



OPEN ACCESS

EDITED BY

Tiziana Sanavia,
University of Torino, Italy

REVIEWED BY

Marcelo Adrian Marti,
University of Buenos Aires, Argentina
Swaminathan Venkatraman,
SASTRA University, India

*CORRESPONDENCE

Zixuan Li,
✉ 2293159762@qq.com
Zhiguo Yu,
✉ 2016869119@qq.com
Peng Li,
✉ 004570@hnuucm.edu.cn

RECEIVED 23 October 2025
REVISED 18 February 2026
ACCEPTED 24 February 2026
PUBLISHED 16 March 2026

CITATION

Li Z, Yu Z and Li P (2026) A CSGNN model-based method for essential protein identification. *Front. Bioinform.* 6:1731178. doi: 10.3389/fbinf.2026.1731178

COPYRIGHT

© 2026 Li, Yu and Li. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A CSGNN model-based method for essential protein identification

Zixuan Li*, Zhiguo Yu* and Peng Li*

School of Informatics, Hunan University of Chinese Medicine, Changsha, Hunan, China

Identification of essential proteins is fundamental for understanding cellular processes and disease mechanisms. However, many existing computational methods do not adequately model dynamic expression activity and often underutilize global network context, which limits prediction accuracy. To address these issues, we propose a Correlation-guided Subgraph Graph Neural Network (CSGNN) for essential protein identification by integrating correlation-guided graph construction with attention-based representation learning. First, we derive an activity-aware expression matrix from periodic gene expression patterns, and we construct a weighted protein network by computing Pearson correlation coefficients between gene pairs. This correlation-guided network further defines first-order and second-order neighborhoods, which provide multi-scale subgraph contexts for each protein. Next, we employ a two-layer attention-based graph convolution to learn node embeddings by aggregating information within these correlation-defined neighborhoods. Finally, we form an interaction-aware node representation by integrating each protein embedding with its neighborhood context, and we use a lightweight multilayer perceptron to output an essentiality probability for each protein. Proteins are then ranked by the predicted scores to identify essential candidates. Experiments on yeast and *E. coli* datasets demonstrate that CSGNN consistently outperforms traditional baselines, indicating improved accuracy and robustness for essential protein identification.

KEYWORDS

dynamic network, dynamic thresholding, essential protein prediction, graph attention network, similarity score

1 Introduction

Essential proteins represent a class of key proteins that are indispensable for maintaining cellular viability and normal biological functions. In biological systems, disruption of such proteins, through deletion, inhibition, or functional impairment, often leads to severe growth defects or loss of viability, particularly in microorganisms, indicating their fundamental role in sustaining life processes (Pan and Finkel, 2017). As a result, essential proteins are closely associated with the core proteome that supports basic cellular survival across conditions.

The identification of essential proteins is therefore a core task in life science research and practical applications, playing an important role in understanding biological processes, disease mechanisms, drug development, and targeted therapy. Proteins in living organisms participate in a wide range of functions, including structural support, signal transduction, metabolic regulation, and cellular maintenance (Jones and Thornton, 1996). Identifying those proteins that are essential within complex biological networks enables researchers to gain deeper insights into signaling pathways, metabolic systems, and gene regulatory mechanisms, thereby revealing fundamental principles underlying cellular organization and function.

In the medical domain, essential protein identification provides a foundation for exploring disease mechanisms, discovering pathogenic factors, and developing therapeutic targets (Pan and Finkel, 2017). For example, in cancer and neurodegenerative diseases, identifying essential proteins offers valuable guidance for the design of targeted drugs, improving the precision and effectiveness of diagnosis and treatment (Walkinshaw, 1992). In biotechnology, screening essential proteins is also beneficial for optimizing industrial fermentation processes and developing novel enzyme preparations (Savini et al., 2008). Therefore, essential protein identification constitutes a crucial step in advancing life sciences, both in basic research and in translational and applied studies.

2 Related work

Essential proteins are the direct products of gene expression. They play central roles in key cellular processes such as growth, metabolism, and signal transduction (Liu et al., 2024). Accurate identification of essential proteins helps us understand the relationship between genes and phenotypes. It also provides important clues for disease mechanism analysis and drug target discovery.

Traditional experimental methods can validate protein functions with relatively high accuracy. Gene editing (Pace et al., 2024) and protein assembly techniques (Goh et al., 2024) support functional analysis. Gene knockout (Shor and Schneidman-Duhovny, 2024) and RNA interference (Ballantyne et al., 2024) allow direct observation of functional changes *in vivo*. However, these methods are expensive, time-consuming, and limited in throughput. They are not suitable for large-scale systematic screening (Wang et al., 2024). With the development of high-throughput technologies and the accumulation of interaction and expression data, computational identification of essential proteins has become an important research direction. A common paradigm is to use protein-protein interaction (PPI) networks as the backbone and integrate multiple types of biological information (Saha et al., 2024).

In the early stage of computational research, most studies focused on the topological structure of PPI networks. Essential protein identification was treated as a key node detection problem. Proteins were ranked using centrality-based scoring strategies. Degree centrality (DC) was widely adopted because it is intuitive and interpretable. It became a classic baseline in this field (Hahn and Kern, 2005). Later, some methods introduced edge weights or neighborhood structure information. These strategies aimed to model interaction strength and local cooperative patterns. They also improved robustness to noisy edges and network sparsity. WDC (Tang et al., 2013) and W5N (Li et al., 2017) are representative examples. They extend the centrality framework by incorporating weights or expanded neighborhood information. As a result, key node detection no longer relies only on static topological features.

Beyond centrality-based and weighted scoring methods, machine learning further promoted the shift from rule-based scoring to data-driven discriminative learning (Inzamam-Ul-Hossain and Islam, 2023). These approaches usually combine

PPI topological features with biological features such as Gene Ontology (GO) annotations and expression profiles. A classifier is trained to learn the decision boundary between essential and non-essential proteins. This strategy better captures nonlinear relationships among multiple features. For example, Acencio et al. used several network topological features and two types of GO annotations as inputs to a decision tree model (Inzamam-Ul-Hossain and Islam, 2023). This work provided an early supervised learning framework for multi-source feature integration. Later, Zeng et al. built an ensemble classifier based on gradient boosting decision trees to improve feature representation and prediction accuracy (Zeng et al., 2021). Although these methods improved feature utilization and discriminative performance, they still rely heavily on explicit feature construction and selection. When dealing with high-dimensional, heterogeneous, and noisy biological data, their ability for automatic feature extraction and cross-dataset generalization remains limited (Portugal et al., 2018).

With the rapid development of deep learning in bioinformatics, researchers began to explore end-to-end representation learning frameworks. These frameworks reduce dependence on manual feature engineering. They also integrate network structure, dynamic expression, and sequence semantics in a more systematic way (Zhig et al., 2025). Zeng et al. proposed the DeepEP model. It uses node2vec to learn topological representations from PPI networks. It also applies a convolutional neural network to extract features from gene expression profiles. The two representations are fused to predict essential proteins. This design reflects cooperative modeling of topology and expression features (Ali et al., 2024). To address biological dynamics, Lu et al. proposed the AG-GATCN model. It enhances temporal convolutional networks with attention and gating mechanisms to model dynamic expression patterns. It also uses a graph attention network to extract structural features from PPI networks. This design better captures state-dependent interaction patterns and improves prediction robustness (Yang et al., 2023). Recent studies further incorporate higher-order structural information. For example, Tian et al. proposed the HCNS model. It constructs a hypergraph based on weighted PPI networks and protein complex information. It also integrates sequence features into representation learning. This model captures cooperative structures at the complex level and sequence semantic signals at the same time. It shows strong performance in accuracy and robustness (Tian et al., 2025). Current deep learning models have moved beyond shallow topology-based paradigms. They now adopt multi-scale and multi-modal biological modeling strategies. These models emphasize functional module cooperation, dynamic regulatory context, and sequence-level semantics. As a result, they provide a representation that is closer to the biological mechanisms of essential protein formation and cellular survival dependency.

