_a\in V_A}At_{\boldsymbol{B},A}\cdot H_{A,v_a}^{(k)}$

Concatenation: $H_{\text{Con}}=\text{CONCAT}(H_A^{(K)},H_B^{(K)})$

Final Layer: Use the concatenated node embeddings for PPI prediction;

$\hat{Y}=\text{SOFTMAX}(W^{(K+1)}\cdot H_{\text{Con}}+b^{(K+1)})$,

$\hat{Y}$ is the Prediction Probability.

The model leverages the inductive capability of GraphSAGE with attention to capture complex interactions in dynamic protein structures by iterative sampling and aggregating data from nearby nodes. The mutual attention mechanism makes more precise interaction predictions possible, which enhances the model's comprehension of pertinent features between protein pairs.

3.4. GIN with mutual attention

The GIN architecture is a permutation-equivariant design that generates the same nodes regardless of the node order in the input graph. GIN is premised on the graph "isomorphism principle" that determines whether the two graphs are isomorphic, or structurally identical. GIN is an improvement on the Weisfeiler–Lehman (WL) test, a graph isomorphism test that repeatedly improves node representations by taking into account the neighbors. The WL test for the two proteins in *A* and *B* is depicted in Fig. 3.

GIN compiles data from adjacent nodes by summing the node features, followed by the transformation. Through this aggregation process, GIN identifies the global patterns within the graph. GIN has better expressive capability than GCN. It capture intricate relationships and higher-order dependencies in graph-structured data due to the isomorphism layer and aggregation mechanism. The paper presents an innovative approach that enhances interaction prediction by integrating the expressive capability of GIN with a unique mutual attention mechanism. The expressiveness and power of GIN are among its prominent features. GIN is a resilient and efficient model for dynamic graph-structured protein data, and it is as powerful as the WL method with many iterations. The overall methodology of GIN with attention is shown in Fig. 4.

In the PPI, GIN aggregates data from nearby represented nodes by utilizing a message-passing mechanism and the attention enables to focus on neighboring nodes that are more pertinent throughout the process of aggregating information. The output of the model is invariant to node ordering since this procedure is permutation-equivariant. Multiple GIN layers are used in the model's design to acquire hierarchical information, with mutual attention to improve predictions. We use *k*-fold cross-validation on a protein pair dataset to evaluate the predictive performance of GIN-Mutual. The detailed algorithm of GIN with mutual attention for two proteins: *Protein* $A\{G_{P_A}=(V_A,E_A)\}$ and *Protein* $B\{G_{P_B}=(V_B,E_B)\}$ is presented as Algorithm 2.

Algorithm 2. GIN with attention for CPPI prediction.

Input: Two PPI Graphs G_{P_A} and G_{P_B} where, $G_{P_A}=(V_A,E_A)$ with Node features feature $\{X^A_{v_p},\forall v_a\in V_A\}$ and Graph $G_{P_B}=(V_B,E_B)$ with Node features feature $\{X^B_{v_p},\forall v_b\in V_B\}$,

Training Parameters*: Weight matrices W^k with depth K,* $\forall k\in\{1,2,...,K\}$,

Neighboring function of Graph $G_{P_A}(N_A):v_a\to 2^{V_A}$

Neighboring function of Graph $G_{P_B}(N_B):v_b\to 2^{V_B}$

Initialize: Randomly initialize node embeddings $H^{(0)}_A$ and $H^{(0)}_B$; $H^{(0)}_A\to X^A_{v_a},\forall v_a\in V_A, H^{(0)}_B\to X^A_{v_b},\forall v_b\in V_B$

for $k=1$ **to** K **do**

K is the depth

Aggregation for G_{P_A}:

for $v_a\in V_A$ **do**

Update node embedding $H^{(k)}_{A,v_a}$ **using GIN layer;**

$$H^{(k)}_{A,v_a}=\mathrm{GIN}\left(\{H^{(k-1)}_{A,v_a},\forall u_a\in\mathrm{Neighbors}(v_a)\}\right)$$

Aggregation for G_{P_B}:

for $v_b\in V_B$ **do**

Update node embedding $H^{(k)}_{B,v_b}$ **using GIN layer;**

$$H^{(k)}_{B,v_b}=\mathrm{GIN}\left(\{H^{(k-1)}_{B,v_b},\forall u_b\in\mathrm{Neighbors}(v_b)\}\right)$$

Attention Computation: Compute attention weights between nodes in G_{P_A} and G_{P_B};

Compute attention weights $At_{A,B}$ *for node* v_a **in** G_{P_A};

$$At_{A,B}=\mathrm{ATTENTION}(H^{(k)}_{A,v_a},H^{(k)}_B)$$

Compute attention weights $At_{B,A}$ *for node* v_b **in** G_{P_B};

$$At_{B,A}=\mathrm{ATTENTION}(H^{(k)}_{B,v_b},H^{(k)}_A)$$

Mutual Attention: **Update node embeddings using mutual attention;**

$$H^{(k)}_{A,v_a}=H^{(k)}_{A,v_a}+\sum\nolimits_{v_b\in V_B}At_{A,B}\cdot H^{(k)}_{B,v_a},$$

$$H^{(k)}_{B,v_a}=H^{(k)}_{B,v_a}+\sum\nolimits_{v_a\in V_A}At_{B,A}\cdot H^{(k)}_{A,v_a}$$

Concatenation:

$$H_{\mathrm{Con}}=\mathrm{CONCAT}(H^{(K)}_A,H^{(K)}_B)$$

Final Layer: Use the concatenated node embeddings for PPI prediction;

$$\hat{Y}=\mathrm{SOFTMAX}(W^{(K+1)}\cdot H_{\mathrm{Con}}+b^{(K+1)}),$$

$\hat{Y}$ is the Prediction Probability

Several essential steps are necessary to extend the approaches to datasets beyond cancer-specific PPIs including comprehending the dataset, modifying the GNN architecture, and executing efficient training and evaluation methodologies. Initially, curation of dataset that precisely represents the intended biological context, such as PPIs for a particular disease or organism, is essential. Thereafter, it is essential to analyze the features of dataset, encompassing the structure and characteristics of nodes (e.g., categorical, numerical, or textual) and edges (e.g., weights or labels), in addition to the overall graph structure and domain-specific information to guide feature selection and model development. The feature engineering is vital to modify input features, such as amino acid sequences or structural aspects, to accommodate context-specific variations. The GNN architecture must be adapted by choosing a suitable variant of GNN for broad applications to prioritize essential neighbors. Attention mechanisms, such as scalar and multi-head attention, are employed to improve node representation learning while pooling strategies facilitate the effective aggregation of node representations. Similarly, the training encompasses dataset preprocessing, selecting an appropriate loss function for the task, implementing techniques such as mini-batch training or graph sampling for extensive graphs, and performance evaluation using relevant metrics. Moreover, fine-tuning the model and encompassing hyperparameter adjustments and alterations to the attention process is essential to enhance performance for novel datasets. Ultimately, comprehensive validation by testing on separate datasets is crucial to ascertain the robustness and generalizability of the methodology.

