Targeted Pathway Inference for Biological Knowledge Bases via Graph Learning and Explanation

Rikuto Kotoge[1], Ziwei Yang[1,2], Zheng Chen[1], Yushun Dong[3], Yasuko Matsubara[1], Jimeng Sun[4], Yasushi Sakurai[1]

[1]SANKEN, The University of Osaka, Japan
[2]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan
[3]Department of Computer Science, Florida State University, USA
[4]Department of Computer Science, University of Illinois Urbana-Champaign, USA
{rikuto88, chenz, yasuko, yasushi}@sanken.osaka-u.ac.jp,
yang.ziwei.37j@st.kyoto-u.ac.jp, yd24f@fsu.edu, jimeng@illinois.edu

Abstract

Retrieving targeted pathways in biological knowledge bases, particularly when incorporating wet-lab experimental data, remains a challenging task and often requires downstream analyses and specialized expertise. In this paper, we frame this challenge as a solvable graph learning and explaining task and propose a novel subgraph inference framework, EXPATH, that explicitly integrates experimental data to classify various graphs (bio-networks) in biological databases. The links (representing pathways) that contribute more to classification can be considered as targeted pathways. Our framework can seamlessly integrate biological foundation models to encode the experimental molecular data. We propose ML-oriented biological evaluations and a new metric. The experiments involving 301 bio-networks evaluations demonstrate that pathways inferred by EXPATH are biologically meaningful, achieving up to 4.5× higher Fidelity+ (necessity) and 14× lower Fidelity- (sufficiency) than explainer baselines, while preserving signaling chains up to 4× longer.

Introduction

Decades of research have revealed that systems, from cells to organisms, can be considered biological networks (Ideker and Krogan 2012). These networks have been compiled into public knowledge bases such as KEGG (Kanehisa and Sato 2020) and STRING (Szklarczyk et al. 2023), which document molecular (e.g., among genes or proteins) interactions and their roles in cellular functions. While knowledge bases are continuously updated, a primary concern remains: *they lack specificity for experimental data*. The main objective of biological knowledge bases is to cover all possible interactions in a system. These networks are general and static. In contrast, experimental studies focus on one specific condition or dataset, where only a subset of the network is actually relevant. Our objective is to identify which interactions are active, meaningful, or target-specific in the given data, as shown in Figure 1. In this paper, we propose to infer the bio-networks that capture targeted interactions from experimental data, thereby facilitating downstream analyses.

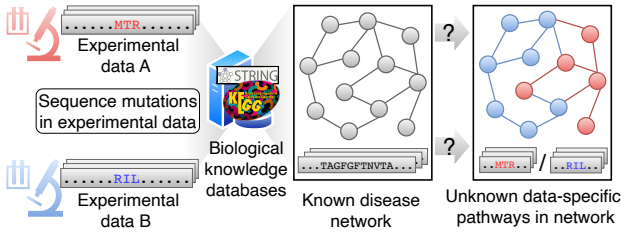

Figure 1: This example illustrates two experimental datasets with different mutations (red and blue) that are mapped onto the same disease network, yet fail to reveal the distinct interactions that account for their differences.

Many researchers have formulated this bio-network inference as a graph learning problem. In this setting, interactions in a bio-network are modeled as graph edges, and experimental data are embedded as node features. Various computational and machine learning methods have been proposed to infer meaningful targeted graph structures. Computational methods often rely on statistical node-centric metrics (Nacher and Akutsu 2016) to evaluate the importance of nodes. Edges connected to highly ranked nodes are considered more important. However, such objectives lack explicit inference of interactions and are computationally intractable for large bio-networks (Yang et al. 2025). Machine learning methods, particularly graph neural networks (GNNs), define explicit objectives such as link prediction or graph reconstruction, enabling direct inference of network structure (Ravindra et al. 2020). Importantly, experimental data influence the learning process via node feature aggregation, making the inferred interactions more specific to the dataset. However, existing methods are still in the early stages of exploration, and several key limitations remain unaddressed.

- **Implicit targeted interaction inference.** Their objectives aim to reconstruct the general graph structure accurately, including irrelevant interactions. Some works propose to gradually infer subgraph structure, weakening the influence of prior general bio-network information (Li et al. 2024). However, they still fail to explicitly identify the distinct interactions unique to different experimental data.

- **Lacking pathway modeling.** Existing works treat all interactions equally and independently, overlooking long-range dependencies in biological pathways. In reality, biological systems typically exhibit multi-step interactions, where one protein interaction triggers another, eventually leading to specific cellular outcomes.
- **Inadequate biologically plausible evaluation.** Existing methods typically require downstream biological analysis to qualitatively interpret the inferred interactions, which requires domain expertise. There is a lack of quantitative evaluation methods tailored for machine learning models.

