applied sciences

MDPI

Article

GBNSS: A Method Based on Graph Neural Networks (GNNs) for Global Biological Network Similarity Search

Yi Wang [1], Feng Zhan [1,2,*], Cuiyu Huang [3] and Yiran Huang [1,2,*]

[1] School of Computer and Electronics Information, Guangxi University, Nanning 530004, China
[2] Guangxi Key Laboratory of Multimedia Communications Network Technology, Nanning 530004, China
[3] Tianjin Key Laboratory of Biosensing and Molecular Recognition, College of Chemistry, Nankai University, Tianjin 300071, China
* Correspondence: zf@gxu.edu.cn (F.Z.); hyr@gxu.edu.cn (Y.H.)

Abstract: Biological network similarity search plays a crucial role in the analysis of biological networks for human disease research and drug discovery. A biological network similarity search aims to efficiently identify novel networks biologically homologous to the query networks. Great progress has been achieved in biological network similarity searches. However, it remains a challenge to mine the biological network information fully to improve the accuracy of query results without increasing time overheads. In this study, we propose a biological network similarity search method based on graph neural networks named GBNSS, which combines topological and biological information (GO annotations) of biological networks into graph neural networks to find topologically and biologically similar biological networks in the database. Additionally, GBNSS is a topology-free biological network similarity search method with an arbitrary network structure. The experimental results on four benchmark datasets show that GBNSS outperforms the existing methods in terms of computational efficiency and search accuracy. Case studies further demonstrate that GBNSS is capable of searching similar networks in real-world biological networks.

Keywords: biological network search; graph neural networks; graph convolutional networks; GO annotations

check for updates

Citation: Wang, Y.; Zhan, F.; Huang, C.; Huang, Y. GBNSS: A Method Based on Graph Neural Networks (GNNs) for Global Biological Network Similarity Search. *Appl. Sci.* **2024**, *14*, 9844. https://doi.org/10.3390/app14219844

Academic Editor: Antonio Pagliaro

Received: 31 August 2024
Revised: 17 October 2024
Accepted: 21 October 2024
Published: 28 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Biological networks are intricate systems comprising molecular interactions within organisms [1], including protein–protein interactions [2], gene regulatory mechanisms [3], and metabolic reactions [4]. Biological network similarity search is a crucial component of biological network analysis. This approach has become a significant research field as large-scale biological data accumulate. It aims to identify common features and functional patterns among different biological networks, facilitating the application of established knowledge to uncharted biological networks and providing novel perspectives on unknown properties [5–8].

The growing availability of biological network data has made biological network similarity searches increasingly relevant across various research areas. A significant challenge in the post-genomics era is deciphering evolutionary mechanisms through molecular interactions and reactions within cells. Global biological network similarity search plays a crucial role in advancing our ability to understand molecular evolution [9], identify functional orthologs [10], recognize enzyme clusters [11], aid in disease diagnosis [12], and pinpoint drug targets [13,14]. Despite considerable advancements in biological network similarity search techniques, there remains a balance to be struck between speed and accuracy. New strategies for biological network similarity searches are needed to meet diverse requirements.

Biological network similarity searches are typically categorized into two types: exact query and approximate query. Each is designed to meet specific research needs. Exact query

methods insist that the search results must match a subgraph of the original query network, focusing on graph isomorphism—an NP-complete problem [15] that becomes impractical as network size increases. In contrast, approximate query methods are more flexible, prioritizing similarity in network topology or biological characteristics. These methods are particularly useful for accommodating human error and enhancing computational speed. Existing biological similarity search methods focus on reducing query time and improving query accuracy. These methods to minimize query time can be categorized into building indexes and leveraging parallel computing resources. For instance, [16–18] reduce query time by creating a biological database network index and minimizing network size. Similarly, methods [19,20] enhance efficiency by utilizing CPU computing resources through parallelism. Methods to improve query accuracy are generally classified into three categories: those based on biological network topology, those based on biological sequence similarity, and those combining both. For example, [21–24] use network topology such as important nodes or converting network topology to trees, while [25–27] leverage biological sequence similarity for biochemical accuracy. Methods like [28–30] combine subgraph signature and sequence similarity through model training for more accurate results.

Despite progress, existing methods for biological network similarity search have notable drawbacks. Index-based query methods are limited by the topology of the network. Parallel computing approaches, while effective, are limited by the number of network nodes and struggle with networks exceeding 100 nodes. Moreover, methods leveraging topological and biological information enhance accuracy but face significant increases in query time with larger networks.

To address these issues, we propose GBNSS (graph neural networks for biological network similarity search), a graph neural network-based query algorithm for biological networks. GBNSS is a topology-free method capable of handling biological networks with arbitrary structures. GBNSS extracts topological features and biochemical information from nodes in a biological network, builds a feature representation for each network, and fully captures the network's information by aggregating neighboring nodes' features and assigning higher weights to important nodes. Self-supervised training was performed using the biological network alignment metrics EC (Edge Correctness), LCCS (Largest Common Connected Component), S3 (Symmetric substructure score), and GOSim (GO annotation similarity) as training labels. The proposed scheme has been tested across all available biological networks for four different species, showing that it performs competitively when compared to the latest methods.

Our contributions are as follows:

1. We introduce GBNSS, a novel method utilizing machine learning for biological network similarity search. Post training, the model significantly reduces query time to within 1 s.
2. GBNSS integrates topological and biological features and employs graph neural networks to automatically learn feature representations, capturing hidden patterns and correlations of biological networks to enhance query accuracy.
3. Experimental results on four benchmark datasets reveal that GBNSS surpasses existing methods in both computational efficiency and search accuracy. Additionally, case studies highlight GBNSS's effectiveness in identifying similar networks within real-world biological datasets.

2. Materials and Methods

2.1. Graph Representations

Let $D = \{P_1, P_2, P_3, \ldots, P_n\}$ be a repository of biological networks where n denotes the total number of networks within the database. Additionally, $S = \{Q_1, Q_2, Q_3, \ldots, Q_m\}$ is a collection of biological query networks with m being the count of networks in the set. Consider $P = (V, E)$ a directed graph that embodies a biological network. In this context, the node set V signifies molecules such as proteins, genes, and DNA, while the edge set E represents interactions among these molecules. Similarly, let $Q = (V', E')$ represent a

query network defined by its node set V' and the edge set E'. The aim of biological network similarity searches is to identify and return all networks from the database that exhibit topological and biological analogies to the query biological network Q.

2.2. Topological Features

In biological networks' biological feature extraction process, a hierarchical feature extraction protocol is adopted to characterize both node and global network attributes. The following provides a formalized description of the methodology:

- Node Feature Extraction: Each node in a biological network database is assigned a feature vector, with the vector's dimensionality equal to the total number of unique networks in the biological network datasets. The vector represents the presence or absence of the node in each network, with 1 indicating presence and 0 indicating absence.
- Graph Feature Extraction: Node-specific feature vectors are combined to form a composite representation of the graph by concatenation or pooling mechanisms such as summation or average calculation. If a biological network has n nodes, its resulting graph feature vector will have a dimensionality of $n \times m$ upon concatenation. The relationship between the graph and the features is illustrated in Figure 1.