3.5. Role of attention mechanism in enhancing GNN interpretability and performance

The attention mechanism is essential for improving the efficacy and interpretability of GNN, especially within the GIN and GraphSAGE framework. The GIN design employs an injective multiset function for neighbor aggregation, to acquire the intricate structural information from the graph. The attention mechanism is incorporated into GIN to dynamically assess the contributions of neighboring nodes according to their significance to the target node. It enables GIN to concentrate on influential neighbors, improving the quality of node embeddings obtained from local structures. In GraphSAGE, the attention mechanism allows the model to acquire attention weights for various neighbors during the aggregation phase, indicating that not all neighbors contribute uniformly to the updated representation of a node; instead, more significant neighbors may be assigned more weights, resulting in more informative node embeddings.

Enhanced Interpretability: Attention scores indicate the significant nodes that contribute to a specific prediction. Through analyzing these scores, researchers can acquire insights into the fundamental relationships and dependencies within the network. The attention mechanism enables to dynamically discern the most

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

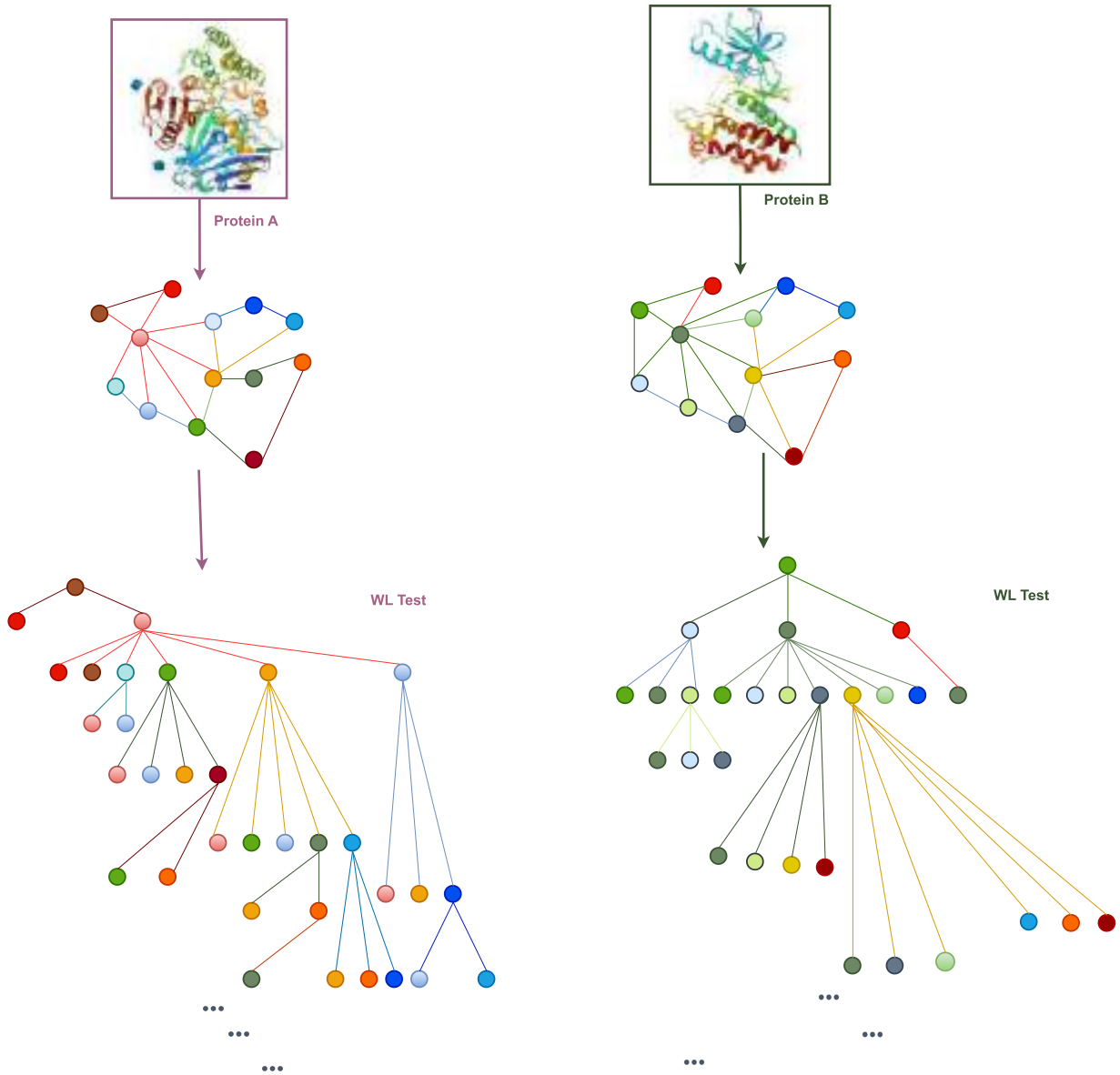


Fig. 3. (Color online) GIN WLTEST.

significant features and associations pertinent to the task. In conventional GNNs devoid of attention mechanisms, all neighbors are regarded uniformly during aggregation, perhaps failing to encapsulate the intricacies of complex interactions efficiently. Attention techniques improve computational efficiency by enabling models to concentrate on a specific selection of pertinent nodes instead of indiscriminately aggregating input from all neighbors. This selective aggregation minimizes computational overhead and memory consumption, enabling the application of these models to more enormous datasets or more intricate graphs.

Incorporating attention processes into GIN and GraphSAGE markedly improves their interpretability and predictive efficacy relative to conventional GNN methodologies. These variants enhance accuracy by enabling models to concentrate on pertinent neighboring nodes while offering significant insights into underlying predictions' decision-making processes. The attention mechanism improves interpretability by allocating weights to the edges and nodes that significantly influence the model's predictions. The attention mechanism enhances prediction performance by allowing the model to concentrate on the interaction graph's most pertinent aspects, thereby minimizing noise from less significant connections. This focused technique enhances predictions' accuracy and biological relevance compared to conventional GNN methods.