To tackle the above challenges, we propose EXPATH, a deep learning framework for inferring targeted data-specific pathways in bio-networks, with the following contributions:

- **Graph explanation formulation for explicit interaction inference.** we formulate bio-network inference as a subgraph learning and explanation task, and hence propose a graph-based model equipped with GNNExplainer. Subgraphs, contributing most significantly to the learning objective, are explicitly identified as targeted interactions.
- **Pathway-level encoding and explaining.** To ensure these subgraphs capture high-order pathways, technically, we propose two novel models: PATHMAMBA, a hybrid learning model, combines GNNs with state-space sequence modeling (Mamba) to learn both local interactions and global pathway-level dependencies; PATHEXPLAINER identifies objective-critical pathways by learning novel pathway masks. We also provide a theoretical analysis of EXPATH's expressiveness and show that identified pathways capture higher-order structural patterns.
- **A novel ML-oriented biological evaluation.** We propose an evaluation workflow that directly incorporates model-derived subgraph importance scores to quantitatively assess their biological relevance.

EXPATH can *seamlessly integrate biological foundation models*, and in this work, we use the large protein language model, ESM-2 (Lin et al. 2023), as a case encoder. We collect all available human pathway networks from KEGG (Kanehisa et al. 2024), resulting in 301 bio-network, using amino acid (AA) sequences as reference experimental data. Extensive experiments demonstrate that the pathways inferred by EXPATH are biologically meaningful, achieving up to 4.5× higher Fidelity+ (necessity) and 14× lower Fidelity- (sufficiency) than explainer baselines, while preserving signaling chains up to 4× longer.

Related Work

Existing methods can be grouped into statistical topology-driven and data-driven deep graph learning methods.

- Topology-driven Methods. They use statistical metrics on structural properties of graphs, such as node degrees (Karger 1994; Kashtan et al. 2004a), centrality (Haynes, Hedetniemi, and Slater 2013; Wang et al. 2014; Nacher and Akutsu 2016), betweenness, or PageRank scores (Iván and Grolmusz 2011; Maehara et al. 2014) to infer which substructures exhibit a more significant influence on the overall topology, thereby identifying more targeted interactions.

- Deep Graph Learning Methods. They incorporate experimental data during the learning process by embedding data as node representations. They train GNNs with suitable objectives, such as link prediction or graph reconstruction (Yue et al. 2020; Zhang et al. 2021; Muzio, O'Bray, and Borgwardt 2021), and the links that contribute most to these objectives can be considered the targeted interactions. For instance, the works of (Hamilton, Ying, and Leskovec 2017; Gligorijević et al. 2021; Cheng et al. 2021) have been validated to predict protein functions within protein-protein interaction (PPI) networks. Moreover, GNN models have been applied to incorporate RNA-Seq data, for tasks like predicting disease states and cell-cell relationships (Wang et al. 2021; Ravindra et al. 2020).

Limitations of Previous Work. The topology-driven methods focus only on graph edges. They cannot incorporate experimental data to infer biological networks. While GNN-based methods can generate targeted interactions in a data-driven manner, their objectives do not explicitly focus on inferring networks and are typically task-specific. In contrast, our method focuses on directly explaining graph representations of bio-networks under specific experimental data.

Problem Formulation

Definition 1 (Knowledge bio-networks). The bio-networks can be represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the vertices, each representing a molecule such as a gene or protein, and $\mathcal{E}$ is the set of edges, representing molecular interactions. Let $\mathbf{G} = \{\mathcal{G}^{(m)}\}_{m=1}^{M}$ denote a dataset comprising M bio-networks. Each $\mathcal{G}^{(m)}$ is associated with a label $y^{(m)} \in \mathbf{Y}$, indicating its primary biologically functional class such as metabolism or human diseases.

Definition 2 (Molecular experimental data). For each node $v \in \mathcal{V}$, we are given a condition-specific feature vector $\mathbf{x}_v \in \mathbb{R}^d$ derived from molecular experiments (e.g., amino-acid sequence embeddings, gene-expression counts, or protein abundances). Collecting all nodes yields the matrix $\mathbf{X}^{(m)} = \left[\mathbf{x}_v\right]_{v \in \mathcal{V}^{(m)}}$ for network $\mathcal{G}^{(m)}$.

Problem 1 (Bio-network inference). Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a curated graph and the node features $\mathbf{X} = \{\mathbf{x}_v\}_{v \in \mathcal{V}}$ obtained from condition-specific *experimental data* (e.g., amino-acid embeddings). Although $\mathcal{G}$ is static, the pair $(\mathcal{G}, \mathbf{X})$ constitutes a *data-specific graph* that reflects the molecular state of the same pathway under the given experiment. To this end, we formulate this problem as a *two-stage sub-graph learning and explaining* task.