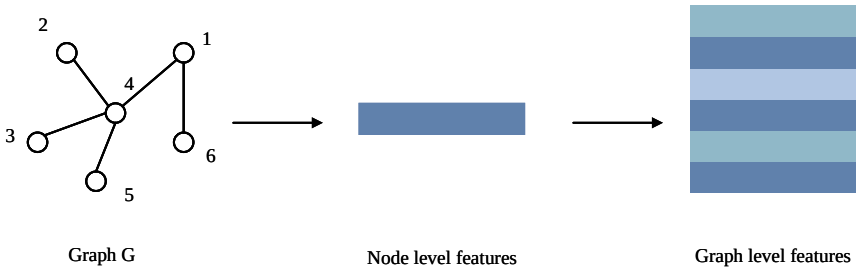


Figure 1. Node-level features and graph-level features.

2.3. Biological Features

GO annotation is the process of assigning functional terms from the Gene Ontology (GO) database to genes or gene products such as proteins. Gene Ontology is a large-scale, structured, precisely defined, and dynamically updated bioinformatics resource that provides a set of controlled vocabulary to describe gene functions across different species.

Gene Ontology annotation (GO annotation) provides a unified and standardized language for describing the attributes of genes and gene products [31], enabling cross-species data understanding and comparison, and reflecting the functional roles of molecules within cells. Using GO annotations to construct biochemical features, our method uses a three-dimensional vector to represent a molecule's biochemical characteristics. Each dimension represents a molecular function, cellular component, and biological process, and its value is equal to the number of associated GO annotations. For example, hsa:3081 has a feature vector of [2, 3, 2].

The biochemical features provide insights into the functional roles of nodes within biological networks, complementing the topological features of each node in capturing the underlying biological mechanisms.

2.4. The Label of the GBNSS Model

Network alignment seeks to identify analogous elements between two networks, such as biological pathways. Topological measurement indices evaluate the quality of network alignment. The definitions of three widely used network alignment metrics, EC, LCCS, and S3, are detailed as follows:

- Edge Correctness (EC): *EC* is a quantitative measure that assesses the degree of edge alignment between two networks based on a node mapping [32], and provides insight

into the structural similarity and accuracy of mapping between the two networks. The EC is defined as

$$EC(P,Q) = \frac{|\{(u,v) \in E_1 \wedge (f(u),f(v)) \in E_2\}|}{min(|E_1|,|E_2|)}, \quad (1)$$

where E_1 and E_2 are the set of edges in network P and Q, respectively. u and v are nodes in network P; $f(u)$ and $f(v)$ are the corresponding nodes in network Q that u and v are mapped to. $min(|E_1|,|E_2|)$ takes the minimum of the number of edges in P and Q. The EC value ranges from 0 to 1, where 1 signifies complete overlap and 0 indicates no overlap at all. A higher EC value generally indicates a greater overlap of edges between the two networks after mapping, which typically suggests a higher degree of structural similarity between them.

- Largest Common Connected Subgraph (LCCS): LCCS is the largest subgraph that is shared between two or more graphs, where the subgraph is both connected and identical in terms of its nodes and edges across the graphs. It is calculated by counting the edges in the largest connected subgraph between two networks [33]. A larger and denser subgraph reflects the network's topological importance. LCCS is used to avoid problems with high Edge Count (EC) scores that may not necessarily reflect a connected substructure. In the case of two different alignments with the same EC score, the alignment with a higher LCCS is better.
- Symmetric substructure score (S3): *S3* effectively measures how well edges are conserved in a global network alignment [17] and has been shown to be a superior metric for comparing sparser network regions to denser ones, unlike EC. The S3 is defined as follows:

$$S3(P,Q) = \frac{|\{(u,v) \in E_1 \wedge (f(u),f(v)) \in E_2\}|}{|E_1| + |E(Q[f(V_1)])| - |\{(u,v) \in E_1 \wedge (f(u),f(v)) \in E_2\}|}, \quad (2)$$

where $|E_1|$ is the total number of edges in the network P; V_1 is the set of nodes in the network P. $E(Q[f(V_1)])$ denotes the edges in the graph of network Q induced by $f(V_1)$, which are the edges between the images of the nodes in V_1 under the mapping f. A higher *S3* indicates a greater degree of symmetry and structural alignment between the two substructures. Conversely, a lower score suggests less symmetry and structural similarity.

- GOSim (GO annotation similarity): *GOSim* is a similarity metric that is defined to compare the similarity of two biological networks above their biochemical performance. Comparing the GO annotation of two biological networks reflects the difference in biochemical performance between two biological networks, which is defined as follows:

$$GOSim(P,Q) = \frac{1}{m}\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n} GOSemSim(p_i,q_j), \quad (3)$$

where m denotes the number of molecules contained in the biological network P; n denotes the number of molecules contained in the biological network Q. p_i denotes the i-th molecule in the biological network P, and q_j denotes the j-th molecule in biological network Q. GOSemSim [34] is a tool for measuring the similarity between two biological molecules based on their associated Gene Ontology (GO) terms. It calculates a numerical score that reflects the functional similarity between the molecules by evaluating how closely related their GO annotations are.

2.5. Graph Convolutional Networks (GCNs)

Both topological and biological features require node embedding computation to generate graph-level embeddings, which are obtained by aggregating node-level embeddings. Among the various existing node embedding algorithms, we choose to use graph convolutional networks (GCNs) because they are invariant to graph representation, provided

that the initialization is carefully designed [35]. GCNs are also inductive, allowing us to compute node embeddings for any unseen graph by applying the GCN operation. GCNs have become one of the most popular models for node embeddings and are part of the family of neighbor aggregation-based methods. Their key operation graph convolution is defined as [36]

$$conv(u_n) = f_l(\sum_{m \in \mathcal{N}(n)} \frac{1}{\sqrt{d_n d_m}} u_m W_l^l + b_l^l), \quad (4)$$

where $\mathcal{N}(n)$ is the set of first-order neighbors of node n, which includes all nodes that are directly connected to node n. $\frac{1}{\sqrt{d_n d_m}}$ is a normalization factor, d_n is the degree of node n, and d_m is the degree of node m. u_m is the feature vector of a neighbor node m and W_1^l is the weight matrix used in the l-th GCN; it transforms the feature vectors of the neighbors. b_l^l is the bias added during the transformation in the l-th layer, and $f_l(\cdot)$ is an activation function applied element-wise to the result of the linear transformation and aggregation such as $ReLU(x) = max(0, x)$.