Despite significant progress, essential protein identification still faces several common challenges. First, PPI data are heterogeneous and contain noise and missing interactions. False connections and incomplete interactions can affect network modeling and prediction stability. Models need to suppress noisy edges and avoid blindly propagating incorrect information across the whole network. Second, protein interactions and gene expression are

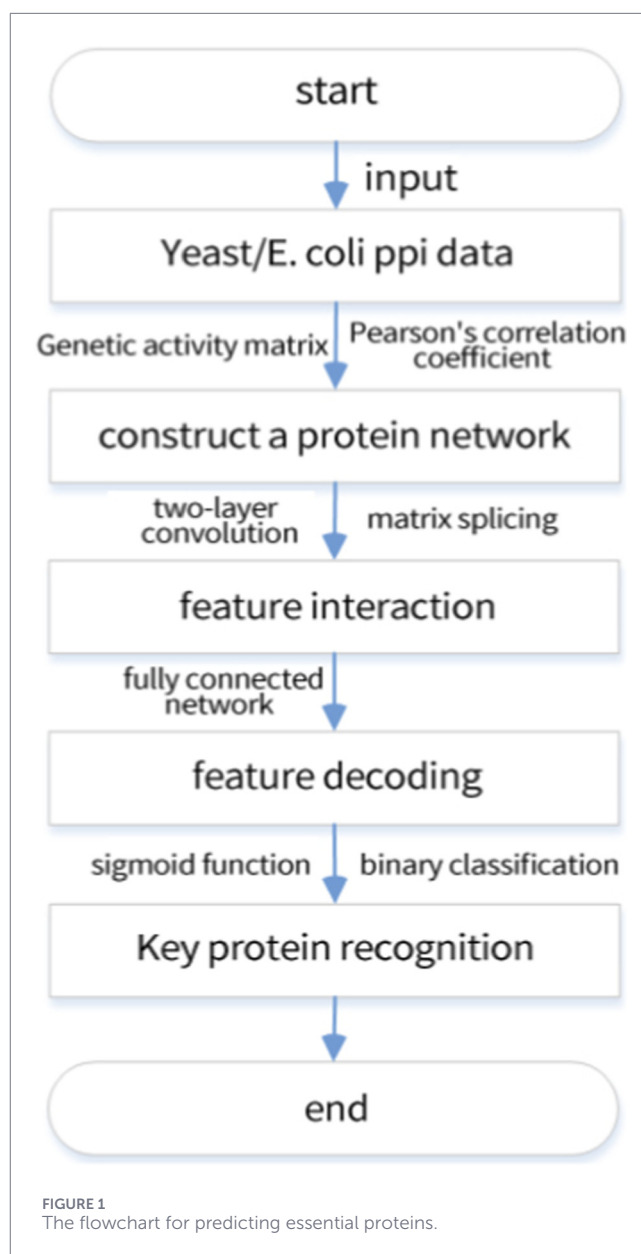
dynamic and context-dependent. Static networks cannot fully reflect essentiality under different cellular states. Models should use expression activity to capture state differences and learn representations within context-consistent neighborhoods. Third, multi-source biological data are highly heterogeneous. Effective integration while maintaining biological interpretability remains difficult. Models should perform selective aggregation under structural constraints to improve controllability and interpretability. Finally, many methods cannot balance global cooperative structures and local interactions. This limitation weakens the modeling of module-level essentiality mechanisms. An effective model should capture both local interaction details and module-level cooperation patterns.

To address these challenges, we propose the Correlation-guided Subgraph Graph Neural Network (CSGNN) for node-level essential protein prediction. The model is built on a weighted PPI network. It introduces correlation information derived from gene expression time series into the structural modeling process. We construct correlation-guided neighborhood subgraphs to capture interaction patterns that are consistent with specific biological states. The model then performs multi-layer graph representation learning on the constructed subgraphs. It gradually integrates first-order and higher-order neighborhood information to generate node embeddings. A feature fusion and classification module outputs the final essentiality prediction. In this unified framework, CSGNN jointly models PPI topology and state information reflected by expression correlations. Essential protein evaluation is performed within a state-consistent network context rather than relying only on static interactions. This design provides a more direct representation basis for modeling cooperative effects of essential proteins within functional modules.

3 Methods

The overall pipeline of the proposed Correlation-guided Subgraph Graph Neural Network (CSGNN) for node-level essential protein prediction is illustrated in Figure 1. Given a protein–protein interaction (PPI) network and time-series gene expression data, CSGNN first derives activity-aware expression signals to emphasize biologically meaningful temporal variations. Based on these signals, a correlation-guided association structure is constructed to define context-specific neighborhoods (subgraphs) for each target protein, providing a state-consistent local context for representation learning.

CSGNN then performs multi-order neighborhood propagation on the resulting graph structure to capture both direct interactions and broader functional dependencies within local neighborhoods. The learned multi-scale node embeddings are further combined with informative neighborhood evidence in a node-centric manner to form interaction-aware representations. Finally, a lightweight decoder maps these representations to essentiality probabilities, which are converted to binary predictions only during discrete evaluation. Overall, CSGNN integrates temporal activity cues with correlation-guided network topology to learn biologically coherent and discriminative representations for essential protein identification.



3.1 Constructing the protein activity feature matrix

Temporal gene expression profiles are shaped by coordinated regulatory programs, including cell-cycle progression and stimulus-responsive transcriptional control (Cubuk et al., 2021). Consequently, observed trajectories typically reflect a mixture of condition-relevant expression changes and background variability. Let M denote the number of measured time points, and let N denote the number of protein nodes in the interaction network.

Using raw expression values $\{x_{i,t}\}_{t=1}^M$ directly as node features may introduce low-amplitude variability into neighborhood aggregation, thereby weakening the contrast of condition-relevant temporal patterns and compromising the stability of correlation-based topology construction. We therefore construct an activity-filtered representation by applying a protein-specific dynamic

threshold to each temporal profile. This transformation retains expression values that exceed a protein-adaptive baseline and suppresses values that remain within the range of background fluctuation, resulting in node features that better reflect temporally salient regulatory changes.

For protein i , its expression profile is denoted by $d_i = [x_{i,1}, x_{i,2}, \dots, x_{i,M}]$, where $x_{i,t}$ is the original expression value at time point t . We define the protein-specific dynamic threshold as shown in Equation 1:

$$\text{Threshold}_i = \mu_i + \frac{2\sigma(d_i)}{1 + \text{Var}(d_i)} \quad (1)$$

where $\mu_i = \text{mean}(d_i)$, $\sigma_i = \text{std}(d_i)$, and $\text{Var}(d_i)$ denote the mean, standard deviation, and variance of d_i , respectively. The threshold comprises a baseline term μ_i and an adaptive margin controlled by temporal fluctuation. The margin increases with σ_i , requiring stronger deviations from baseline for profiles with greater dispersion. The factor $1 + \text{Var}(d_i)$ moderates the margin under large overall variability, preventing the filtering operation from becoming overly stringent for strongly varying trajectories. Accordingly, the activity-filtered expression is defined as shown in Equation 2:

$$x'_{i,t} = \begin{cases} x_{i,t}, & x_{i,t} \geq \text{Threshold}_i \\ 0, & x_{i,t} < \text{Threshold}_i \end{cases} \quad (2)$$

Finally, we construct the protein activity feature matrix $X' \in \mathbb{R}^{N \times M}$, as shown in Equation 3:

$$X' = \begin{bmatrix} x'_{1,1} & \cdots & x'_{1,M} \\ \vdots & \ddots & \vdots \\ x'_{N,1} & \cdots & x'_{N,M} \end{bmatrix} \quad (3)$$

Each row corresponds to a protein node and each column corresponds to a time point. The matrix X' is used as the node feature input in the subsequent graph neural network.