- **Task-1: Graph representation learning and classification.** Learn a classifier $F(\mathcal{G}, \mathbf{X})$ that predicts the functional label $y \in \mathbf{Y}$ of an unseen data-specific graph.
- **Task-2: Targeted subgraph explanation.** Develop an explainer $E(\cdot)$ that identifies the smallest subgraph $\hat{\mathcal{G}} \subseteq \mathcal{G}$ such that $F(\hat{\mathcal{G}}, \mathbf{X})$ still outputs y.

Problem 2 (Pathway modeling). Many biological functions arise from *long, multi-step reaction pathways* that span several hops in $\mathcal{G}$. Capturing long-range dependencies is essential: (1) for functional prediction, as perturbation effects often propagate across distant nodes; and (2) for mechanistic

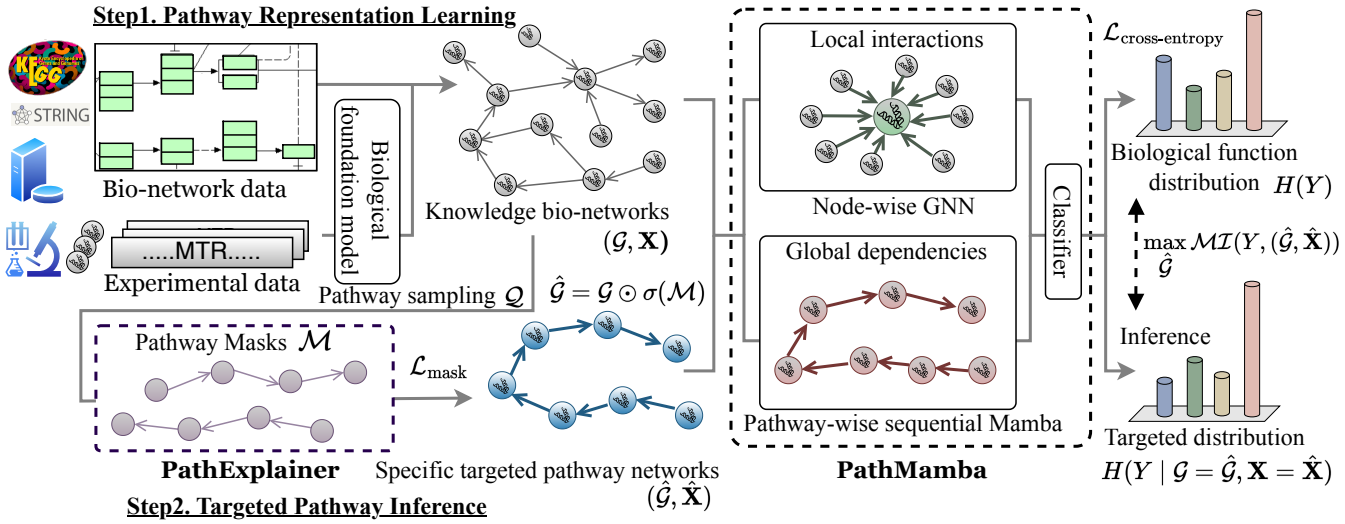

Figure 2: Overview of ExPath. Our method comprises two novel components. (1) PathMamba combining graph neural networks with state-space sequence modeling (Mamba) to capture both local interactions and global pathway-level dependencies for pathway information learning; and (2) PathExplainer identifies functionally critical nodes and edges through trainable pathway masks for targeted pathway inference.

insight into causal pathways beyond local interactions. Different experimental conditions on the *same* network, therefore yield distinct, data-specific subgraphs $\hat{\mathcal{G}}$, each revealing targeted pathway most responsible for the given data. Hence,

- **Expectation: Pathway-level encoding and explaining.** The classifier $F(\cdot)$ and explainer $E(\cdot)$ leverage both graph topology and node features. $F(\cdot)$ aims to capture long-range dependencies, yielding high and class-balanced accuracy. Also, $E(\cdot)$ extracts subgraphs that retain biologically meaningful information of long pathways.

Proposed Method

Framework

ExPath comprises two components: graph-based classification and post-hoc subgraph explanation, as shown in Figure 2. To tackle Task-1, PathMamba, a classifier combining GNNs with state-space sequence modeling, is to capture both local node-pair interactions and global pathway-level dependencies. To address Task-2, PathExplainer, a graph explainer with pathway-wise masking, aims to identify the most influential subgraphs. We explicitly integrate pathway information into both models to meet Expectation.