2.6. The Architecture of GBNSS

The GBNSS model is a graph-based neural network (GNN) designed for biological network similarity search. It integrates both topological and biological information using graph convolutional networks (GCNs), an attention mechanism, and a tensor network module to effectively capture species-specific biological network features. The architecture is structured as follows and the architecture of GBNN is shown in Figure 2:

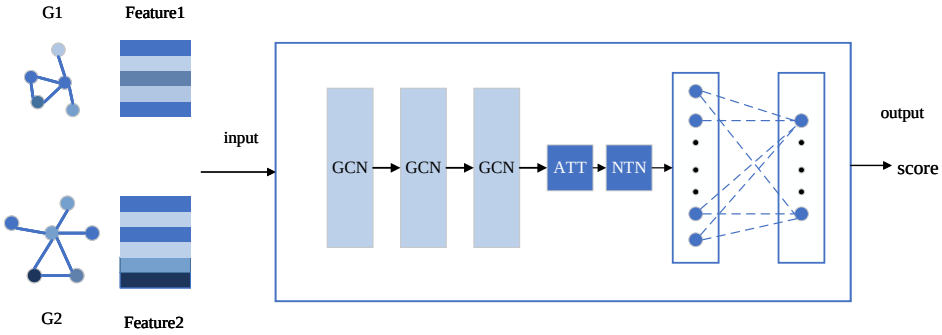


Figure 2. The architecture of a graph neural network in GBNSS.

- Input Layers: The model operates on pairs of biological networks, where the edge indices and feature matrices of both networks are provided as inputs. Each node is represented by a feature vector of a size derived from the topological features and biological features.
- Graph Convolution Layers: The model consists of three layers of graph convolutions, each responsible for extracting increasingly abstract representations from the input networks. These three GCN layers ensure that the model captures both local information (immediate neighbors) and global patterns (up to three-hop neighbors), which are essential for biological network similarity search.
- Attention Mechanism: The abstracted node features from the GCN layers are passed through an attention mechanism, which helps in identifying the most important nodes in each graph. This mechanism generates a set of pooled features for each graph, effectively compressing the node features into a global graph representation.
- Neural Tensor Network: The pooled features from both graphs are passed through a tensor network layer, which calculates a pairwise similarity score between the two sets of graph representations. The layer uses 16 tensor neurons, which allows it to combine the two feature sets and capture complex interactions between the two graphs.
- Fully Connected Layer: The output from the neural tensor network is passed through a fully connected layer. This layer reduces the feature dimension and prepares the features for the final scoring layer.

- Output layers: The final layer is a single neuron (output = 1), which outputs a similarity score, representing the predicted similarity between the two biological networks.

2.7. The Model of GBNSS

GBNSS leverages GNNs to learn effective representations of biological networks, integrating both topological and biological features. In particular, GBNSS consists of three main components: feature generation, model training, and similarity query. In the feature generation phase, both topological and biological features are extracted from each node in the biological networks. The model training phase employs a GNN architecture comprising multiple graph convolution layers and a graph attention layer, which is trained to predict the similarity scores between pairs of biological networks. Finally, in the similarity query phase, the trained model is utilized to compute the similarity scores between a query network and a database of biological networks, facilitating efficient and accurate similarity searches. Figure 3 visualizes the scheme of GBNSS.

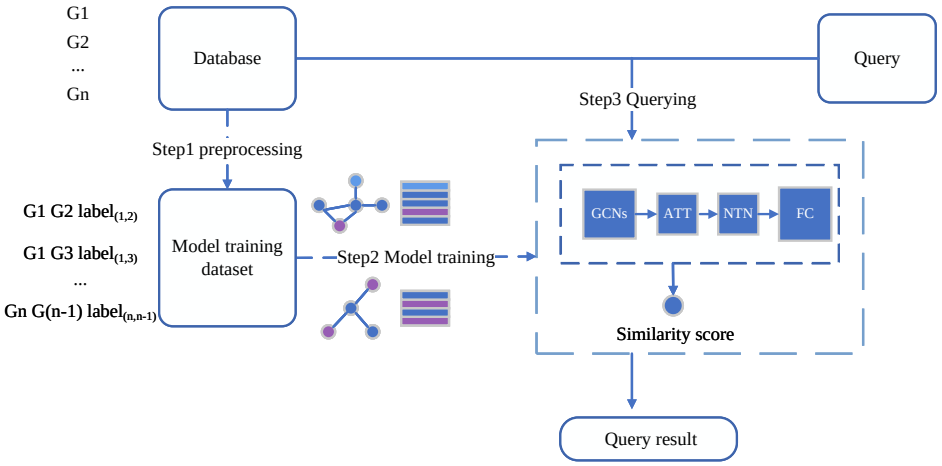


Figure 3. The scheme of GBNSS. The dotted line in Step 1 indicates the data preprocessing, the dotted line in Step 2 indicates the model training, and the solid line in Step 3 indicates the querying.

The GBNSS framework encompasses three pivotal stages—feature generation, model training, and similarity query—with each contributing to the overarching goal of facilitating efficient and accurate similarity searches among biological networks.

- Feature Generation Stage: At the feature generation stage, both topological and biochemical features are extracted from the biological networks. Topological features, inspired by the bag-of-words model, capture the presence or absence of nodes across a set of biological networks. Meanwhile, biochemical features are derived from Gene Ontology annotations, categorizing nodes based on molecular function, cellular component, and biological process. Together, these features provide comprehensive representations of the structural and functional characteristics of nodes in biological networks.
- Model Training Stage: The model training stage harnesses the power of graph neural networks (GNNs) to learn effective representations of biological networks. GBNSS employs a GNN architecture comprising multiple graph convolutional layers and a graph attention layer. During training, the model is optimized to predict similarity scores between pairs of biological networks, using metrics EC, LCCS, and S3 for topological similarity and GOSim for biochemical similarity. By minimizing the mean squared error (MSE) loss, the model learns to capture intricate relationships and nuances within and between biological networks. Specifically, when predicting the topological similarity between two networks, we use topological features as input because they directly capture the structural aspects relevant to the task. Similarly, for biochemical feature similarity prediction, using biological features as input aligns

better with the nature of the target. The pseudo-code of GBNSS's model training is presented in Algorithm 1.

- Similarity Query Stage: In the final stage, the trained model is deployed for similarity query tasks. Given a query biological network and a database biological network, the model GBNSS computes their similarity score. These scores are then used to rank the database networks in descending order of similarity, yielding the final similarity query results. By leveraging the model's learned representations and predictive capabilities, GBNSS enables efficient and accurate identification of biologically relevant network relationships. Four distinct models are provided after model training. Three models focus on topological similarity, using EC, LCCS, and S3 as labels, while the fourth model targets biochemical similarity with GOSim as the label. During inference, users can choose the model that best suits their specific needs and provide the relevant input data accordingly.

Algorithm 1 Model training Procedure of GBNSS.