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

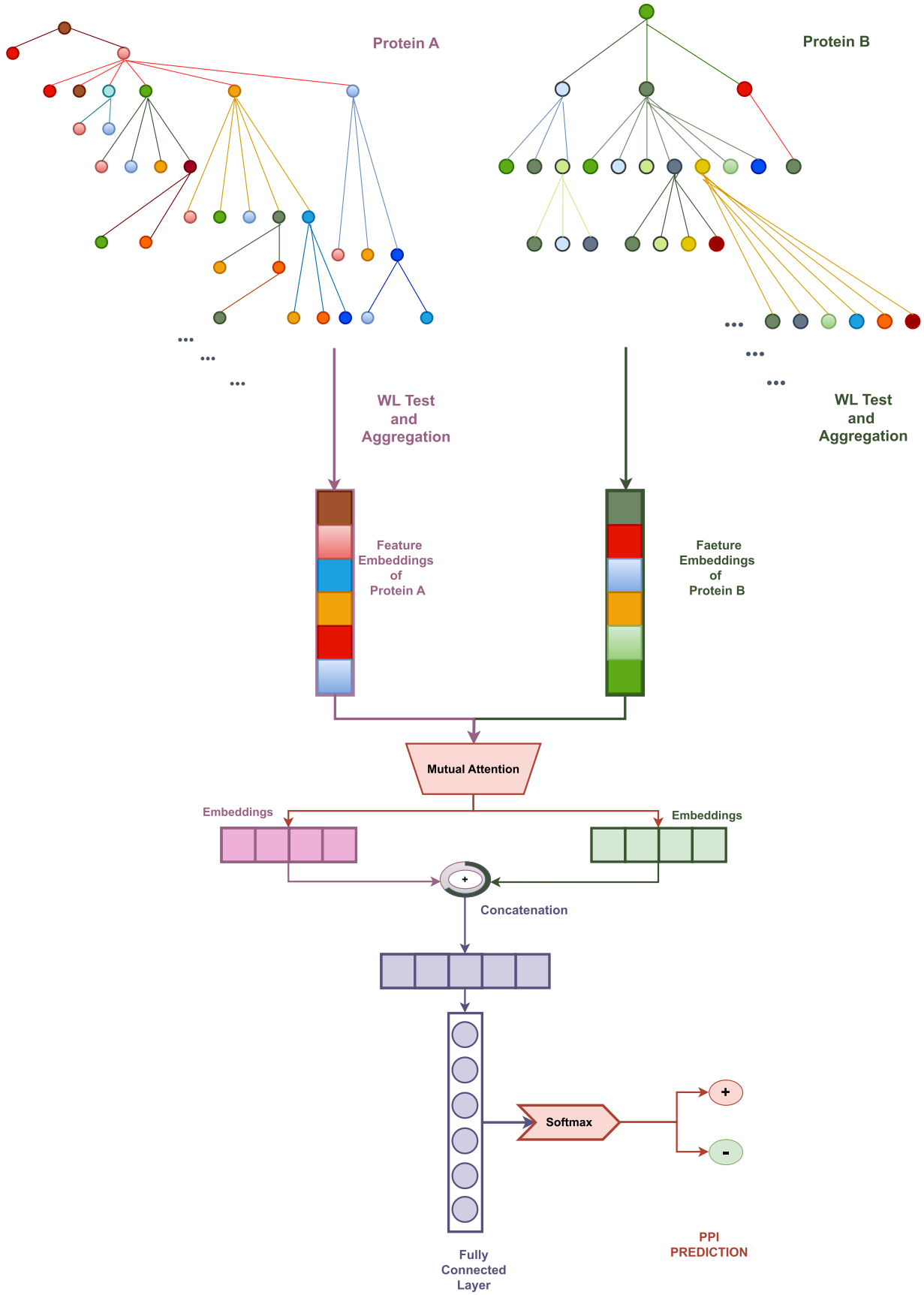


Fig. 4. (Color online) GIN with attention.

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

3.6. Hyperparameters and training procedures for reproducibility

The essential hyperparameters for training encompass model architecture details, including the number of GNN layers, hidden units per layer, and the number of attention heads. The model emphasizes embedding node features, iterative message passing, and attention-based interaction modeling. Each protein graph is depicted using node embeddings (dimension=20), with feature aggregation executed over two GNN layers (layer_gnn=2). The mutual attention mechanism entails mapping node features from two protein graphs into a shared space using two linear layers (W1_attention and W2_attention). A learnable weight vector (w) calculates attention ratings to determine the significant graph interactions. The resultant attention-based representations are concatenated and transmitted through a fully connected output layer (W_out) for binary interaction categorization. The dataset for training is constructed by encoding each protein network with integer-indexed node characteristics and adjacency matrices that depict links. The interaction labels are binary, indicating whether a specific pair of proteins interact. The model is trained via the cross-entropy loss function and optimized with the Adam optimizer. The learning rate is set to 1e−3 and decreases by a factor of 0.5 per 10 epochs. Training occurs over 30 epochs, with each epoch randomly picking 800 data points from the training dataset to ensure diversity. To assess the model, 5-fold cross-validation is utilized to guarantee reliable performance measures.

The model undergoes validation and testing in each fold, and metrics are calculated and recorded at each epoch, and the model's weights are saved to guarantee reproducibility. The attention score matrix is generated by merging projected node features from both graphs and utilizing softmax across rows and columns to calculate attention weights. The resultant weighted total generates context-sensitive representations of the proteins, subsequently employed for the interaction prediction task. The random seeds are uniformly employed throughout all processes for deterministic outcomes during dataset shuffle and model initialization. Incorporating all these hyperparameters and the training settings guarantees the subsequent researchers to effectively reproduce and expand upon this study.

3.7. Scalability and generalization to larger datasets

The scalability of GNNs augmented with attention mechanisms, whether applied to larger datasets or diverse protein interactions beyond cancer-specific situations, presents both opportunities and challenges. A primary obstacle is computational complexity. Traditional GNNs, even when augmented with attention methods, pose significant overhead due to their message-passing nature. As the graph size increases, the computing needs grow exponentially, leading to extended training times and increased memory demands. Thus, require numerous iterations to converge, making them less viable for large-scale applications. To address these challenges, different new solutions have emerged:

- Mini-batch training method[37]: It enhances the effective processing of large graphs by partitioning them into smaller, manageable batches, thus preserving crucial information while reducing memory usage and training time.
- Distributed training approaches parallelize the training process across multiple devices or nodes, significantly enhancing scalability but require careful management of communication overhead and work distribution to ensure adequate resource utilization.[38]
- Innovative, scalable designs are also employed to overcome these limitations. Models like SHINE[39] are designed to handle heterogeneous larger graphs and imbalanced datasets effectively, maintaining performance. These advancements are crucial for expanding the application of GNNs to diverse protein interaction datasets beyond cancer-specific contexts.

4. RESULTS

Accuracy, F_1-score, Precision, Recall, and area under the receiver operating characteristic curve (AUC-ROC) are used to evaluate the model's efficacy and quantify how well the model predicts the PPIs. The SAGE and GIN, with attention, excel in state-of-the-art prediction performance on a balanced set with an equal number of positive and negative pairs and accurately predict PPIs through extensive training, testing, and validation. The ROC-AUC Graphs of the two approaches are shown in Fig. 5. The efficacy of the proposed model strategy emerges from the unique capabilities of the GNN and the attention mechanism. The proposed methods demonstrate efficacy in interaction prediction by concentrating on biologically relevant patterns and merging prediction accuracy with explainability. The attention mechanism inherently emphasizes the residues with significant interaction potential, and