Notably, our ExPath is compatible with large biological foundation models for encoding experimental data. In this work, we leverage large protein language model encodings to investigate the mapping from amino acid (AA) sequences to corresponding pathway bio-networks. Learning from AA sequence data is challenging due to its inherent complexity. Even slight variations can lead to significant structural changes, potentially disrupting protein functionality within pathways. Several studies focus on feature extraction in AA sequences, like AlphaFold (Jumper et al. 2021).

Feedforward Process. We first encode experimental data into node attributes using ESM-2 (Lin et al. 2023). It is pretrained on over 60 million AA sequences with parameter scaling up to 15 billion. We then train PathMamba to learn pathway-level information and perform bio-network classification. Finally, we apply PathExplainer to selectively highlight the minimal subgraphs that drive the final prediction, offering interpretable insights into key pathways.

PathMamba: Pathway Representation Learning

PathMamba integrates the Graph Isomorphism Network (GIN) with a novel pathway-wise Mamba model. It leverages the strengths of both global selective modeling mechanisms and message-passing GNNs. Specifically, inspired by GPS (Rampášek et al. 2022), our model avoids early-stage information loss that could arise from using GNNs in the initial layers. We employ novel pathway-wise global aggregation in efficient combination with random pathway sampling and sequential Mamba (Gu and Dao 2024) modeling. At each layer, node and edge features are updated by aggregating the outputs of a pathway-wise Mamba as:

$$X^{l+1}, = \texttt{PathMamba}^l\left(X^l, A\right), \quad (1)$$

computed as

$$X_L^{l+1}, = \texttt{LocalGIN}^l\left(X^l, A\right), \quad (2)$$

$$X_G^{l+1}, = \texttt{GlobalMamba}^l\left(X^l, A\right), \quad (3)$$

$$X^{l+1}, = \texttt{MLP}^l\left(X_L^{l+1} + X_G^{l+1}\right), \quad (4)$$

where $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix of a graph with N nodes and E edges; $X^l \in \mathbb{R}^{N \times d}$ represents the d-dimensional node features at layer l; $\texttt{LocalGIN}^l$ is a GIN; $\texttt{GlobalMamba}^l$ is a global pathway-wise aggregation layer; and $\texttt{MLP}^l$ is a two-layer multilayer perceptron (MLP) used to combine local and global features.

Positional Encoding. To address a fundamental limitation of GNNs (Xu et al. 2019) or hybrid models (Rampášek et al. 2022) to distinguish nodes with identical local structures, The node embedding $\mathbf{h}_i \in \mathbb{R}^d$ and the positional encoding $\mathbf{p}_i \in \mathbb{R}^K$ are concatenated and passed through a linear layer to obtain the final representation: $\mathbf{x}_i = \text{Linear}([\mathbf{h}_i \| \mathbf{p}_i])$, where $[\mathbf{h}_i \| \mathbf{p}_i] \in \mathbb{R}^{d+K}$ denotes the concatenation.

Node-wise local aggregation. The GINs update Node features by aggregating information from their local neighbors. The GIN operation can be expressed as:

$$X_L^{l+1} = \text{ReLU}\left(W^l \cdot \left((1+\epsilon)X^l + \sum_{j \in \mathcal{N}(i)} X_j^l\right)\right), \quad (5)$$

where $\mathcal{N}(i)$ represents the set of neighbors of node i, W^l is the learnable weight matrix at layer l, and ϵ is a trainable parameter controlling the importance of self-loops.

Pathway-wise global aggregation. To capture long-term dependencies, we propose random pathway sampling and sequential pathway modeling in PATHMAMBA.

- *Random Pathway Sampling.* Formally, for each node v_i, we randomly sample a varied, single pathway with a maximum length of L. The sampling process is defined as:

$$\mathcal{Q} = \left\{\mathbf{q^i} \mid \mathbf{q^i} \sim \text{Pathway}(\mathbf{v_i}, \mathbf{L}), |\mathbf{q^i}| \leq \mathbf{L}\right\}_{i=1}^{N}, \quad (6)$$

where N is the number of nodes in the graph, and $\mathbf{q^i}$ represents the sampled pathway for node v_i. Each pathway $\mathbf{q^i}$ is a sequence of nodes $\{v_i, v_{i_1}, v_{i_2}, \ldots, v_{i_L}\}$, sampled according to a random walk process (Tonshoff et al. 2023). The sampling process Pathway(v_i, L) involves selecting a sequence of connected nodes starting from v_i. The selection of each subsequent node is determined probabilistically, guided by the graph adjacency structure.