Input: $D = \{P_1, P_2, \ldots, P_n\}$:database networks
$L_{DB} = \{A_{P_1}, A_{P_2}, \ldots, A_{P_n}\}$:adjacency matrix of database networks
TOF: the topological feature of biological networks
BOF: the biological feature of biological networks

for $P_i, P_j \in D$ **do**
  Compute EC, LCCS, S3, GOSim label $EC_{(P_i,P_j)}$, $LCCS_{(P_i,P_j)}$, $S3_{(P_i,P_j)}$, $GOSim_{(P_i,P_j)}$ between P_i and P_j
end for
if metric == EC **then**
  label = $EC_{(Q,DB)}$, feature = TOF
  Training the model Θ on $(A_{P_i}, TOF_{P_i}, A_{P_j}, TOF_{P_j}, EC_{(P_i,P_j)})$
else if metric == LCCS **then**
  label = $LCCS_{(Q,DB)}$, feature = TOF
  Training the model Θ on $(A_{P_i}, TOF_{P_i}, A_{P_j}, TOF_{P_j}, LCCS_{(P_i,P_j)})$
else if metric == S3 **then**
  label = $S3_{(Q,DB)}$, feature = TOF
  Training the model Θ on $(A_{P_i}, TOF_{P_i}, A_{P_j}, TOF_{P_j}, S3_{(P_i,P_j)})$
else if metric == GOSim **then**
  label = $GOSim_{(Q,DB)}$, feature = BOF
  Training the model Θ on $(A_{P_i}, BOF_{P_i}, A_{P_j}, BOF_{P_j}, GOSim_{(P_i,P_j)})$
end if

Output: A GBNSS model Θ

3. Results

3.1. Experiment Setup

- Experimental platform and computational resources: The experiments were conducted on a Windows personal computer using an Intel® Core™ i5-10210U processor (Intel, Santa Clara, CA, USA) and 8 GB of memory.
- Compared methods: We compare the experimental results of GBNSS with NSSRF [28] and PathEmb [30] to highlight its effectiveness. These methods were chosen for comparison due to their robustness, efficiency, and widespread use in biological network similarity searches, making them strong benchmarks for evaluating the performance of GBNSS. NSSRF, PathEmb, and GBNSS use the global similarity score to work out the similarity of biological networks. NSSRF2, NSSRF3, and NSSRF4 are different versions of the NSSRF method, characterized by the size of the subgraph signatures they use. Specifically, NSSRF2 uses a two-node signature, NSSRF3 uses a three-node signature, and NSSRF4 uses a four-node signature.

- Experimental Datasets: GBNSS was examined on all of the available databases (Bos Taurus, Homo Sapiens, Mus Musculus, and Saccharomyces cerevisiae), which were retrieved from the KEGG (Kyoto Encyclopedia of Genes and Genomes) PATHWAY database. [37]. The pathway was transformed into a network where nodes represent molecules and edges represent interactions between them. The data statistics for the four KEGG datasets are presented in Table 1. We measured GBNSS biological similarity performance using Gene Ontology terms. The GO terms were retrieved from UniProt [38].

Table 1. Data statistics for four species of KEGG PATHWAY datasets.

Species		Bos Taurus	Homo Sapiens	Mus Musculus	Saccharomyces c.
Database	Networks	40	270	260	70
	Nodes	4–28	4–470	3–431	3–128
	Edges	3–145	3–4480	3–3390	3–373
Query Set	Networks	5	44	41	11
	Nodes	6–51	6–422	6–478	6–62
	Edges	5–259	4–1984	4–2057	4–110

3.2. Data Processing

We processed four datasets, and each dataset was randomly split into query and database networks. We calculated the similarity between every pair of database networks using EC, LCCS, and S3 metrics via NETAL and HubAlign, selecting the maximum values as the final similarity measures. These computed values served as self-supervised labels. The details of the dataset information are listed in Table 2.

Table 2. Statistics of model datasets.

Dataset	Graphs	Pairs	Training Set	Testing Set
Bos Taurus	40	800	560	240
Homo Sapiens	270	36,450	25,515	10,935
Mus Musculus	260	33,800	23,660	10,140
Saccharomyces c.	70	2450	1715	735

The datasets were split into training and test sets using a 7:3 ratio, with 70% of the data allocated for training and 30% for testing. To ensure the robustness of our model, we also applied five-fold cross-validation, which helps validate the model's performance across different data splits. In GBNSS, separate GNN models are trained for each species to capture species-specific biological features. This is essential, as biological networks differ significantly across species, not only in their topological structures but also in the biological processes they represent. We applied the same training procedure for each species, including Homo Sapiens, Bos Taurus, Mus Musculus, and Saccharomyces cerevisiae. This approach ensures that the models are trained on biologically relevant data, capturing both the unique topological patterns and biological characteristics specific to each species.

3.3. Paremeter Setting

This section provides a detailed overview of the parameter settings used in the GBNSS model. These parameters were selected through empirical tuning to ensure optimal performance in predicting biological network similarity. The GBNSS architecture incorporates multiple layers, each designed to capture distinct aspects of the network structure and interactions.

The model begins with three graph convolutional network (GCN) layers, each progressively capturing deeper neighborhood information. These layers enable the model to learn both local and global structural features from the biological networks. The output dimensions for the three GCN layers are 128, 64, and 32, respectively.

Following the GCN layers, we incorporate a neural tensor network (NTN) to compute similarity scores between pairs of biological networks. This layer enhances the model's ability to identify meaningful relationships between networks. The NTN layer has an output dimension of 16.

To refine the representations and generate the final similarity score, the model includes two fully connected (FC) layers. These layers convert the high-dimensional outputs from previous layers into a single scalar similarity score. The output dimensions for the first and second FC layers are 16 and 1, respectively.

To prevent overfitting, we applied dropout with a rate of 0.5 between layers. This technique deactivates 50% of the neurons at random during training, which helps ensure the model does not rely too heavily on any particular neurons. As a result, the model is better able to generalize to new data. In addition to dropout, we implemented five-fold cross-validation to further safeguard against overfitting. This method splits the data into five parts, using four for training and reserving one for validation in each iteration. By repeating this process five times, we ensure that every subset is used for validation once. This approach offers a more comprehensive evaluation of the model's performance and reduces the risk of overfitting to a single subset.

3.4. Topological Similarity Evaluation

Firstly, we assess the performance of GBNSS for biological network topological similarity in Bos Taurus, Homo Sapiens, Mus Musculus, and Saccharomyces cerevisiae four databases. The performance is measured by the Simpson overlap coefficient value of the top K similar biological networks in sets sim_NA and sim_M, which are defined as follows:

$$Simpson = \frac{|sim_{NA} \cap sim_M|}{min(|sim_{NA}|, |sim_M|)}, \quad (5)$$

where sim_{NA} is the top K similar network returned by ranking the network alignment metrics EC, LCCS, and S3. sim_M is the top K similar networks from the predicted set provided by the biological network similarity search method. A higher Simpson overlap coefficient value indicates better performance of the method. Figure 4 illustrates the topological performance measured by the Simpson overlap coefficient value on the Bos Taurus and Saccharomyces cerevisiae datasets. In most cases, GBNSS outperforms NSSRF and PathEmb in terms of Simpson overlap for these two datasets.

GBNSS consistently shows higher Simpson overlaps than NSSRF and PathEmb across all cases for the Homo Sapiens and Mus Musculus datasets. Notably, GBNSS achieves an overlap of over 0.5 within the top 10% similarity range for these datasets, highlighting its effectiveness with larger datasets. Figure 5 displays the topological performance measured by Simpson overlap for Homo Sapiens and Mus Musculus.