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

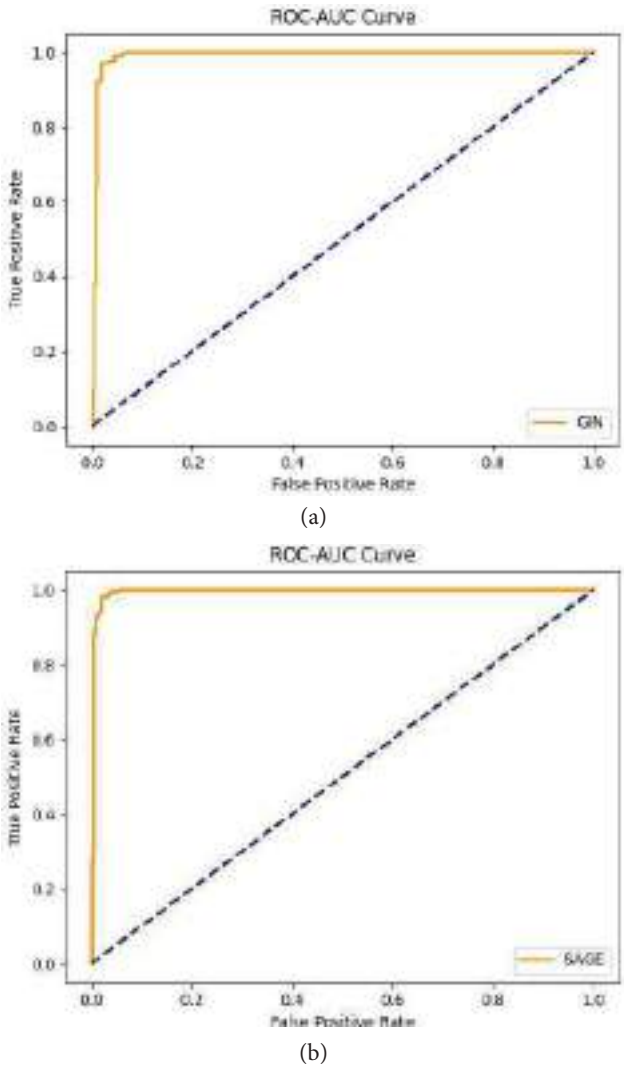


Fig. 5. (Color online) ROC-AUC graphs: (a) GIN with attention and (b) SAGE with attention.

thus, the model's congruence between model design and biological aspects improves both interpretability and predictive capability.

The Figure 6 presents the comprehensive comparative analysis of the three models by visual representation of their confusion matrices and provides insights for selecting the most optimal method for the prediction.

The comparison of the results on the novel approaches viz: GraphSAGE and GIN with attention and existing model[23] approach on the novel Cancer dataset is shown in Fig. 7. It shows that *GraphSAGE with mutual Attention* outperforms current models on several assessment parameters when predicting PPIs. The acquired results demonstrate its effectiveness and offer enhanced performance compared to the state-of-the-art method for prediction of CPPIs. *GIN with Attention* also performs better on evaluation parameters and shows encouraging results in predicting PPIs. The proposed models captures intricate interactions within protein structures with robustness and effectiveness, offering a cutting-edge approach to protein interaction prediction.

This section further presents answers to the research questions outlined in Sec.1.

RQ1: How the graph-based methods viz: GraphSAGE and GIN effectively determine the critical nodes and edges in the protein interaction graph?

The proposed methods offer a novel approach to PPI prediction by leveraging graph-based representation learning techniques, allowing them to capture complex relationships within protein interaction networks. GraphSAGE employs an innovative

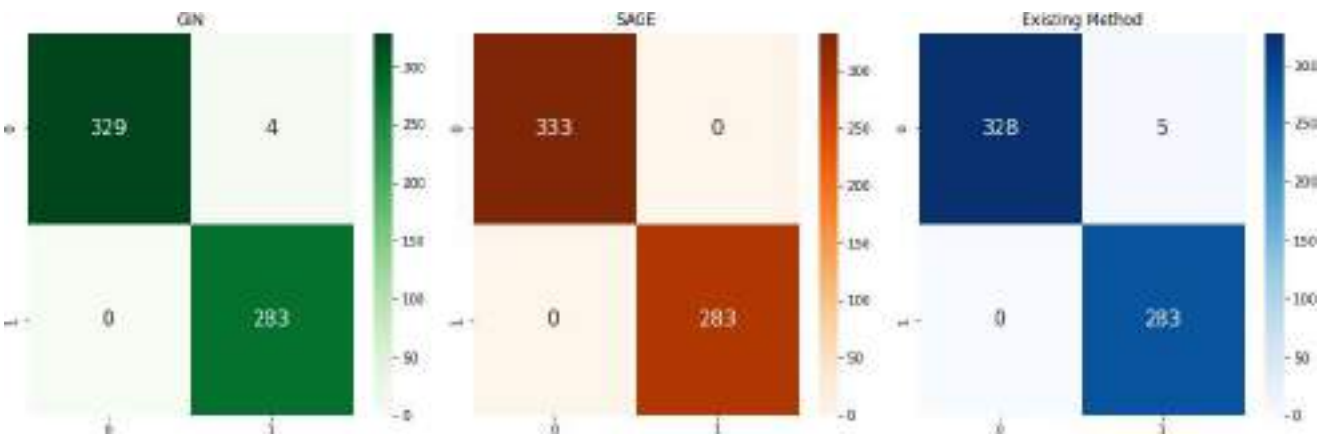


Fig. 6. (Color online) Confusion matrices comparison.

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

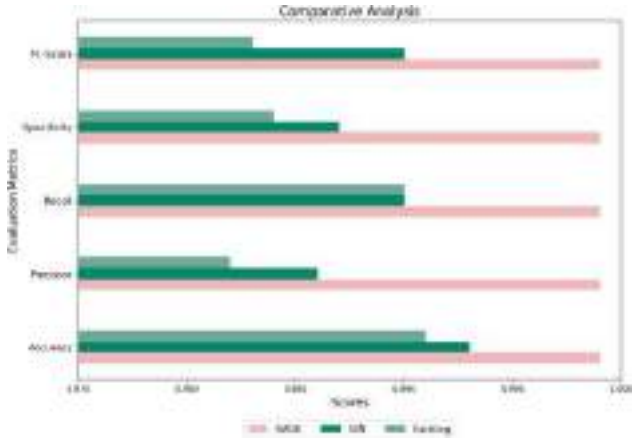


Fig. 7. (Color online) Performance comparison.

graph-based representation learning algorithm to enhance PPI prediction, enabling it to effectively capture intricate relationships within protein interaction networks. GraphSAGE enhances the ability to identify complex network patterns by effectively sampling and aggregating data from nearby nodes. It ensures both scalability and improved prediction accuracy. GIN utilizes a modified variant of the WL test, designed explicitly for graph categorization. GIN acquires node representations within the graph, enabling efficient capture of data structured in a graph format. It captures complex graph features and acquire valuable node representations effectively. The rigorous experimentation shows that proposed methods outperform existing approaches by efficiently leveraging the protein network structure and integrating node properties, resulting in more precise PPI predictions and showing better results with standard approaches on evaluation metrics.

RQ2: Why is the attention mechanism incorporated in the proposed GNN-based models for predicting CPPIs?