- *Sequential Pathway Modeling.* The forward propagation of the Mamba layer aggregates long-range dependencies along the sampled pathways. The selective sequential modeling of Mamba is well-suited for capturing such path information. For each sampled pathway $\mathbf{q^i} \in \mathcal{Q}(\mathbf{X^l})$, the Mamba layer processes the pathway sequentially as:

$$\begin{aligned} &\Delta_t = \tau_\Delta(f_\Delta(\mathbf{x}_t^l)), \quad \mathbf{B}_t = f_B(\mathbf{x}_t^l), \quad \mathbf{C}_t = f_C(\mathbf{x}_t^l), \\ &\mathbf{h}_t^l = (1 - \Delta_t \cdot \mathbf{D})\mathbf{h}_{t-1}^l + \Delta_t \cdot \mathbf{B}_t \mathbf{x}_t^l, \quad X_G^{l+1} = C \cdot h_L^{l+1}, \end{aligned} \quad (7)$$

where $\mathbf{x}_t^l$ is the t-th input node feature matrix in pathway $\mathbf{q^i}$ at layer l. f_* are learnable projections and $\mathbf{h}_t^e$ is hidden state. τ_Δ is the softplus function. The forgetting term $(1 - \Delta_t^e \cdot \mathbf{D})$ implements a selective mechanism analogous to synaptic decay or inhibitory processes that diminish outdated or irrelevant information. Conversely, the update term $\Delta_t^e \cdot \mathbf{B}_t^e$ mirrors gating that selectively reinforces and integrates salient new information. The projection $\mathbf{C}_t^e$ translates the internal state into observable outputs. By processing each sampled pathway individually, the Mamba layer effectively aggregates information along each pathway. The aggregated pathway representations are then combined to form the updated node features X_G^{l+1} for the next layer.

Afterward, we apply max pooling over the node features, i.e., $\{h_{v_i}\}_{i=1}^N$, followed by an MLP and softmax activation for the classification task.

PATHEXPLAINER: Targeted Pathway Inference

PATHEXPLAINER directly infers subgraphs to generate targeted pathways by leveraging the interpretability of PATHMAMBA. Vallina GNNexplainers (Ying et al. 2019; Luo et al. 2020), which focus primarily on the node or edge level, often struggle to capture the global structures at the pathway level. In contrast, PATHEXPLAINER introduces **novel pathway mask training**, where entire pathways (i.e., connected nodes and edges) are selectively masked during training to evaluate their contributions to PATHMAMBA.

Theoretical Objective. PATHEXPLAINER formalizes the identification of important subgraphs as an optimization problem. For a given graph $\mathcal{G}$ and its features $\mathbf{X}$, the explanation is defined as $(\hat{\mathcal{G}}, \hat{\mathbf{X}})$, where $\hat{\mathcal{G}} \subseteq \mathcal{G}$ is the subgraph and $\hat{\mathbf{X}}$ represents the selected features. The explanation is derived by optimizing the mutual information $\mathcal{MI}(\cdot)$ between the subgraph and the model's prediction, aiming to identify $\hat{\mathcal{G}}$ that captures the predictive rationale of PATHMAMBA:

$$\max_{\hat{\mathcal{G}}} \mathcal{MI}(Y, (\hat{\mathcal{G}}, \hat{\mathbf{X}})) = H(Y) - H(Y \mid \mathcal{G} = \hat{\mathcal{G}}, \mathbf{X} = \hat{\mathbf{X}}), \quad (8)$$

where $H(Y)$ is the entropy of the predictions Y and $H(Y \mid \mathcal{G} = \hat{\mathcal{G}}, \mathbf{X} = \hat{\mathbf{X}})$ is the conditional entropy given the explanation. A lower conditional entropy indicates a more faithful and informative representation of the prediction.

Optimization Framework. The optimization is approached by learning a pathway mask $\mathcal{M}$ for the sampled pathway's edges and nodes based on random pathways $\mathcal{Q}$ as described in Section . For each node v_i, a random pathway q_i of length up to L is sampled. These pathways are then used to restrict the mask learning process within the sampled pathways, ensuring that the learnable pathway mask $\mathcal{M}$ focuses on them. Specifically, the targeted subgraph $\hat{\mathcal{G}}$ is inferred based on $\mathbf{M}$ as: $\hat{\mathcal{G}} = \mathcal{G} \odot \sigma(\mathcal{M})$, where σ denotes the sigmoid function. The loss function for PATHEXPLAINER combines two components: a cross-entropy term for prediction consistency and regularization terms for sparsity:

$$\mathcal{L}_{\text{mask}} := -\sum_{c=1}^{C} \mathbb{1}[y = c] \log P_\Phi(Y = y \mid \mathcal{G} = \hat{\mathcal{G}}, \mathbf{X} = \hat{\mathbf{X}}) + \lambda \|\mathcal{M}\|, \quad (9)$$

where $\|M\|$ encourages sparsity in the edge selection, Y is a random variable representing labels $\{1, 2, \ldots, \text{C}\}$, and λ balances the trade-off between the prediction consistency and the sparsity regularization. Hence, the identified important subgraphs and node features that contribute most to specific bio-networks are considered as targeted pathways.