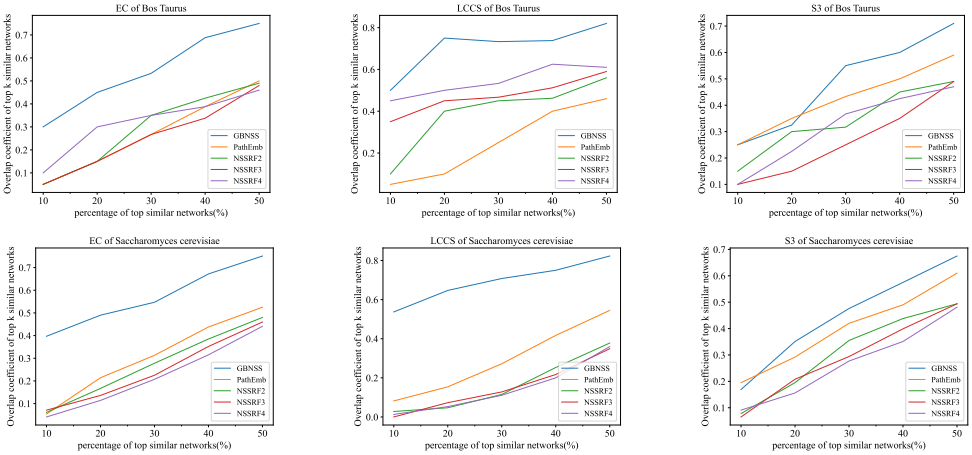


Figure 4. The topological performance is examined by Simpson overlap on the Bos Taurus and Saccharomyces cerevisiae datasets.

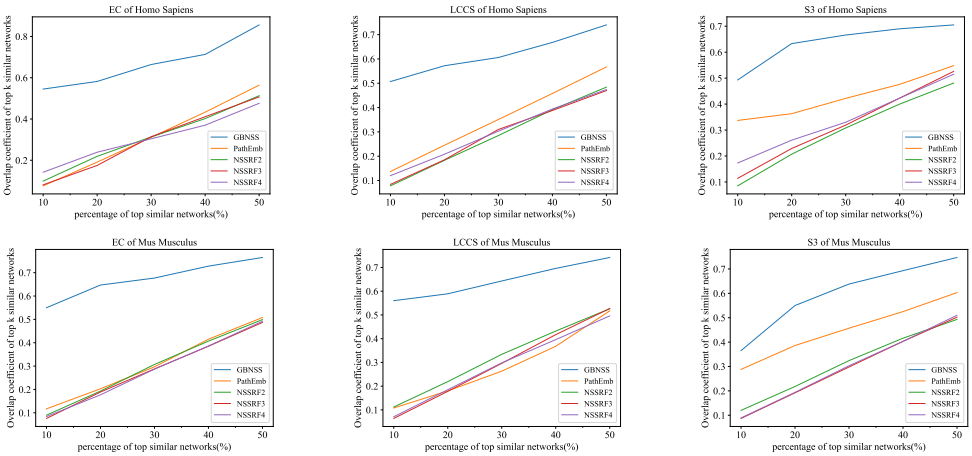


Figure 5. The topological performance is examined by Simpson overlap on the Homo Sapiens and Mus Musculus datasets.

3.5. Biological Similarity Evaluation

To evaluate the biological similarity performance of GBNSS, we use two existing measures: the Simpson overlap of Gene Ontology and functional coherence (FC).

As depicted in Figure 6, the Simpson overlap coefficient was used to evaluate the performance of GOSim across four different datasets. Comparative analysis revealed that GBNSS surpassed NSSRF2, NSSRF3, NSSRF4, and PathEmb on all four datasets. Strikingly, within the top decile of ranked similarities on the Mus Musculus dataset, the Simpson overlap coefficient achieved by GBNSS was marginally less than 0.9, demonstrating a high level of concurrence between the predicted functional similarities.

Functional coherence [39] evaluates biological relevance by analyzing the semantic relationships of GO terms associated with the relevant protein-coding genes. The similarity between two biological networks is computed as

$$FC(u,v) = \frac{|S_u \cap S_v|}{|S_u \cup S_v|}, \tag{6}$$

where S_u and S_v are the GO term set of network u and network v. The higher the functional coherence, the more similar the two networks are. The corresponding biological performance is shown in Figure 7. The FC of GBNSS on Homo Sapiens, Mus Musculus, and Saccharomyces cerevisiae are larger than those of NSSRF and PathEmb. As a result,

GBNSS consistently outperforms the other methods, showing that it can produce more biologically relevant outcomes in the context of biological network similarity searches.

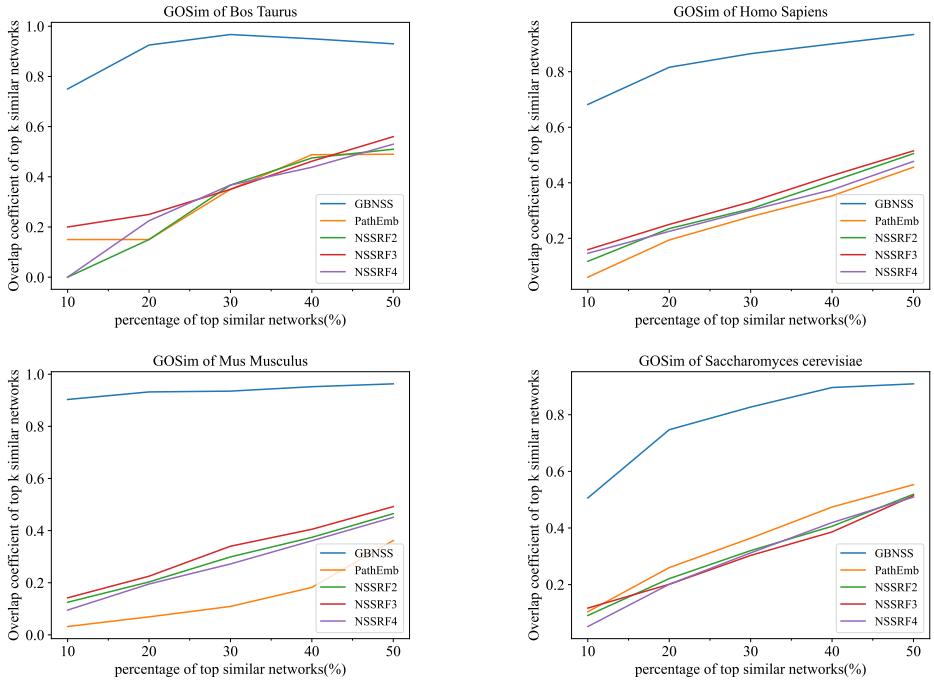


Figure 6. Simpson overlap coefficient performance for GOSim on four species datasets.