3.2 Constructing protein networks

Cellular functions are rarely executed by isolated proteins; instead, they emerge from coordinated regulatory programs in which groups of proteins exhibit temporally coherent expression patterns (Maier et al., 2020). Such coordinated dynamics reflect shared regulatory control, pathway-level organization, or participation in common functional processes. To capture this structured temporal coordination at the network level, we translate activity trajectories into pairwise associations and construct a protein association graph (Carthew, 2021).

Given the activity-filtered expression matrix $X' \in \mathbb{R}^{N \times M}$ obtained in Section 2.1, the i -th row vector x'_i represents the temporal activity profile of protein i across M time points. We quantify the association strength between proteins i and j using the Pearson correlation coefficient computed over their activity trajectories, as shown in Equation 4:

$$\text{Sim}(i, j) = \frac{\text{Cov}(x'_i, x'_j)}{\sigma(x'_i)\sigma(x'_j)} \quad (4)$$

This similarity measure evaluates the consistency of temporal co-variation, thereby capturing coordinated activation and repression patterns under the observed biological condition.

Based on these similarity scores, we construct a static adjacency matrix $A = [a_{ij}]$ via threshold-based binarization, as shown in Equation 5:

$$a_{ij} = \begin{cases} 1, & \text{sim}(i, j) > \tau \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where τ controls the sparsity and connectivity of the resulting association network. Since Pearson correlation is symmetric, the adjacency matrix satisfies $a_{ij} = a_{ji}$.

The threshold τ determines the balance between preserving coordinated temporal structures and suppressing weak or potentially spurious associations. We systematically evaluated model performance across a range of τ values. Predictive performance improves as weak correlations are filtered out, reaches a maximum around $\tau = 0.6$, and declines when the network becomes overly sparse. Smaller thresholds retain excessive low-strength associations and increase noise, whereas larger thresholds fragment the graph and limit the modeling of structured temporal dependencies. Based on this empirical analysis, we set $\tau = 0.6$, achieving the best trade-off between predictive accuracy and topological stability for subsequent graph convolution.

3.3 Graph convolution with multi-order, attention-guided aggregation

Essentiality is a systems phenotype that emerges from coordinated cellular organization. Proteins tend to be essential when they execute non-redundant roles in core processes, such as macromolecular complex assembly, metabolic throughput, and regulatory control (Ljubojevic et al., 2021). In expression-derived association graphs, these roles are typically reflected by coherent contextual patterns: proteins within the same functional unit show tightly coupled temporal activity, while weaker associations may result from indirect effects, transient coupling, or measurement variability (Day et al., 2022). As a consequence, essential-protein prediction benefits from an aggregation scheme that emphasizes functionally coherent context and attenuates weak dependencies that can propagate non-specific signals.

To achieve this, CSGNN adopts attention-based message passing and integrates contextual evidence at two structural orders. Let $A = [a_{ij}]$ be the first-order adjacency matrix defined in Section 2.2, and define a second-order connectivity matrix as shown in Equation 6:

$$A^{(2)} = A \cdot A \quad (6)$$

Here $(A^{(2)})_{ij}$ indicates the number of length-two paths between proteins i and j ; in this work, we use the condition $(A^{(2)})_{ij} > 0$ only to identify two-hop reachable neighbors. The first-order structure captures direct co-expression coupling, whereas the two-hop structure captures shared partners and module proximity that arise from modular biological organization.

Let $H^{(0)} = X' \in \mathbb{R}^{N \times M}$ denote the input node feature matrix (Section 2.1), and let $h_i^{(0)} \in \mathbb{R}^M$ be the i -th row vector. We use $\text{ReLU}(\cdot)$ as the activation function and $\phi(\cdot)$ as the LeakyReLU used in attention scoring.

3.3.1 First-layer graph convolution

The first graph convolution layer is designed to extract contextual evidence that reflects the local functional environment of each protein. In cellular systems, essential proteins are rarely isolated entities; rather, they are typically embedded in tightly coordinated functional modules, including macromolecular complexes, regulatory circuits, and central metabolic pathways. Within such modules, proteins often exhibit synchronized temporal activity patterns because they are co-regulated, physically coupled, or functionally interdependent (Dudek et al., 2021). Consequently, direct neighbors in the correlation-guided association graph frequently correspond to proteins whose activity trajectories are strongly coordinated with that of the target protein, providing immediate evidence of functional coupling.

However, functional organization in biological networks extends beyond direct interactions. Proteins connected through shared partners may belong to overlapping sub-complexes, participate in adjacent steps of a pathway, or be co-regulated under a common upstream mechanism. Such two-hop connectivity captures a broader layer of module-level organization and may reveal indirect but biologically meaningful dependencies (Guimaraes, 2020). To reflect this hierarchical organization, we explicitly distinguish first-order and second-order structural contexts.

Formally, given the adjacency matrix $A = [a_{ij}]$, the first-order neighborhood of node i is defined as $\mathcal{N}_1(i) = \{j \mid a_{ij} = 1\}$, and the second-order neighborhood is defined as $\mathcal{N}_2(i) = \{j \mid A_{ij}^{(2)} > 0, j \neq i\}$, where $A^{(2)} = A^2$ and the condition $(A^{(2)})_{ij} > 0$ indicates that proteins i and j are connected by at least one two-hop path. The explicit separation of these two neighborhoods enables the model to treat direct co-expression coupling and broader module proximity as structurally distinct sources of contextual evidence.

Because not all neighbors contribute equally to essentiality inference, we introduce attention-based weighting within each structural order. Even within a coherent functional module, some neighbors exhibit stronger temporal coordination or tighter regulatory dependence with the target protein, whereas others may represent peripheral or condition-specific associations. To capture this heterogeneity, the model first projects node features using a learnable matrix $W^{(1)} \in \mathbb{R}^{d \times M}$. For each order $m \in \{1, 2\}$, a compatibility score between node i and its neighbor $j \in \mathcal{N}_m(i)$ is computed as shown in Equation 7:

$$e_{ij}^{(m)} = \phi\left(a^T \left[W^{(1)} h_i^{(0)} \parallel W^{(1)} h_j^{(0)} \right]\right) \quad (7)$$

where $q_m \in \mathbb{R}^{2d}$ is a learnable attention vector specific to structural order m , \parallel denotes concatenation, and $\phi(\cdot)$ is the LeakyReLU activation. The attention coefficient is then obtained by neighborhood-wise normalization, as shown in Equation 8:

$$a_{ij}^{(m)} = \frac{\exp\left(e_{ij}^{(m)}\right)}{\sum_{k \in \mathcal{N}_m(i)} \exp\left(e_{ik}^{(m)}\right)} \quad (8)$$

These coefficients quantify the relative contribution of each neighbor when updating node i , thereby enabling the model to emphasize neighbors whose representations are more compatible with the target in the projected feature space.