Theoretical Analysis for Targeted Pathway Fidelity

In this section, we place our method within the Weisfeiler–Lehman (WL) hierarchy to characterize its expressive

Methods	Human Diseases		Metabolism		Organismal Systems		Molecular & Cellular Processes		Overall
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Accuracy
GCN	0.632 ± 0.013	0.669 ± 0.022	0.895 ± 0.009	0.958 ± 0.007	0.644 ± 0.037	0.630 ± 0.023	0.570 ± 0.033	0.357 ± 0.025	0.683 ± 0.056
GraphSAGE	0.583 ± 0.020	0.633 ± 0.072	0.890 ± 0.007	0.959 ± 0.014	0.553 ± 0.041	0.575 ± 0.031	0.526 ± 0.059	0.337 ± 0.062	0.632 ± 0.037
GAT	0.630 ± 0.015	0.643 ± 0.036	**0.932 ± 0.017**	0.970 ± 0.008	0.659 ± 0.015	**0.703 ± 0.010**	0.560 ± 0.058	0.370 ± 0.025	0.690 ± 0.018
GIN	0.688 ± 0.023	0.697 ± 0.014	0.912 ± 0.016	0.944 ± 0.022	0.629 ± 0.025	0.638 ± 0.041	0.606 ± 0.032	0.497 ± 0.027	0.717 ± 0.013
GPS	0.744 ± 0.018	0.729 ± 0.024	0.893 ± 0.006	0.955 ± 0.014	0.634 ± 0.026	0.658 ± 0.011	0.629 ± 0.060	**0.507 ± 0.019**	0.726 ± 0.014
Graph-Mamba	0.707 ± 0.024	0.712 ± 0.024	0.897 ± 0.009	0.967 ± 0.007	0.626 ± 0.021	0.663 ± 0.033	**0.700 ± 0.021**	0.463 ± 0.032	0.723 ± 0.014
PATHMAMBA	**0.786 ± 0.029**	**0.800 ± 0.033**	0.915 ± 0.011	**0.972 ± 0.005**	**0.670 ± 0.026**	**0.703 ± 0.010**	0.667 ± 0.035	0.497 ± 0.028	**0.744 ± 0.015**
w/ 3B	0.752 ± 0.022	0.726 ± 0.027	0.917 ± 0.008	0.973 ± 0.010	0.661 ± 0.017	0.663 ± 0.023	0.656 ± 0.032	0.550 ± 0.042	0.742 ± 0.009
w/ 150M	0.764 ± 0.031	0.764 ± 0.011	0.906 ± 0.011	0.975 ± 0.013	0.639 ± 0.023	0.688 ± 0.025	0.653 ± 0.029	0.510 ± 0.030	0.728 ± 0.013
w/ 35M	0.748 ± 0.033	0.751 ± 0.019	0.914 ± 0.005	0.969 ± 0.007	0.634 ± 0.028	0.663 ± 0.028	0.633 ± 0.055	0.510 ± 0.049	0.722 ± 0.013
w/o ESM-2	0.380 ± 0.008	0.585 ± 0.015	0.669 ± 0.015	0.585 ± 0.015	0.241 ± 0.007	0.063 ± 0.019	0.378 ± 0.030	0.377 ± 0.043	0.440 ± 0.010

Table 1: Baseline comparison results on bio-network classification. The best and second-best results are highlighted in **bold** and underline, respectively. The gray-shaded rows indicate PATHMAMBA with different ESM-2 (encoder) parameter settings.

power. By proving that our explainer goes beyond the 1-WL limitation, we ensure that the extracted pathways capture higher-order structural patterns, establishing a theoretical upper bound on fidelity and inference ,supporting the empirical results.

Lemma 1. *(Expressiveness for explanations). When combined with higher expressive models (*e.g.*, it distinguishes more graphs),* PATHEXPLAINER *can generate more finely differentiated (and potentially more "faithful") explanation pathways (subgraphs). In contrast, a less expressive models merges different graphs into larger equivalence classes, leading to non-unique, less granular explanations.*

We prove this by showing that the expressiveness of the underlying a graph classifier f determines the granularity of equivalence classes, with more expressive models enabling finer distinctions between graphs.