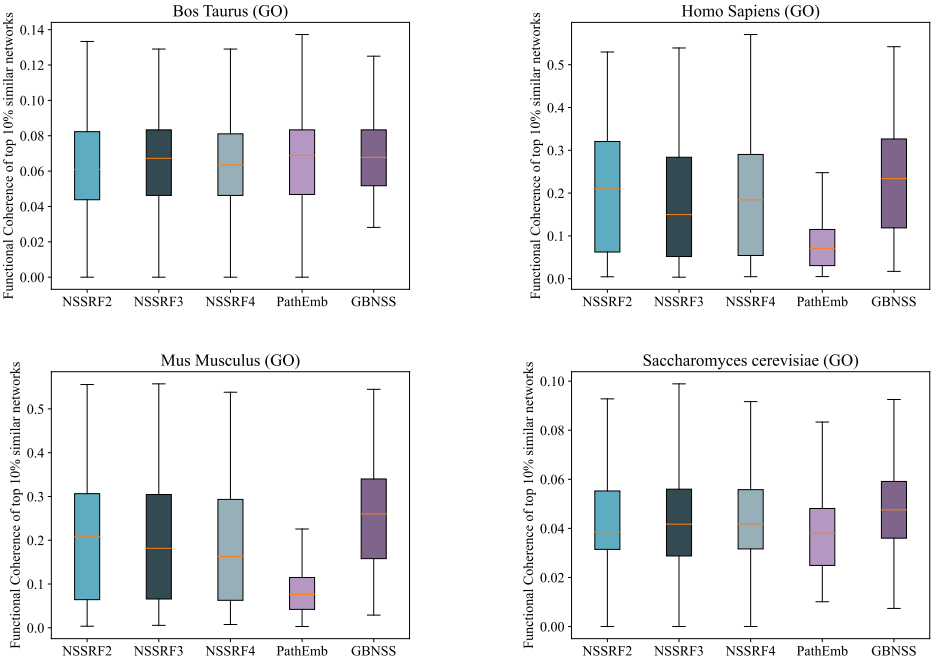


Figure 7. Functional coherence of top 10% similar biological networks on four species datasets.

3.6. Computational Cost

The GBNSS algorithm comprises three distinct phases: feature construction, offline model training, and similarity search. To assess the computational efficiency associated with generating similar biological networks, we compiled the necessary computational costs. It is important to note that GBNSS, NSSRF, and PathEmb utilize varying numbers of training datasets; therefore, we confined our comparison to the query times of these methods. The similarity search querying time for NSSRF2, NSSRF3, NSSRF4, PathEmb,

and GBNSS is shown in Table 3. Notably, GBNSS exhibits a reduced time during similarity queries, and its query time remains unaffected by the size of the database.

Table 3. The similarity search querying time.

Dataset	PathEmb	NSSRF2	NSSRF3	NSSRF4	GBNSS
Bos Taurus	2.4	2.1	2.5	2.3	0.14
Homo Sapiens	39.6	3.4	10	9.2	0.2
Mus Musculus	36.5	2	5.5	6.1	0.2
Saccharomyces c.	11.3	2.2	3	2.9	0.16

3.7. Ablation Experiment

We conducted two ablation experiments to assess how node features and the attention mechanism contribute to the performance of the GBNSS model. Specifically, we evaluated the effect of topological features and biochemical features on two similarity search tasks: topological similarity and biochemical similarity. These experiments were designed to determine how each component influences the model's predictions, with the Wilcoxon signed-rank test used to assess the statistical significance of the results.

To ensure a focused analysis, all experiments were conducted on the Homo Sapiens dataset, a widely used and comprehensive biological network. This choice minimizes the influence of dataset characteristics, allowing us to isolate the effects of the model components.

The first experiment analyzed the impact of topological node features and the attention mechanism on the model's ability to predict topological similarity. The four model configurations and their results are summarized in Table 4.

Table 4. Attention mechanism and features used in different GBNSS models about topological similarity searches in Bos Taurus species.

Models	Features	Attention Mechanism	Result
$GBNSS_e1$	None	NO	0.191
$GBNSS_e2$	None	Yes	0.239
$GBNSS_e3$	TOF	NO	0.22
$GBNSS$	TOF	YES	0.304

The full model GBNSS achieves the highest topological similarity score of 0.304, demonstrating that the combined use of topological features and an attention mechanism enhances performance. In contrast, $GBNSS_e1$performs the worst, with a score of 0.191, indicating that the absence of both components negatively impacts the model's predictions.

The intermediate scores for $GBNSS_e2$ (0.239) and $GBNSS_e3$ (0.220) suggest that topological features and attention mechanisms each independently contribute to performance improvements, but the greatest accuracy is achieved when both are used together.

We used the Wilcoxon signed-rank test to compare the full model with each ablated version. The results show that both topological features and the attention mechanism enhance performance, with the attention mechanism providing a statistically significant improvement ($p < 0.05$).

The second experiment examined how biochemical node features and the attention mechanism affect the model's ability to predict biochemical similarity. The four model configurations and their results are presented in Table 5.

The full model GBNSS achieves the highest biochemical similarity score of 0.75, indicating that both biochemical features and attention mechanisms are essential for optimal performance. The $GBNSS_g1$ configuration produces the lowest score (0.43), showing that the absence of both components significantly reduces prediction accuracy.

The intermediate scores of $GBNSS_g2$ (0.56) and $GBNSS_g3$ (0.51) suggest that both biochemical features and attention mechanisms independently enhance performance. However, the combination of both components yields the best results.

We used the Wilcoxon signed-rank test to compare the full model with each ablated version. The results confirm that both biological features and the attention mechanism contribute meaningfully to performance improvements, with each showing a statistically significant effect ($p < 0.05$).

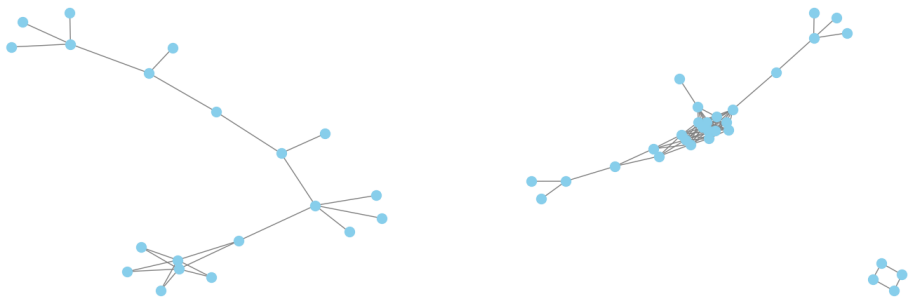
Table 5. Attention mechanism and features used in different GBNSS models about metric GOSim about topological similarity searches in Bos Taurus species.

Models	Features	Attention Mechanism	Result
$GBNSS_g1$	None	NO	0.43
$GBNSS_g2$	None	Yes	0.56
$GBNSS_g3$	BOF	NO	0.51
$GBNSS$	BOF	YES	0.75

3.8. Case Study

A case study aims to capture the unique aspects of a single case, providing insight into its behavior in significant situations. The case studies presented here offer detailed information on each method's performance in specific contexts. To illustrate the competitive performance of GBNSS, we show two case studies with concrete metric values EC, LCCS, S3, and GOSim.