Incorporating an attention mechanism is intended to improve the model's efficacy by capturing contextual information and focusing on critical residues and relationships within the protein network. Attention enhances the model's ability by assigning different weights and priorities to the various interactions. It allows the model to flexibly alter the significance of protein residues, leading to more precise PPIs and concentrating on crucial nodes and edges to capture intricate relationships between proteins.

RQ3: In terms of predictive performance, how does the GraphSAGE and GIN model attention compare to existing state-of-the-art methods?

The primary objective is to evaluate the competitiveness of the proposed models through an extensive comparative analysis with established state-of-the-art approaches, showcasing its predictive capabilities and adaptability. When predicting the CPPIs, the proposed models, achieve state-of-the-art prediction performance on a balanced set with an equal number of positive and negative pairs and accurately predict PPIs through extensive testing and validation. The comparison of the results on the novel approaches and existing model are shown in Fig. 7.

RQ4: How are the attention weights visualized or interpreted biologically?

The attention mechanism incorporated with the proposed methods, especially for PPI networks, enhance interpretability by elucidating the significance of different nodes and edges throughout the learning process. Visualizing attention weights on network topologies enables researchers to examine the influence of various nodes on one another by superimposing attention scores on the graph, emphasizing significant nodes and edges crucial in the interaction network. Attention mechanisms clarify pathways by emphasizing crucial proteins involved in specific interactions and discern alterations in interaction patterns that may correlate with biological events, such as disease progression or treatment response.

Figure 8 shows two plots illustrating attention weights for Protein 1 and Protein 2, emphasizing the residues of the protein sequences considered most significant by the model during interaction prediction. Attention weights determine the degree of focus the model gives to each protein segment during prediction. These plots provide insights to researchers about the critical protein segments for interactions. For Protein 1, the attention weights demonstrate considerable variability with pronounced peaks and troughs, signifying that the model allocates differing degrees of significance throughout the sequence. The spikes indicate segments of the protein sequence that the model considers more significant for PPIs, whereas the troughs denote less critical portions. In contrast, Protein 2 has a more uniform pattern, with a substantial rise in interest towards the later portion of the sequence, suggesting a particular region that may be crucial for interactions. In both plots, the x-axis represents the residues within the protein sequence, while the y-axis illustrates the attention weight, indicating the significance attributed to the model. The statistical analysis identifies critical interaction areas or functional domains necessary for

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

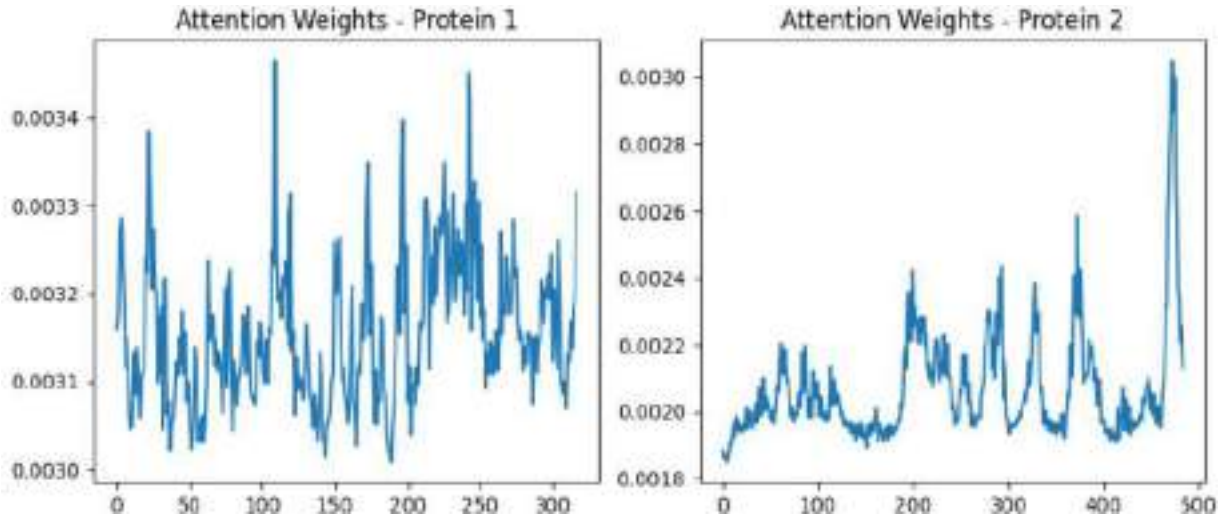


Fig. 8. (Color online) Visualization of attention weights of two interaction proteins.

determining protein behavior, developing pharmaceuticals, or recognizing therapeutic targets.

5. DISCUSSION

The paper assesses two datasets: the novel Cancer dataset with an existing dataset, employing both the innovative graph-based techniques GraphSAGE and GIN augmented with attention and the conventional method.[23] The assessment performance metrics: Accuracy, F_1-score, Precision, Recall, and area AUC-ROC are used for evaluation.

5.1. Comparison of techniques on existing dataset

The results derived from the current dataset demonstrate that our strategy resulted in enhanced and comparable outcomes compared to the existing technique. The comparable enhancement confirms the technique's efficacy and establishes a robust standard for comparison with conventional methods. The comparative study, specifically in F_1-score and MCC, as depicted in Figs. 9(a) and 9(b), respectively, emphasizes how the proposed method outperforms others.

5.2. Comparison of techniques on novel curated cancer dataset

The application of novel techniques to the Cancer dataset resulted in significant enhancements. The results highlight our technology's flexibility and scalability and confirm its ability to extract valuable insights from newly generated data domains. The comparative study, specifically in F_1-score and MCC, as depicted in Figs. 10(a) and 10(b), respectively, on our novel curated CPPI dataset emphasizes how the method outperforms others. The novel cancer dataset demonstrates its value as a valuable addition to the field by attaining comparable enhanced scores in all analyzed measures.