Lemma 2 (Comparison with k-WL test)**.** *For every $k \geq 1$ there are graphs that are distinguishable by* PATHMAMBA*, but not by k-WL (and hence not by k-WL GNNs).*

Proof. The proof of this theorem directly comes from the recent work (Tonshoff et al. 2023; Behrouz and Hashemi 2024). They prove a similar theorem using 1-d CNNs (Tonshoff et al. 2023) or SSM (Behrouz and Hashemi 2024) with randomly sampled subgraphs. Since our method adopts Mamba (an SSM architecture) combined with the random sampling strategy, their theoretical results are directly applicable to our setting. □

Lemma 3 (Comparison with 1-WL test)**.** PATHMAMBA *is strictly more expressive than 1-WL GNNs.*

Proof. We first note that PATHMAMBA contains the GIN as a sub-module, which has the same expressive power as the 1-WL test (Xu et al. 2019). Therefore, PATHMAMBA is at least as expressive as 1-WL GNNs. By Lemma 2, there are graphs that cannot be distinguished by 1-WL GNNs, but can be distinguished by PATHMAMBA. Consequently, PATHMAMBA is strictly more expressive than 1-WL GNNs. □

Theorem 1 (Explanations of EXPATH)**.** *Based on Lemma 1, 2, and 3,* EXPATH *can generate more finely differentiated (and potentially more "faithful") explanation pathways (subgraphs) than 1-WL GNN-based methods, and not bounded by any WL GNN methods.*

Methods	Training Time (msec)	Inference Time (msec)
GPS	29.2 ± 2.3	10.3 ± 0.3
Graph-Mamba	34.8 ± 0.4	9.5 ± 0.2
PATHMAMBA	$\mathbf{24.4 \pm 0.9}$	$\mathbf{6.9 \pm 0.2}$

Table 2: The computational efficiency comparison with hybrid models, including both training and inference runtime.

Experiments and Results

Dataset and Preprocessing. We collected all available human pathway networks from the widely used knowledge database, KEGG (Kanehisa and Goto 2000). Our dataset consists of four main classes: Human Diseases, Metabolism, Molecular and Cellular Processes, and Organismal Systems, covering 301 bio-networks. For nodes, we ensured that all protein nodes in the network were linked to their reference AA sequence data.

Experimental Setup. We conducted 10-fold stratified K-Fold cross-validation repeated five times. The optimal hyperparameters were determined using grid search. Training for all models was implemented on NVIDIA A6000 GPU and Xeon Gold 6258R CPU.

Experiment-I: Pathway Learning

Objective. This experiment aims to evaluate whether EXPATH can classify diverse bio-networks and benchmark its performance against baseline models.

Baselines and Metrics. We collected baselines from both message-passing GNNs and more advanced graph models, including GCN (Kipf and Welling 2017), GraphSAGE (Hamilton, Ying, and Leskovec 2017), GAT (Veličković et al. 2018), GIN (Xu et al. 2019), GPS (Rampášek et al. 2022), and Graph-Mamba (Wang et al. 2024). We employed precision, recall, and overall accuracy for the performance evaluation. We used 650M ESM-2 for PATHMAMBA and all baselines as the node feature encoding model.

Results. Table 1 demonstrates that PATHMAMBA achieves the highest accuracy (0.744), outperforming all GNNs, GPS (0.726), GraphMamba (0.723). Furthermore, it secures best

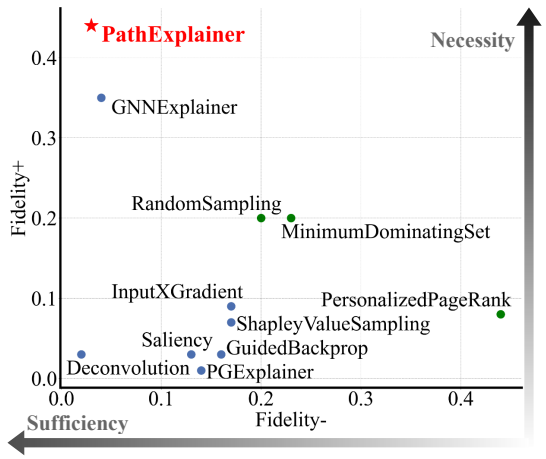

Figure 3: Fidelity+ (necessity ↑) and Fidelity- (sufficiency ↓) scores of extracted subgraphs. Our PATHEXPLAINER achieves the best performance on both metrics.

or second-best positions across all functional categories, demonstrating its robust ability to generalize across diverse pathway structures. The gray-shaded rows indicate the results of removing ESM-2 and modifying the model size in terms of F1 scores. When ESM-2 is removed, the accuracy decreases significantly (0.74 → 0.44). The results highlight the importance of AA-seq and the limitations of prior studies that were unable to leverage this information.

Table 2 compares the training and inference times of our model with other expressive hybrid models, using a batch size of 32. Our training time is 30% faster than GPS, and inference time is 27% faster than Graph-Mamba.