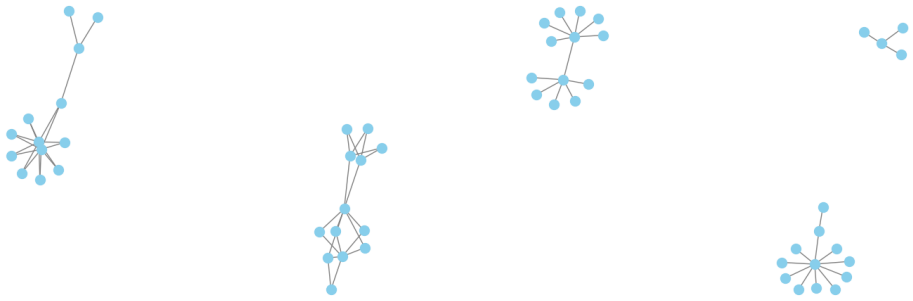
The first case study was conducted using the Saccharomyces cerevisiae dataset, which includes 81 biological networks. The pathway sce04130 (SNARE interactions in vesicular transport) was randomly picked up from the biological query networks and used to search against the Saccharomyces cerevisiae dataset. Another pathway, sce00520 (Amino sugar and nucleotide sugar metabolism), was randomly chosen from the dataset and ranked within the top 10% of pathways similar to sce04130. The EC, LCCS, S3, and GOSim scores between sce04130 and sce00520 are 0.78, 10, 0.4, and 28.05, respectively. When GBNSS was applied to search the Saccharomyces cerevisiae dataset, it successfully identified sce00520 within its top 10% of similar pathways. In contrast, NSSRF and PathEmb could only detect sce00520 within their top 50% of similar pathways. The visual comparison of pathways sce04130 and sce00520 is shown in Figure 8.



(**a**) (**b**)

Figure 8. Visualization of (**a**) sce04130 and (**b**) sce00520.

The second case study was performed using the Homo Sapiens dataset, which includes 314 biological networks. The pathway hsa05410 (Hypertrophic cardiomyopathy) was randomly picked up from the biological query networks to search against the Homo Sapiens dataset. Another pathway, hsa04976 (Bile secretion), was randomly chosen from the dataset and ranked within the top 10% of pathways similar to hsa05410. The EC, LCCS, S3, and GOSim scores between hsa05410 and hsa04976 are 0.6, 10, 0.3, and 63.46, respectively. When GBNSS was applied to search the Homo Sapiens dataset, it successfully identified hsa04976 within its top 10% of similar pathways, while PathEmb and NSSRF identified hsa04976 within their top 30% and top 90% of similar pathways, respectively. The visual comparison of pathways hsa05410 and hsa04976 is shown in Figure 9.



(**a**) (**b**)

Figure 9. Visualization of (**a**) hsa05410 and (**b**) hsa04976.

3.9. Limitations and Future Works

Although GBNSS performs well in biological network similarity searches, several areas require improvement.

Firstly, the use of one-hot encoding for topological features poses scalability challenges. As the dataset size grows, the dimensionality of these vectors increases, leading to higher memory usage and lower computational efficiency. To address this, we plan to explore more efficient feature representations. In future work, we aim to incorporate intrinsic topological features such as the number of neighbors, in-degree, out-degree, and betweenness centrality to represent nodes more effectively.

Secondly, GBNSS performed well in predicting accuracy for topological similarity queries and biochemical similarity queries of biological networks. There are too many similarity metrics used by our model to determine the similarity of two biological networks, which leads to the complicated training process of the model. To simplify our model and methodology, we plan to design a metric that can represent the biochemical similarity and topological similarity of biological networks in our future work.

Thirdly, GBNSS focuses on completed biological networks to ensure the biological relevance and reliability of the predictions. We will carry out future studies on incomplete biological networks.

Finally, although GBNSS predicts similarity scores to assess network similarity, these predictions have not yet been experimentally validated. A thorough evaluation and analysis will be essential to confirm the reliability and robustness of the model. Future work will involve comprehensive experiments to verify the accuracy and effectiveness of the predicted similarity scores.

4. Discussion

In this study, we introduced GBNSS, a novel approach for global biological network similarity search. The GBNSS method consists of three phases: the graph feature construction phase, the model training phase, and the similar network querying phase. GBNSS has been extensively evaluated on four species, considering both topological and biological aspects. Across multiple biological network databases, GBNSS consistently outperforms existing methods such as NSSRF and PathEmb in terms of topological accuracy. Notably, GBNSS retrieves more relevant pathways at various biological levels compared to these methods, demonstrating its utility in accurately identifying similar networks.

Although GBNSS requires more model training time than NSSRF and PathEmb, it significantly reduces query time, making it more efficient for rapid searches. This trade-off between training time and query speed positions GBNSS as a suitable method for applications requiring fast similarity queries in large-scale biological databases.

Due to its incorporation of multiple metrics for comparing the topological similarity of two networks, GBNSS effectively predicts topological similarity. However, there are currently limited metrics for assessing biochemical functional similarity.

Beyond the current scope of biological network similarity search, GBNSS holds the potential for broader applications across several important fields. For example, in drug discovery, GBNSS could be instrumental in identifying similar biological networks related to disease pathways. In evolutionary biology, GBNSS can be applied by comparing biological networks across species, helping to identify evolutionary conserved pathways and functions.