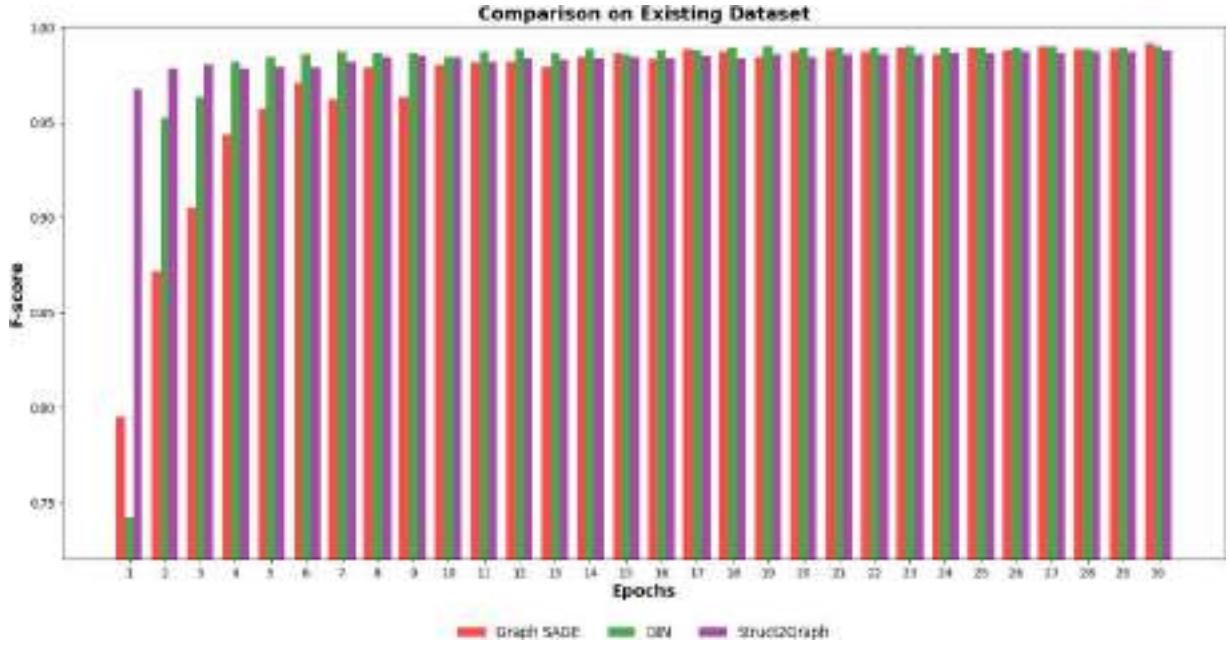
5.3. Cross-validation of techniques using both datasets

The cross-validation strategy is employed to ensure the reliability and consistency of the approaches, utilizing both existing and innovative methodologies on the two datasets. The cross-validation findings demonstrate that our innovative approach achieves high performance with the established dataset parameters and effectively adapts to new, unexplored contexts. Employing a dual-dataset validation strategy enhances the applicability, reliability and emphasizes on the adaptability of the proposed approaches.

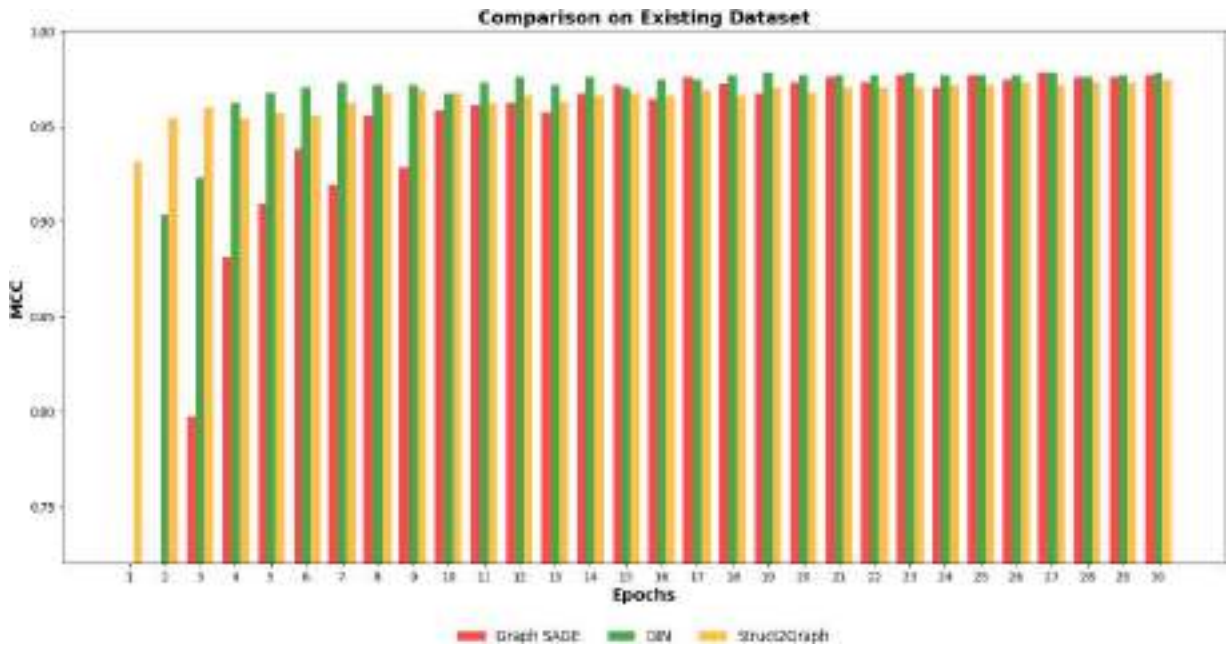
5.4. Computational requirements for Large-Scale PPI networks

The computational requirements of the methods are premised on the dimensions of the PPI graph and the intricacy of the GNN model. The computing demands of the proposed models are substantial, particularly as the network scale expands. The attention coefficients are derived from the embeddings of a node and its neighbors, utilizing linear transformations succeeded by nonlinear activation functions. The computational

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332



(a) Comparison of F_1-scores on existing dataset



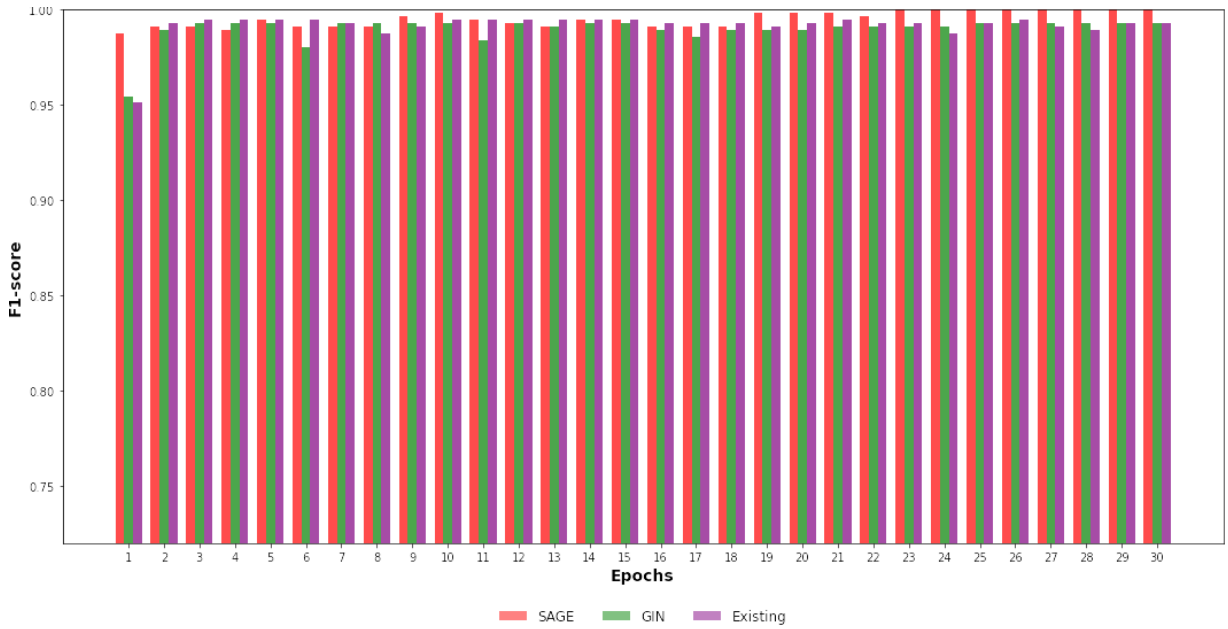
(b) Comparison of MCC-scores on existing dataset

Fig. 9. (Color online) Comparison of techniques on existing dataset.