The order-specific aggregated representations are then computed as shown in Equation 9:

$$h_{i,m}^{(1)} = \text{ReLU} \left(\sum_{k \in \mathcal{N}_m(i)} a_{ij}^{(m)} W^{(1)} h_j^{(0)} \right) \quad (9)$$

Finally, the first-order and second-order representations are concatenated as shown in Equation 10:

$$h_i^{(1)} = \left[h_{i,1}^{(1)} \parallel h_{i,2}^{(1)} \right] \quad (10)$$

yielding a multi-scale embedding that simultaneously encodes immediate functional coupling and broader module-level proximity. In biological terms, this layer consolidates coherent local coordination while attenuating weak or inconsistent associations that may otherwise obscure functionally organized patterns relevant to essentiality.

3.3.2 Second-layer graph convolution

While the first layer primarily captures local functional coherence and module-proximal context, essentiality often depends on a protein's position within a broader and more distributed functional organization (Fuchs et al., 2020). In cellular systems, essential proteins frequently serve as structural scaffolds of complexes, bottlenecks in metabolic pathways, or coordinators of regulatory programs. Their indispensability may therefore arise from coordinated activity patterns spanning multiple interconnected substructures rather than from a single tightly coupled neighborhood (Morris et al., 2022).

After the first layer, the node representations $h^{(1)}$ already encode refined local consistency and two-hop contextual information. The second layer operates on this enriched representation space and extends information propagation, allowing the model to further consolidate distributed dependencies across overlapping modules or sub-complexes. In this stage, neighbor importance is re-evaluated in light of functionally integrated embeddings rather than raw activity trajectories.

Specifically, a second projection matrix $W^{(2)} \in \mathbb{R}^{d \times 2d}$ is applied to first-layer embeddings. For each structural order $m \in \{1, 2\}$, attention coefficients $\beta_{ij}^{(m)}$ are computed analogously to the first layer, but using $h^{(1)}$ as input. The order-specific updates are defined as shown in Equation 11:

$$h_{i,m}^{(2)} = \text{ReLU} \left(\sum_{j \in \mathcal{N}_m(i)} \beta_{ij}^{(m)} W^{(2)} h_j^{(1)} \right) \quad (11)$$

The final node embedding is obtained by concatenation, as shown in Equation 12:

$$h_i^{(2)} = \left[h_{i,1}^{(2)} \parallel h_{i,2}^{(2)} \right] \quad (12)$$

Through this second aggregation stage, contextual evidence is refined at a higher level of organization. The first layer emphasizes immediate coordination and module adjacency, whereas the second layer integrates more distributed patterns across connected functional regions. The resulting embeddings thus capture essentiality-related signals at complementary biological scales and provide a multi-scale, biologically informed representation for downstream node-level prediction.

3.4 Essential protein identification

After two graph convolution layers, each protein is represented by a multi-scale embedding that integrates direct co-expression coupling and two-hop module-proximal context. In cellular systems, essentiality typically emerges from a protein's participation in coordinated functional organization rather than from isolated molecular activity. Many essential proteins are indispensable because they serve as core components of macromolecular complexes, maintain flux through critical metabolic routes, or coordinate central regulatory programs (Monzel et al., 2023). These roles are inherently contextual: the functional impact of a protein is expressed through stable and consistent relationships with its partners and surrounding functional neighborhood. Accordingly, a reliable essentiality predictor should capture not only node-intrinsic signals but also how coherently a protein aligns with informative contextual proteins in its neighborhood.

To incorporate such neighborhood-consistency evidence while maintaining a strictly node-level prediction target, we construct interaction-aware descriptors between each target protein and its contextual proteins, and aggregate them into a single node-centric representation.

Let $H_1^{(2)} \in \mathbb{R}^{N \times d}$ and $H_2^{(2)} \in \mathbb{R}^{N \times d}$ denote the second-layer outputs obtained under the first-order and second-order structural contexts, respectively. We concatenate them to obtain the final embedding matrix, as shown in Equation 13:

$$H_{final} = [H_1^{(2)} \| H_2^{(2)}] \in \mathbb{R}^{N \times 2d} \quad (13)$$

The embedding of protein i is $h_i \in \mathbb{R}^{2d}$, i.e., the i -th row of H_{final} . We define the contextual protein set for i as $\mathcal{C}(i) = \mathcal{N}_1(i) \cup \mathcal{N}_2(i)$, which includes proteins connected to i either directly or through two-hop reachability in the correlation-guided association graph. This contextual set provides a structured approximation of the functional neighborhood in which the essentiality of protein i is manifested.

For each contextual protein $j \in \mathcal{C}(i)$, we construct an interaction descriptor that summarizes complementary aspects of representation compatibility between the target and its context. Specifically, element-wise addition captures shared trends in embedding magnitude and direction, as shown in Equation 14:

$$h_{add}^{(ij)} = h_i + h_j \in \mathbb{R}^{2d} \quad (14)$$

while element-wise multiplication emphasizes feature-wise agreement by highlighting dimensions that are simultaneously strong in both proteins, as shown in Equation 15:

$$h_{prod}^{(ij)} = h_i \odot h_j \quad (15)$$

In biological terms, this feature-wise agreement is consistent with the fact that essentiality-related evidence often concentrates on coherent functional signals shared by proteins within tightly coordinated modules. We further include ordered concatenation, as shown in Equation 16:

$$h_{cat}^{(ij)} = [h_i \| h_j] \in \mathbb{R}^{4d} \quad (16)$$

which preserves the target and context embeddings without forcing early mixing. Although the association graph is undirected, the

descriptor is constructed in a target-context manner (centered at protein i) to characterize how well each neighbor j matches the functional context of i , rather than to impose directionality on edges.

We then concatenate the interaction terms to form the pair-level compatibility descriptor, as shown in Equation 17:

$$h_{pair}^{(ij)} = [h_{add}^{(ij)} \| h_{prod}^{(ij)} \| h_{cat}^{(ij)}] \in \mathbb{R}^{8d} \quad (17)$$

Since the prediction target remains protein i , we aggregate these pair descriptors across $\mathcal{C}(i)$ into a single interaction-aware representation, as shown in Equation 18:

$$\bar{h}_i = \sum_{\mathcal{C}(i)} w_{ij} h_{pair}^{(ij)} \quad (18)$$

The weights w_{ij} are derived from the learned attention coefficients in the second graph convolution layer and then re-normalized over the union context set $\mathcal{C}(i)$. This design ensures that neighbors exhibiting stronger functional consistency with protein i contribute more strongly to the aggregated descriptor, aligning the final decision with coherent module-level evidence rather than with weak or noisy associations. Importantly, although interaction descriptors are constructed at the protein-pair level, the aggregation step yields exactly one representation \bar{h}_i per node, and the prediction therefore remains strictly node-level.

Finally, we map \bar{h}_i to an essentiality probability using a two-layer decoder, as shown in Equations 19, 20:

$$z = \text{ReLU}(W_1 \bar{h}_i + b_1) \quad (19)$$

$$\hat{y}_i = \text{Sigmoid}(W_2 z_i + b_2) \quad (20)$$

where $W_1 \in \mathbb{R}^{d_h \times 8d}$, $b_1 \in \mathbb{R}^{d_h}$, $W_2 \in \mathbb{R}^{1 \times d_h}$, $b_2 \in \mathbb{R}$, and d_h is the hidden dimension. For discrete evaluation, we produce hard labels using a decision threshold δ , as shown in Equation 21:

$$\tilde{y}_i = \begin{cases} 1, & \hat{y}_i \geq \delta \\ 0, & \hat{y}_i < \delta \end{cases} \quad (21)$$

In this study, we use $\delta = 0.5$ as the default decision threshold. Since \hat{y}_i is produced by a Sigmoid function, $\delta = 0.5$ corresponds to the neutral decision boundary where the model assigns equal posterior preference to the essential and non-essential classes.