Author Contributions: Methodology, Y.W.; Software, Y.W.; Validation, Y.W.; Formal analysis, Y.H.; Writing—original draft, Y.W.; Writing—review & editing, F.Z., C.H. and Y.H.; Supervision, Y.H.; Funding acquisition, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Guangxi Province (No. 2020GXNSFAA159074) and the National Natural Science Foundation of China (No. 61862006 and No. 62261003).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request. Data are available from: https://github.com/hyr0771/GBNSS, accessed on 27 October 2024.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alon, U. Biological networks: The tinkerer as an engineer. *Science* **2003**, *301*, 1866–1867. [CrossRef]
2. Nepusz, T.; Yu, H.; Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* **2012**, *9*, 471–472. [CrossRef]
3. Vaishnav, E.D.; de Boer, C.G.; Molinet, J.; Yassour, M.; Fan, L.; Adiconis, X.; Thompson, D.A.; Levin, J.Z.; Cubillos, F.A.; Regev, A. The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **2022**, *603*, 455–463. [CrossRef]
4. Muchowska, K.B.; Varma, S.J.; Moran, J. Nonenzymatic metabolic reactions and life's origins. *Chem. Rev.* **2020**, *120*, 7708–7744. [CrossRef]
5. Huang, Y.; Bin, Y.; Zeng, P.; Lan, W.; Zhong, C. NetPro: Neighborhood interaction-based drug repositioning via label propagation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 2159–2169. [CrossRef]
6. Huang, Y.; Wu, Z.; Lan, W.; Zhong, C. Predicting disease-associated N7-methylguanosine (m7G) sites via random walk on heterogeneous network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 3173–3181. [CrossRef]
7. Huang, Y.; Chen, F.; Sun, H.; Zhong, C. Exploring gene-patient association to identify personalized cancer driver genes by linear neighborhood propagation. *BMC Bioinform.* **2024**, *25*, 34. [CrossRef]
8. Huang, Y.; Zeng, P.; Zhong, C. Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning. *BMC Bioinform.* **2024**, *25*, 132. [CrossRef]
9. Lynch, M.; Walsh, B. *The Origins of Genome Architecture*; Sinauer Associates: Sunderland, MA, USA, 2007; Volume 98.
10. Glover, N.; Dessimoz, C.; Ebersberger, I.; Forslund, S.K.; Gabaldón, T.; Huerta-Cepas, J.; Martin, M.J.; Muffato, M.; Patricio, M.; Pereira, C.; et al. Advances and applications in the quest for orthologs. *Mol. Biol. Evol.* **2019**, *36*, 2157–2164. [CrossRef]
11. Panditrao, G.; Bhowmick, R.; Meena, C.; Sarkar, R.R. Emerging landscape of molecular interaction networks: Opportunities, challenges and prospects. *J. Biosci.* **2022**, *47*, 24. [CrossRef]
12. Tian, Y.; Yan, T.; Zhao, X.; Huang, K.; Zhou, X. A learned index for exact similarity search in metric spaces. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 7624–7638. [CrossRef]
13. Djeddi, W.E.; Hermi, K.; Ben Yahia, S.; Diallo, G. Advancing drug–target interaction prediction: A comprehensive graph-based approach integrating knowledge graph embedding and ProtBert pretraining. *BMC Bioinform.* **2023**, *24*, 488. [CrossRef] [PubMed]
14. Hopkins, A.L. Network pharmacology: The next paradigm in drug discovery. *Nat. Chem. Biol.* **2008**, *4*, 682–690. [CrossRef] [PubMed]
15. Yu, X.; Liu, Z.; Fang, Y.; Zhang, X. Learning to count isomorphisms with graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 4845–4853.
16. Hasan, M.M.; Kahveci, T. Indexing a protein-protein interaction network expedites network alignment. *BMC Bioinform.* **2015**, *16*, 1–17. [CrossRef]
17. Tian, Y.; Mceachin, R.C.; Santos, C.; States, D.J.; Patel, J.M. SAGA: A subgraph matching tool for biological graphs. *Bioinformatics* **2007**, *23*, 232–239. [CrossRef]
18. Dost, B.; Shlomi, T.; Gupta, N.; Ruppin, E.; Bafna, V.; Sharan, R. QNet: A tool for querying protein interaction networks. *J. Comput. Biol.* **2008**, *15*, 913–925. [CrossRef]
19. Bonnici, V.; Busato, F.; Micale, G.; Bombieri, N.; Pulvirenti, A.; Giugno, R. APPAGATO: An APproximate PArallel and stochastic GrAph querying TOol for biological networks. *Bioinformatics* **2016**, *32*, 2159–2166. [CrossRef]

20. Bombieri, N.; Bonnici, V.; Giugno, R. Parallel searching on biological networks. In Proceedings of the 2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP), Pavia, Italy, 13–15 February 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 307–314.
21. Sahraeian, S.M.E.; Yoon, B.J. RESQUE: Network reduction using semi-Markov random walk scores for efficient querying of biological networks. *Bioinformatics* **2012**, *28*, 2129–2136. [CrossRef]
22. Jeong, H.; Yoon, B.J. SEQUOIA: Significance Enhanced Network Querying Through Context-Sensitive Random Walk and Minimization of Network Conductance. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Seattle, WA, USA, 2–5 October 2016; pp. 535–536.
23. Woo, H.M.; Yoon, B.J. MONACO: Accurate biological network alignment through optimal neighborhood matching between focal nodes. *Bioinformatics* **2021**, *37*, 1401–1410. [CrossRef]
24. Jeong, H.; Qian, X.; Yoon, B.J. CUFID-query: Accurate network querying through random walk based network flow estimation. *BMC Bioinform.* **2017**, *18*, 133–146. [CrossRef]
25. Ye, J.; McGinnis, S.; Madden, T.L. BLAST: Improvements for better sequence analysis. *Nucleic Acids Res.* **2006**, *34*, W6–W9. [CrossRef] [PubMed]
26. Kelley, B.P.; Yuan, B.; Lewitter, F.; Sharan, R.; Stockwell, B.R.; Ideker, T. PathBLAST: A tool for alignment of protein interaction networks. *Nucleic Acids Res.* **2004**, *32*, W83–W88. [CrossRef] [PubMed]
27. Kalaev, M.; Smoot, M.; Ideker, T.; Sharan, R. NetworkBLAST: Comparative analysis of protein networks. *Bioinformatics* **2008**, *24*, 594–596. [CrossRef] [PubMed]
28. Zhang, J.; Kwong, S.; Jia, Y.; Wong, K.C. NSSRF: Global network similarity search with subgraph signatures and its applications. *Bioinformatics* **2017**, *33*, 1696–1702. [CrossRef] [PubMed]
29. Zhang, J.; Kwong, S.; Wong, K.C. Tobio: Global pathway similarity search based on topological and biological features. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *16*, 336–349. [CrossRef]
30. Zhang, J.; Kwong, S.; Liu, G.; Lin, Q.; Wong, K.C. PathEmb: Random walk based document embedding for global pathway similarity search. *IEEE J. Biomed. Health Inform.* **2018**, *23*, 1329–1335. [CrossRef]
31. Dimmer, E.C.; Huntley, R.P.; Alam-Faruque, Y.; Sawford, T.; O'Donovan, C.; Martin, M.J.; Bely, B.; Browne, P.; Mun Chan, W.; Eberhardt, R.; et al. The UniProt-GO annotation database in 2011. *Nucleic Acids Res.* **2012**, *40*, D565–D570. [CrossRef]
32. Kuchaiev, O.; Milenković, T.; Memišević, V.; Hayes, W.; Pržulj, N. Topological network alignment uncovers biological function and phylogeny. *J. R. Soc. Interface* **2010**, *7*, 1341–1354. [CrossRef]
33. Elmsallati, A.; Clark, C.; Kalita, J. Global alignment of protein-protein interaction networks: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *13*, 689–705. [CrossRef]
34. Yu, G.; Li, F.; Qin, Y.; Bo, X.; Wu, Y.; Wang, S. GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **2010**, *26*, 976–978. [CrossRef]
35. Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. Graph convolutional networks: A comprehensive review. *Comput. Soc. Networks* **2019**, *6*, 1–23. [CrossRef] [PubMed]
36. Bai, Y.; Ding, H.; Bian, S.; Chen, T.; Sun, Y.; Wang, W. Simgnn: A neural network approach to fast graph similarity computation. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, VIC, Australia, 11–15 February 2019; pp. 384–392.
37. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **2017**, *45*, D353–D361. [CrossRef] [PubMed]
38. Consortium, U. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515. [CrossRef] [PubMed]
39. Ma, C.Y.; Liao, C.S. A review of protein–protein interaction network alignment: From pathway comparison to global alignment. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 2647–2656. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.