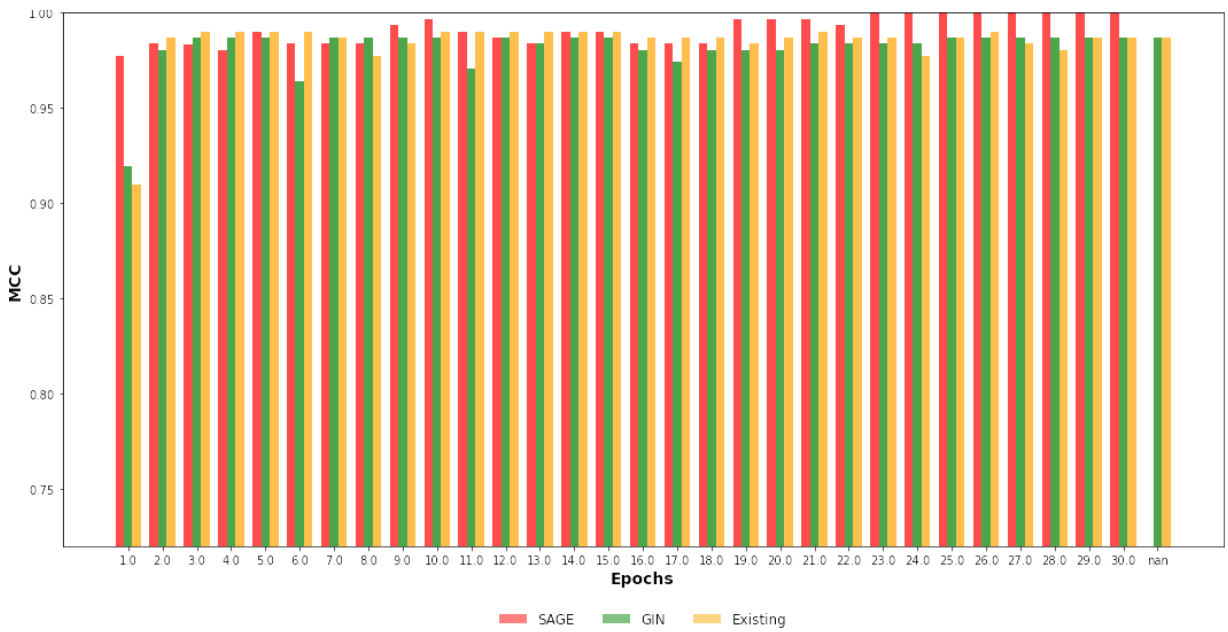
complexity for an individual attention head is $O(N^2D)$, where N represents the number of nodes and D is the dimension of the node embeddings. The methods necessitate considerable memory resources to accommodate the storage of several embeddings and attention weights. Although the method enhanced with attention mechanisms offers robust capabilities for examining extensive PPI networks, it entails considerable computational and memory requirements.

The High-Performance Computing lab facility at our university is employed to facilitate intensive computational research. We utilized a compute node (GPU) from it for the proposed methods. The node is empowered by dual Intel Xeon Gold 6242R CPUs (3.10 GHz, 20 cores/40 threads each), 512 GB of RAM, and 2*960 GB SSD SATA, providing high computational resources. Its integration with a powerful NVIDIA Tesla V100 32GB graphics card enabled the efficient

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332



(a) Comparison of F_1-scores on novel cancer dataset



(b) Comparison of MCC-scores on novel cancer dataset

Fig. 10. (Color online) Comparison of techniques on novel curated cancer dataset.

execution of proposed methods, leveraging high parallel processing capability for the complex computations involved in PPI prediction. The setup ensured robust performance and precise model predictions.

6. CONCLUSION

The research presents a novel CPPI dataset that represents an extensive picture encompassing the different types of cancer proteins and introduces an innovative approach combining graph-based methods viz: GraphSAGE and GIN with attention to identify and interpret intricate connections for predicting CPPIs and enhancing the precision of predictions. The incorporation of attention mechanisms allows the model to dynamically prioritize pertinent information while transmitting messages, resulting in improved interpretability and predictive capabilities.

The enhanced comparative outcomes from the current and novel datasets highlight our research's

DOI: 10.1142/S2737416525400022
J. Comput. Biophys. Chem. **2026**, *25* (2), 315–332

credibility. The successful implementation and verification of the innovative dataset and the graph-based architectural approach create opportunities for further investigations and improvements in this field. The research shows that by combining sophisticated methods like graph-based models and data augmentation, we can significantly enhance structural data's analytical capabilities and results. It helps to understand the complex network of interactions between proteins and thus paves the way for further research in medicine and biology. The novel approach advances bioinformatics and computational biology, providing vital insights into the complex world of CPPIs. Cancer Proteome research has great potential for revealing novel treatment targets and personalized medicine methods across various tissues, developmental stages, and disease states.

STATEMENT OF USAGE OF ARTIFICIAL INTELLIGENCE (AI)

No AI is used to prepare the research, and it is conducted independently by the authors.

DATA AVAILABILITY

The authors curate the novel CPPI dataset, which will be available upon the request to the corresponding author.

AUTHOR CONTRIBUTIONS

Rafiya Jan: Data Analysis and Curation, Design and Conceptualization, Implementation, Manuscript Writing. Ahsan Hussain: Conceptualization, Supervision, Paper writing. Assif Assad: Supervision. Basharat Bhat: Data Analysis and Curation, Supervision.

CONFLICT OF INTEREST

We, the undersigned, declare that all named authors approve this manuscript paper, that it is original and there are no competing interests regarding this work. Signed by all authors as follows:

FUNDING INFORMATION

No Funding Acquired.