4 Experimental data

To evaluate the effectiveness of the proposed CSGNN framework across distinct biological contexts, we conducted experiments on two well-studied model organisms: *Saccharomyces cerevisiae* (yeast) and *E. coli*. Yeast is a canonical eukaryotic system with extensively curated protein-protein interaction (PPI) resources and essentiality annotations, and it has been widely used as a benchmark for essential protein identification. In contrast, *Escherichia coli* is a representative prokaryotic organism with a relatively compact regulatory architecture. Using both organisms allows us to examine the behavior of the proposed method under different network sizes, densities, and biological mechanisms.

The PPI networks of yeast and *E. coli* were obtained from the DIP database (Xenarios et al., 2000). During preprocessing,

duplicate interactions and self-interactions were removed to ensure structural consistency. After filtering, the yeast PPI network contains 5,093 proteins and 24,743 interactions, while the *E. coli* network contains 2,727 proteins and 11,803 interactions.

Time-series gene expression data were collected from the Gene Expression Omnibus (Clough and Barrett, 2016). For yeast, dataset GSE3431 contains 6,777 gene products measured at 36 time points, of which 4,858 genes were mapped to proteins in the yeast PPI network. For *E. coli*, dataset GSE3905 contains 7,312 gene products measured at 8 time points. These temporal profiles provide the basis for constructing correlation-guided interaction graphs.

Essential protein annotations were compiled from curated biological databases. For yeast, essentiality information was integrated from MIPS (Pagel et al., 2005), SGD (Dwight et al., 2002), DEG (Zhang and Lin, 2009), and SGDP (Giaever and Nislow, 2014). After mapping to the PPI network, 1,167 essential proteins were retained in the yeast interaction network. For *E. coli*, 254 essential proteins were identified within the constructed PPI network.

In our learning setting, each protein corresponds to a node in a fixed organism-specific network, and the task is to predict a binary essentiality label for each node. The correlation-guided graph is constructed using the available expression profiles and the organism-level interaction network, and model training is performed as node-level supervised learning on this shared graph. We adopt a stratified random split over nodes, with 70%/10%/20% for training/validation/test, while preserving the ratio of essential to non-essential proteins. Only training labels are used to optimize model parameters, and the validation set is used for model selection. Test labels are held out for final evaluation. This design matches a common biological use case: essentiality is experimentally known for only a subset of proteins, and the goal is to infer essential proteins for the remaining uncharacterized proteins within the same organism-level interaction context.

5 Simulation experiments

5.1 Evaluation indicators

In this study, multiple evaluation metrics are adopted to comprehensively assess the performance of the proposed model on the essential protein prediction task, including Top-N accuracy, area under the ROC curve (AUC), Recall (REC), Matthews correlation coefficient (MCC), F1 score, and Positive Predictive Value (PPV).

Top-N accuracy measures the number of true essential proteins correctly identified among the top N proteins ranked by their predicted essentiality scores, reflecting the model's ability to prioritize the most critical candidate proteins. This metric is particularly relevant for biological screening scenarios, where only a limited number of top-ranked proteins can be experimentally validated.

The AUC quantifies the area under the receiver operating characteristic (ROC) curve and evaluates the overall discriminative capability of the model across different decision thresholds. A higher AUC value indicates stronger robustness and generalization in distinguishing essential proteins from non-essential ones.

Recall (REC) measures the proportion of true essential proteins that are correctly identified by the model, reflecting its sensitivity in

detecting essential proteins. Matthews correlation coefficient (MCC) is a balanced metric that jointly considers true positives, false positives, true negatives, and false negatives, making it particularly suitable for datasets with class imbalance. The F1 score represents the harmonic mean of precision and recall, capturing the trade-off between prediction accuracy and coverage. PPV indicates the proportion of proteins predicted as essential that are indeed essential, reflecting the reliability of the prediction results.

Together, these metrics provide a comprehensive and quantitative evaluation of the model's performance from multiple perspectives, enabling a thorough assessment of its effectiveness in essential protein identification, as shown in Equations 22, 23:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (22)$$

$$PPV = \frac{TP}{TP + FP} \quad (23)$$

5.2 Experimental environment and parameter settings

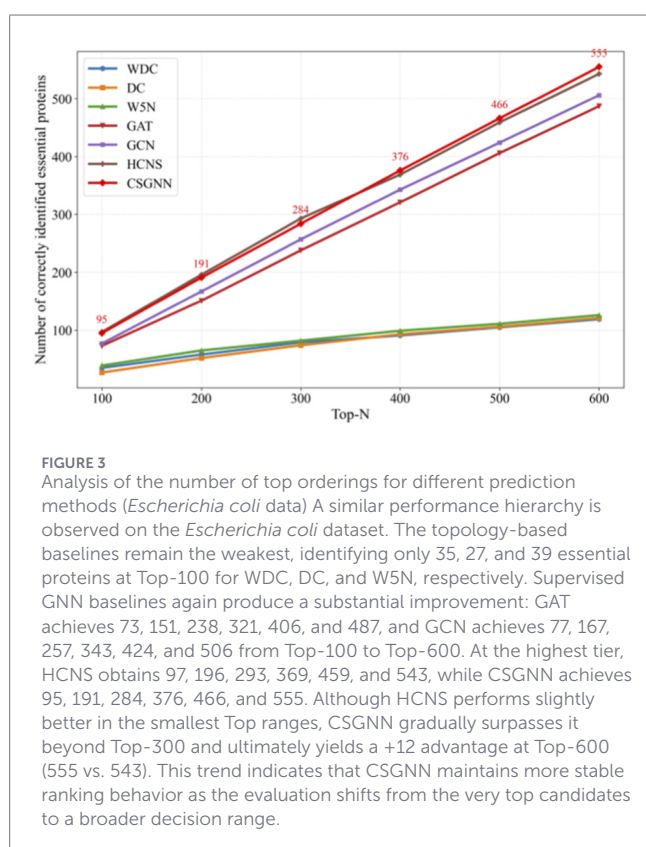
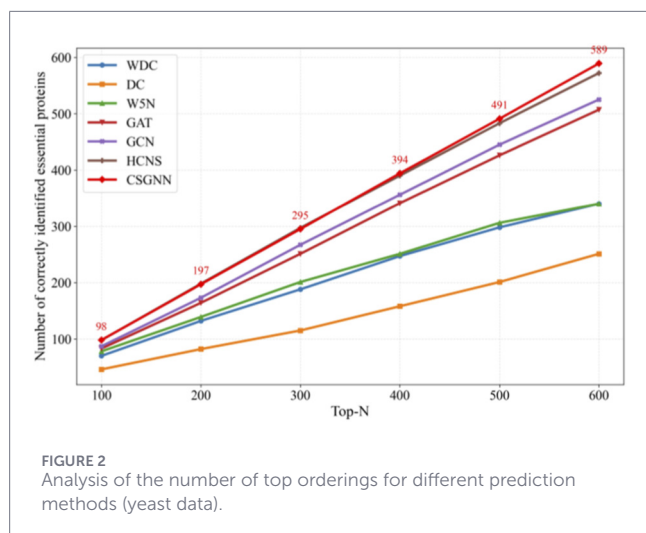
In this experiment, we performed model training based on the PyTorch framework using Nvidia GeForce RTX 3090 GPUs and the CUDA 12.1 environment, which ensured the efficiency of the training process. To optimize the training process of the model, we set the learning rate to 1×10^{-4} , the batch size to 64, and used the AdamW optimizer combined with the weight decay coefficient 1×10^{-4} to enhance the generalization ability of the model. To dynamically adjust the learning rate, we introduced the ReduceLROnPlateau scheduler, which reduces the learning rate to 0.5 times of the original one when the validation set performance (AUROC) is not improved for 3 consecutive cycles. The training process was carried out for 30 cycles and all training data were normalized and combined with a randomized data loading strategy (shuffle = True) to enhance data diversity. The model performance is monitored in real-time via TensorBoard, which records a number of evaluation metrics such as AUC, F1 score, and Top-100 to ensure comprehensive tracking and analysis of the model performance.

5.3 Experimental results and analysis

5.3.1 Top sort analysis

To further evaluate the ranking capability of CSGNN, we conducted a Top-N analysis on the yeast and *E. coli* datasets. For each model, we report the number of correctly identified essential proteins within the Top-100 to Top-600 ranked candidates. In addition to three representative topology-based methods, WDC (Tang et al., 2013), DC (Hahn and Kern, 2005), and W5N (Li et al., 2017), we include two standard supervised GNN baselines, namely, a vanilla two-layer GCN and a two-layer multi-head GAT implemented under the same data splits and training protocol. We further compare against HCNS, a recent state-of-the-art framework integrating hypergraph convolution with deep sequence modeling. The results are illustrated in Figures 2, 3.

On the yeast dataset, traditional topology-driven methods show limited ranking capability. At Top-100, WDC, DC, and W5N



identify 70, 46, and 78 essential proteins, respectively. Although their performance increases gradually with larger Top-N ranges, their overall growth remains considerably below that of learning-based approaches. Introducing supervised message passing substantially improves ranking quality. The two-layer GAT achieves 83, 164, 251, 341, 426, and 507 correct identifications from Top-100 to Top-600, while the two-layer GCN further improves these numbers to 87, 173, 267, 356, 445, and 525. The consistent advantage of GCN over GAT across all Top ranges suggests that, in yeast PPI networks, stable normalized neighborhood aggregation provides a

more robust ranking signal than attention-based reweighting under noisy interaction edges and class imbalance.

Building upon stronger structural modeling, HCNS and CSGNN further enhance the ranking performance. HCNS reaches 98, 198, 297, 390, 483, and 572 across Top-100 to Top-600, benefiting from the integration of hypergraph structure and deep sequence features. CSGNN achieves 98, 197, 295, 394, 491, and 589, remaining competitive in the smallest Top ranges while showing clearer advantages as the ranking scope expands. In particular, at Top-600 CSGNN identifies 589 essential proteins compared with 572 for HCNS, indicating stronger global ranking stability when a broader set of candidate proteins is considered.

Taken together, the results across both datasets reveal a clear methodological progression. Traditional centrality-based approaches rely solely on static topological indicators and therefore provide limited ranking discrimination. Standard GCN and GAT introduce supervised neighborhood aggregation and substantially improve the ordering of candidate proteins, yet remain constrained by shallow message passing on a fixed graph. HCNS further incorporates hypergraph modeling and sequence-level representations, enhancing structural and functional feature integration. CSGNN extends this progression by constructing a correlation-guided dynamic interaction network and propagating information over multi-order neighborhoods, enabling interaction-aware representations that better align with the biological characteristics of essential proteins. This design leads to consistently improved global ranking quality, particularly in larger Top-N regimes where ordering stability becomes critical.

5.3.2 Comprehensive multi-metric performance evaluation

To comprehensively evaluate the classification performance of CSGNN, we conduct a multi-metric assessment on the yeast and *E. coli* datasets. The evaluation metrics include AUC, recall (REC), F1 score, Matthews correlation coefficient (MCC), and positive predictive value (PPV). These metrics measure discrimination ability, coverage of true essential proteins, precision-recall balance, and robustness under class imbalance.

Under the supervised node classification framework, each protein in the PPI network is treated as a node-level sample. Proteins annotated as essential in curated biological databases are labeled as positive samples. All remaining proteins are labeled as negative samples. A stratified random partition strategy is adopted. The data are divided into training, validation, and test sets at a ratio of 70%, 10%, and 20%. The proportion of essential and non-essential proteins is preserved in each subset. Model parameters are optimized using only the training set. The validation set is used for hyperparameter tuning. The test set is strictly held out for final evaluation.

The results on the yeast dataset are shown in Table 1. The three topology-based baselines (WDC, DC, and W5N) show limited performance. Their AUC values range from 0.6705 to 0.7152. Their MCC values remain relatively low. These results indicate that static centrality measures or shallow fusion strategies are insufficient for capturing complex interaction patterns in PPI networks.

Supervised graph learning substantially improves performance. The two-layer GAT achieves an AUC of 0.7832 and an MCC of

TABLE 1 Multi - performance evaluation analysis (yeast data).

Dataset	Method	AUC	REC	MCC	F1	PPV
Yeast	WDC	0.6893	0.4576	0.2967	0.4578	0.4580
	DC	0.6705	0.4002	0.2219	0.4002	0.4002
	W5N	0.7152	0.4747	0.3186	0.4747	0.4747
	GAT	0.7832	0.7212	0.4872	0.6122	0.6035
	GCN	0.8032	0.7412	0.5232	0.6382	0.6286
	HCNS	0.8772	0.8538	0.6326	0.7936	0.7735
	CSGNN	0.8993	0.8682	0.6284	0.8189	0.7966

TABLE 2 Multi-performance evaluation analysis (*Escherichia coli* data).

Dataset	Method	AUC	REC	MCC	F1	PPV
<i>E. coli</i>	WDC	0.6837	0.2323	0.1534	0.2323	0.2322
	DC	0.6849	0.2559	0.1795	0.2559	0.2559
	W5N	0.7243	0.2913	0.2186	0.2913	0.2913
	GAT	0.7531	0.6949	0.4374	0.6328	0.6214
	GCN	0.7721	0.7032	0.4574	0.6609	0.6532
	HCNS	0.8182	0.7430	0.5123	0.7248	0.7339
	CSGNN	0.7973	0.7505	0.4713	0.7394	0.7487

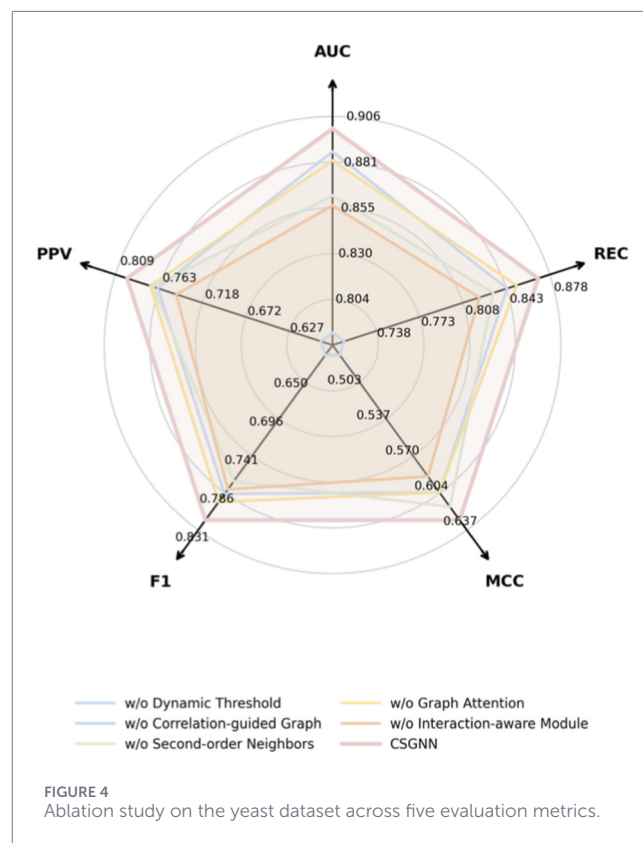
0.4872. The two-layer GCN further improves these values to 0.8032 and 0.5232. GCN consistently outperforms GAT across all metrics. This suggests that normalized neighborhood aggregation provides a stable ranking signal under noisy interactions and moderate class imbalance.