ORCID

Rafiya Jan
https://orcid.org/0000-0002-2901-0005
Ahsan Hussain
https://orcid.org/0000-0003-0721-176X
Assif Assad
https://orcid.org/0000-0003-0323-6863
Basharat Bhat
https://orcid.org/0000-0002-4148-1694

References

1. Skrabanek, L.; Saini, H. K.; Bader, G. D.; Enright, A. J. Computational Prediction of Protein–Protein Interactions. *Mol. Biotechnol.* **2008**, *38*, 1–17.
2. Laddach, A.; Chung, S. S.; Fraternali, F. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. Elsevier, 2018, pp. 834–848.
3. Jumper, J. *et al.* Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.
4. Evans, R. *et al.* Protein Complex Prediction with AlphaFold-Multimer. *bioRXiv* **2021**, doi:10.1101/2021.10.04.463034.
5. Buxbaum, E. *et al. Fundamentals of Protein Structure and Function*, Vol. 31. Springer, 2007.
6. Bonetta, L. Interactome under Construction. *Nature* **2010**, *468*, 851–852.
7. Rao, V. S.; Srinivas, K.; Sujini, G.; Kumar, G. Protein-Protein Interaction Detection: Methods and Analysis. *Int. J. Proteomics* **2014**, *2014*, 147648.
8. Sarkar, D.; Saha, S. Machine-Learning Techniques for the Prediction of Protein– Protein Interactions. *J. Biosci.* **2019**, *44*, 104.
9. You, Z.-H.; Chan, K. C.; Hu, P. Predicting Protein-Protein Interactions from Primary Protein Sequences Using a Novel Multi-Scale Local Feature Representation Scheme and the Random Forest. *PLoS One* **2015**, *10*, e0125811.
10. Huang, Y.-A. *et al.* Using Weighted Sparse Representation Model Combined with Discrete Cosine Transformation to Predict Protein-Protein Interactions from Protein Sequence. *BioMed Res. Int.* **2015**, *2015*, 902198.
11. Guo, Y.; Yu, L.; Wen, Z.; Li, M. Using Support Vector Machine Combined with Auto Covariance to Predict Protein–Protein Interactions from Protein Sequences. *Nucleic Acids Res.* **2008**, *36*, 3025–3030.
12. Zahiri, J.; Yaghoubi, O.; Mohammad-Noori, M.; Ebrahimpour, R.; Masoudi-Nejad, A. PPIevo: Protein–Protein Interaction Prediction from PSSM Based Evolutionary Information. *Genomics* **2013**, *102*, 237–242.
13. Li, B.-Q.; Feng, K.-Y.; Chen, L.; Huang, T.; Cai, Y.-D. Prediction of Protein-Protein Interaction Sites by Random Forest Algorithm with mRMR and IFS. *PLoS One* **2012**, *7* (8), e43927.

14. Rodgers-Melnick, E.; Culp, M.; DiFazio, S. P. Predicting Whole Genome Protein Interaction Networks from Primary Sequence Data in Model and Non-Model Organisms using ENTS. *BMC Genom.* **2013**, *14*, 1–17.
15. Zhao, L.; Wang, J.; Hu, Y.; Cheng, L. Conjoint Feature Representation of GO and Protein Sequence for PPI Prediction Based on an Inception RNN Attention Network. *Mol. Ther. Nucleic Acids* **2020**, *22*, 198–208.
16. Renaud, N.; Geng, C.; Georgievska, S.; Ambrosetti, F.; Ridder, L.; Marzella, D. F.; Réau, M. F.; Bonvin, A. M.; Xue, L. C. DeepRank: A Deep Learning Framework for Data Mining 3D Protein-Protein Interfaces. *Nat. Commun.* **2021**, *12*, 7068.
17. 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* **2019**, *35*, i305–i314.
18. Amidi, A.; Amidi, S.; Vlachakis, D.; Megalooikonomou, V.; Paragios, N.; Zacharaki, E. I. EnzyNet: Enzyme Classification using 3D Convolutional Neural Networks on Spatial Representation. *PeerJ* **2018**, *6*, e4750.
19. Senior, A. W. *et al.* Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577*, 706–710.
20. Kovaćs, I. A. *et al.* Network-based Prediction of Protein Interactions. *Nat. Commun.* **2019**, *10*, 1240.
21. Nasiri, E.; Berahmand, K.; Rostami, M.; Dabiri, M. A Novel Link Prediction Algorithm for Protein-Protein Interaction Networks by Attributed Graph Embedding. *Comput. Biol. Med.* **2021**, *137*, 104772.
22. Linder, J.; Bogard, N.; Rosenberg, A. B.; Seelig, G. A Generative Neural Network for Maximizing Fitness and Diversity of Synthetic DNA and Protein Sequences. *Cell Syst.* **2020**, *11*, 49–62.
23. 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.* **2022**, *23*, 370.
24. Jha, K.; Karmakar, S.; Saha, S. Graph-BERT and Language Model-Based Framework for Protein–Protein Interaction Identification. *Sci. Rep.* **2023**, *13*, 5663.
25. Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; Welling, M. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web: 15th Int. Conf., ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proc. 15*, 2018, pp. 593–607.
26. Gligorijević, V. *et al.* Structure-Based Protein Function Prediction Using Graph Convolutional Networks. *Nat. Commun.* **2021**, *12*, 3168.
27. Li, M.; Zhou, S.; Chen, Y.; Huang, C.; Jiang, Y. EduCross: Dual Adversarial Bipartite Hypergraph Learning for Cross-Modal Retrieval in Multimodal Educational Slides. *Inf. Fusion* **2024**, *109*, 102428.
28. Bai, L.; Cui, L.; Wang, Y.; Li, M.; Li, J.; Philip, S. Y.; Hancock, E. R. HAQJSK: Hierarchical-Aligned Quantum Jensen-Shannon Kernels for Graph Classification. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 6370–6384.
29. Yuan, Q.; Chen, J.; Zhao, H.; Zhou, Y.; Yang, Y. Structure-Aware Protein–Protein Interaction Site Prediction Using Deep Graph Convolutional Network. *Bioinformatics* **2022**, *38*, 125–132.
30. 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.
31. Li, J.; Zheng, R.; Feng, H.; Li, M.; Zhuang, X. Permutation Equivariant Graph Framelets for Heterophilous Graph Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 11634–11648.
32. Szklarczyk, D. *et al.* STRING v11: Protein–Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613.
33. Bateman, A. UniProt: A Universal Hub of Protein Knowledge. *Protein Sci.* **2019** *28*, 32.
34. Burley, S. K. *et al.* RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy. *Nucleic Acids Res.* **2019**, *47*, D464–D474.
35. Alberts, B.; Bray, D.; Hopkin, K.; Johnson, A. D.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *Essential Cell Biology*. Garland Science, 2015.
36. Bajpai, A. K.; Davuluri, S.; Tiwary, K.; Narayanan, S.; Oguru, S.; Basavaraju, K.; Dayalan, D.; Thirumurugan, K.; Acharya, K. K. Systematic Comparison of the Protein-Protein Interaction Databases from a User's Perspective. *J. Biomed. Inform.* **2020**, *103*, 103380.
37. Liu, J.; Hooi, B.; Kawaguchi, K.; Wang, Y.; Dong, C.; Xiao, X. Scalable and Effective Implicit Graph Neural Networks on Large Graphs. In *The Twelfth Int. Conf. Learning Representations*, 2024, pp. 11.
38. Deng, C.; Yue, Z.; Yu, C.; Sarar, G.; Carey, R.; Jain, R.; Zhang, Z. Less is More: Hop-Wise Graph Attention for Scalable and Generalizable Learning on Circuits. In *Proc. 61st ACM/IEEE Design Automation Conf.*, 2024, pp. 1–6.
39. Van Belle, R.; De Weerdt, J. SHINE: A Scalable Heterogeneous Inductive Graph Neural Network for Large Imbalanced Datasets. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 4904–4915.