HCNS further enhances structural modeling by integrating hypergraph convolution with deep sequence encoders. It achieves an AUC of 0.8772 and an MCC of 0.6326. CSGNN achieves an AUC of 0.8993, a recall of 0.8682, an F1 score of 0.8189, and a PPV of 0.7966. Its MCC is 0.6284, which is comparable to HCNS. However, CSGNN shows higher AUC, recall, and F1. These results indicate stronger global discrimination and a more balanced precision–recall trade-off.

The results on the *E. coli* dataset are presented in Table 2. A similar hierarchy is observed. The topology-based methods achieve AUC values between 0.6837 and 0.7243. Their recall values remain low. The two-layer GAT reaches an AUC of 0.7531 and an MCC of 0.4374. The two-layer GCN improves these to 0.7721 and 0.4574. GCN again performs slightly better than GAT.

HCNS achieves the highest AUC (0.8182) and MCC (0.5123) on this dataset. CSGNN achieves an AUC of 0.7973, a recall of 0.7505, an F1 score of 0.7394, and a PPV of 0.7487. Although its AUC and MCC are slightly lower than those of HCNS, CSGNN obtains higher recall and competitive F1 performance. This indicates that CSGNN identifies a larger proportion of true essential proteins while maintaining stable predictive precision under stronger class imbalance.

Across both datasets, a clear methodological progression can be observed. Traditional centrality-based methods rely on static



topology. Their discrimination capacity is limited. Standard GCN and GAT introduce supervised neighborhood aggregation and significantly improve classification performance. However, they are constrained by shallow message passing on a fixed interaction graph. HCNS integrates hypergraph structure and sequence-level representations, which strengthens structural and functional modeling. CSGNN further extends this progression. It constructs correlation-driven interaction graphs and performs multi-order propagation. This design captures dynamic coordination patterns associated with essentiality. As a result, CSGNN achieves strong recall and balanced multi-metric performance across species.

5.4 Ablation experiment

To further investigate the contribution of each architectural component in CSGNN, we conduct a systematic ablation study on the yeast dataset. The evaluation includes AUC, recall (REC), MCC, F1 score, and PPV. The full CSGNN model achieves 0.8993 AUC, 0.8682 recall, 0.6284 MCC, 0.8189 F1, and 0.7966 PPV. Figure 4 illustrates the radar comparison among the full model and its simplified variants.

We first examine the effect of removing the dynamic thresholding mechanism. Without dynamic thresholding, AUC decreases to 0.8865 and recall drops to 0.8432. The decline is consistent across all metrics. This indicates that adaptive correlation filtering improves the quality of the constructed interaction graph. Static thresholds fail to fully capture condition-dependent expression relationships, which weakens the global discrimination ability of the model.

The most significant performance degradation occurs when the correlation-guided graph construction is removed. In this case, AUC drops to 0.7857 and MCC decreases to 0.4783. F1 and PPV are also substantially reduced. This confirms that the correlation-driven graph is the structural foundation of CSGNN. Without this module, the model essentially degenerates into a conventional graph learning framework on a static topology. The results demonstrate that correlation-guided graph construction contributes the largest share of performance gain.

When second-order neighborhood propagation is removed, the performance shows a moderate but consistent decline. AUC decreases to 0.8621 and recall to 0.8321. This suggests that higher-order structural information enhances the model's ability to capture functional coordination patterns beyond immediate neighbors. Essential proteins often participate in coordinated biological processes that span multiple interaction hops. Limiting propagation to first-order neighborhoods reduces the capacity to model these patterns.

Removing the graph attention mechanism also leads to measurable degradation. AUC decreases to 0.8815 and MCC to 0.6032. Although the performance drop is smaller than that observed for graph construction removal, attention-based weighting improves the discrimination of heterogeneous interaction strengths. Uniform aggregation introduces additional noise, especially in densely connected regions of the PPI network.

The interaction-aware representation module also plays an important role. Without this component, AUC decreases to 0.8565 and recall to 0.8209. This module enables context-sensitive feature refinement based on interaction patterns. Its removal weakens the model's ability to align structural signals with functional indispensability.

Overall, the ablation results reveal a clear hierarchy of component importance. Correlation-guided graph construction provides the dominant structural advantage. Dynamic thresholding and multi-order propagation further enhance topological expressiveness. Attention weighting and interaction-aware modeling refine representation quality. The consistent degradation across all simplified variants confirms that the performance improvement of CSGNN does not arise from a single architectural modification, but from the coordinated integration of dynamic graph construction and multi-level representation learning.

6 Conclusion

In this study, an essential protein prediction method based on Correlation-guided Subgraph Graph Neural Network (CSGNN) is proposed considering the advantages of graph neural networks in processing graph data. The experimental results of applying it to yeast and *E. coli* datasets show that the model exhibits high prediction performance in several key indicators, especially in dealing with the dynamic network characteristics, and is able to identify the essential proteins more accurately than the traditional methods. This validates the potential of graph neural networks for applications in complex biological networks. Future research can further advance the progress of essential protein prediction in several directions. The potential of multimodal data fusion can be explored by combining more biological data, such as gene

mutations, protein structural information, and cell type-specific gene expression data, to further improve prediction accuracy. In addition, cross-species prediction of essential proteins can be an important direction to improve the generalization ability of the model by using the shared knowledge between species. With the continuous accumulation of large-scale biological data, the potential of model application on different tasks and datasets can also be enhanced in the future through methods such as migration learning. These directions will promote the widespread application of essential protein prediction techniques to aid biological and medical research.

Data availability statement

The source code and datasets supporting the findings of this study are publicly available at: <https://github.com/Lzx2294159762/CSGNN>.

Author contributions

ZL: Visualization, Software, Data curation, Writing – original draft, Methodology, Formal Analysis, Conceptualization. ZY: Investigation, Validation, Software, Visualization, Writing – original draft, Data curation. PL: Supervision, Writing – review and editing, Funding acquisition, Resources.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was supported by the Hunan Provincial Undergraduate Innovation Training Program (Grant No. S202510541129).

Acknowledgements

The authors would like to thank PL for his valuable guidance and continuous support throughout this research.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Ali, F., Almuhaimeed, A., Khalid, M., Alshabari, H., Masmoudi, A., and Alsin, R. (2024). DEEP-EP: identification of epigenetic protein by ensemble residual convolutional neural network for drug discovery. *Methods* 226, 49–53. doi:10.1016/j.ymeth.2024.04.004
- Ballantyne, C. M., Vasas, S., Azizad, M., Clifton, P., Rosenson, R. S., Chang, T., et al. (2024). Plozasiran, an RNA interference agent targeting APOC3, for mixed hyperlipidemia. *N. Engl. J. Med.* 391 (10), 899–912. doi:10.1056/NEJMoa2404143
- Carthew, R. W. (2021). Gene regulation and cellular metabolism: an essential partnership. *Trends Genet.* 37 (4), 389–400. doi:10.1016/j.tig.2020.09.018
- Clough, E., and Barrett, T. (2016). *The gene expression omnibus database statistical genomics: Methods and protocols*. New York, NY: Springer, 93–110.
- Cubuk, J., Alston, J. J., Incicco, J. J., Singh, S., Stuchell-Brereton, M. D., Ward, M. D., et al. (2021). The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nat. Communications* 12 (1), 1936. doi:10.1038/s41467-021-21953-3
- Day, L., Cakebread, J. A., and Loveday, S. M. (2022). Food proteins from animals and plants: differences in the nutritional and functional properties. *Trends Food Sci. & Technol.* 119, 428–442. doi:10.1016/j.tifs.2021.12.020
- Dudek, M., Angelucci, C., Pathiranage, D., Wang, P., Mallikarjun, V., Lawless, C., et al. (2021). Circadian time series proteomics reveals daily dynamics in cartilage physiology. *Osteoarthritis Cartilage* 29 (5), 739–749. doi:10.1016/j.joca.2021.02.008
- Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., et al. (2002). Saccharomyces genome database (SGD) provides secondary gene annotation using the gene ontology (GO). *Nucleic Acids Research* 30 (1), 69–72. doi:10.1093/nar/30.1.69
- Fuchs, P., Rugen, N., Carrie, C., Elsässer, M., Finkemeier, I., Giese, J., et al. (2020). Single organelle function and organization as estimated from arabidopsis mitochondrial proteomics. *Plant J.* 101 (2), 420–441. doi:10.1111/tpj.14534
- Giaever, G., and Nislow, C. (2014). The yeast deletion collection: a decade of functional genomics. *Genetics* 197 (2), 451–465. doi:10.1534/genetics.114.161620
- Goh, K. J., Stubenrauch, C. J., and Lithgow, T. (2024). The TAM, a translocation and assembly module for protein assembly and potential conduit for phospholipid transfer. *EMBO Reports* 25 (4), 1711–1720. doi:10.1038/s44319-024-00111-y
- Guimaraes, Jr P. R. (2020). The structure of ecological networks across levels of organization. *Annu. Rev. Ecol. Syst.* 51 (1), 433–460. doi:10.1146/annurev-ecolsys-012220-120819
- Hahn, M. W., and Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biology Evolution* 22 (4), 803–806. doi:10.1093/molbev/msi072
- Inzamam-Ul-Hossain, M., and Islam, M. R. (2023). Identification of essential protein using chemical reaction optimization and machine learning technique. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20 (3), 2122–2135. doi:10.1109/TCBB.2022.3233473
- Jones, S., and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl. Acad. Sci.* 93 (1), 13–20. doi:10.1073/pnas.93.1.13
- Li, M., Ni, P., Chen, X., Wang, J., Wu, F. X., and Pan, Y. (2017). Construction of refined protein interaction network for predicting essential proteins. *IEEE/ACM Transactions Computational Biology Bioinformatics* 16 (4), 1386–1397. doi:10.1109/TCBB.2017.2665482
- Liu, Q., Han, R., Yu, D., Wang, Z., Zhuansun, X., and Li, Y. (2024). Characterization of thyme essential oil composite film based on soy protein isolate and its application in the preservation of cherry tomatoes. *Lwt* 191, 115686. doi:10.1016/j.lwt.2023.115686
- Ljubojevic, N., Henderson, J. M., and Zurzolo, C. (2021). The ways of actin: why tunneling nanotubes are unique cell protrusions. *Trends Cell Biol.* 31 (2), 130–142. doi:10.1016/j.tcb.2020.11.008
- Maier, B., Leader, A. M., Chen, S. T., Tung, N., Chang, C., LeBerichel, J., et al. (2020). A conserved dendritic-cell regulatory program limits antitumour immunity. *Nature* 580 (7802), 257–262. doi:10.1038/s41586-020-2134-y
- Monzel, A. S., Enríquez, J. A., and Picard, M. (2023). Multifaceted mitochondria: moving mitochondrial science beyond function and dysfunction. *Nat. Metabolism* 5 (4), 546–562. doi:10.1038/s42255-023-00783-1
- Morris, R., Black, K. A., and Stollar, E. J. (2022). Uncovering protein function: from classification to complexes. *Essays Biochem.* 66 (3), 255–285. doi:10.1042/EBC20200108
- Pacesa, M., Pelea, O., and Jinek, M. (2024). Past, present, and future of CRISPR genome editing technologies. *Cell* 187 (5), 1076–1100. doi:10.1016/j.cell.2024.01.042
- Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., et al. (2005). The MIPS Mammalian protein-protein interaction database. *Bioinformatics* 21 (6), 832–834. doi:10.1093/bioinformatics/bti115
- Pan, H., and Finkel, T. (2017). Essential protein s and pathways that regulate lifespan. *J. Biol. Chem.* 292 (16), 6452–6460. doi:10.1074/jbc.R116.771915
- Portugal, I., Alencar, P., and Cowan, D. (2018). The use of machine learning algorithms in recommender systems: a systematic review. *Expert Systems with Applications* 97, 205–227. doi:10.1016/j.eswa.2017.12.020
- Saha, S., Chatterjee, P., Basu, S., and Nasipuri, M. (2024). Epi-sf: essential protein identification in protein interaction networks using sequence features. *PeerJ* 12, e17010. doi:10.7717/peerj.17010
- Savini, I., Rossi, A., Pierro, C., Avigliano, L., and Catani, M. V. (2008). SVCT1 and SVCT2: essential protein s for vitamin C uptake. *Amino Acids* 34, 347–355. doi:10.1007/s00726-007-0555-7
- Shor, B., and Schneidman-Duhovny, D. (2024). CombFold: predicting structures of large protein assemblies using a combinatorial assembly algorithm and AlphaFold2. *Nat. Methods* 21 (3), 477–487. doi:10.1038/s41592-024-02174-0
- Tang, X., Wang, J., Zhong, J., and Pan, Y. (2013). Predicting essential proteins based on weighted degree centrality. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 11 (2), 407–418. doi:10.1109/TCBB.2013.2295318
- Tian, J., Lu, P., and Sha, H. (2025). HCNS: a deep learning model for identifying essential proteins based on hypergraph convolution and sequence features. *Anal. Biochem.* 707, 115949. doi:10.1016/j.ab.2025.115949
- Walkinshaw, M. D. (1992). Protein targets for structure-based drug design. *Med. Research Reviews* 12 (4), 317–372. doi:10.1002/med.2610120403
- Wang, J., Watson, J. L., and Lisanza, S. L. (2024). Protein design using structure-prediction networks: alphafold and RoseTTAFold as protein structure foundation models. *Cold Spring Harb. Perspect. Biol.* 16 (7), a041472. doi:10.1101/cshperspect.a041472
- Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Research* 28 (1), 289–291. doi:10.1093/nar/28.1.289
- Yang, P., Lu, P., and Zhang, T. (2023). AG-GATCN: a novel method for predicting essential proteins. *Chin. Phys. B* 32 (5), 058902. doi:10.1088/1674-1056/acb9f9
- Zeng, M., Wang, N., and Wu, Y. (2021). “Improving human essential protein prediction using only protein sequences via ensemble learning,” in IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 09–12 December 2021 (IEEE), 98–103.
- Zhang, R., and Lin, Y. (2009). DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Research* 37 (Suppl. 1_1), D455–D458. doi:10.1093/nar/gkn858
- Zhiguo, Y., Zixuan, L., and Peng, L. (2025). MultiRepPI: a cross-modal feature fusion-based multiple characterization framework for plant peptide-protein interaction prediction. *BMC Plant Biol.* 25 (1), 933. doi:10.1186/s12870-025-06878-z