Experiment-II: Pathway Inference

Objective. This experiment aims to quantify the fidelity of extracted subgraphs using PATHEXPLAINER and validate the importance of pathways specific to biological functions.

Baselines. We use three baseline categories–(i) conventional statistical methods: RandomSubgraphSampling (RSS) (Kashtan et al. 2004b), PersonalizedPageRank (PPR) (Iván and Grolmusz 2011), and MinimumDominatingSet (MDS) (Nacher and Akutsu 2016; Wuchty 2014)); (ii) gradient-based methods: Saliency (Simonyan, Vedaldi, and Zisserman 2013), InputXGradient (Shrikumar, Greenside, and Kundaje 2017), Deconvolution (Mahendran and Vedaldi 2016), ShapleyValueSampling (Strumbelj and Kononenko 2010), and GuidedBackpropagation (Springenberg et al. 2014); (iii) GNN-specific explainer method: GNNExplainer (Ying et al. 2019) and PGExplainer (Luo et al. 2020).

Metrics. We evaluated the distinctiveness of the pathways inferred by PATHEXPLAINER using fidelity metrics, *Fidelity+* and *Fidelity-*. Fidelity+ measures how well the selected features support accurate predictions, while Fidelity- checks how much accuracy drops when only those features are kept. We further evaluated the length of pathways. *Max Path Length* captures the longest simple path in each subgraph, reflecting whether long signaling chains are retained. *Average Diameter* measures the typical node-to-node distance, showing how spread out the nodes remain after extraction.

Methods	Max Path Length	Average Diameter
Minimum Dominating Set	6	2.00
Random Sampling	11	2.90
Personalized Page Rank	4	1.53
PATHMAMBA-GNNExplainer	9	2.90
GPS-PATHEXPLAINER	12	3.95
EXPATH (Ours)	**16**	**4.20**

Table 3: Comparison of pathway-preservation ability across subgraph extraction baselines. Higher path length and diameter indicate better retention of long-range interactions.

Results. Figure 3 shows that PATHEXPLAINER achieves the highest fidelity+ and the lowest fidelity-. The main reason is that GNNExplainers optimise a mask for each individual node or edge level, whereas PATHEXPLAINER infers pathways (subgraphs) as a single coherent unit. Deconvolution simply aggregates all edges with large gradients, it almost covers every active edge. This pushes low fidelity$^-$, but since it retains many redundant edges, removing them hardly changes the output, so *fidelity*$^+$ (necessity) stays low. GNN-specific or gradient-based methods (blue points) show lower fidelity- compared to traditional methods (green points), indicating that the learned AA-seq enables the identification of sufficient subgraphs.

Table 3 presents that our method attains up to 4 × longer preserved paths and up to 2.7 × larger diameters than competing approaches. This supports that the identified sufficient and necessary features capture biologically meaningful pathways and meets our Expectation.

Experiment-III: Biological Meaningfulness

Objective. We propose an evaluation workflow to analyze the biological significance of the subgraphs and pathways extracted from our method. This workflow should integrate the *weighting/ranking scores of pathway inferred by* EXPATH into biological metrics, enabling the direct quantification of outputs from the models.

Proposed Evaluation Metrics. We designed experiments centered on Gene Ontology (GO) analysis (Ashburner et al. 2000), focusing on the nodes within the extracted subgraphs. The results provide a list of GO terms highlighting the biological functions most significantly represented in the input gene (corresponding to protein) nodes (Ashburner et al. 2000). Then we proposed Number of Enriched Biological Functions (**#EBF**) and Enrichment Contribution Score (**ECS**) to evaluate breadth and depth of the extracted functions (Yang et al. 2025). A higher #EBF indicates broader functional diversity within the subgraph. ECS evaluates the relative contribution of the top-weighted genes.

Results. Table 4 presents the biological meaningfulness comparison results for subgraphs extracted using different methods. Overall, EXPATH achieves the highest performance across #EBF, ECS, and P-value. This highlights its ability to extract biologically relevant structures within pathway networks, effectively balancing breadth and depth. While conventional methods (RSS, MDS, and PPR) per-

Methods	#EBF (↑)	ECS (↑)	P-value (↓)
RSS	5.29	0.27	0.045
MDS	6.34	0.23	0.043
PPR	6.64	0.23	0.042
GIN-GNNE	6.94	0.59	0.041
GPS-GNNE	8.88	0.22	0.039
GraphMamba-GNNE	10.73	0.21	0.042
PathMamba-GNNE	<u>11.89</u>	<u>0.73</u>	**0.036**
GIN-PathE	11.06	0.69	0.041
GPS-PathE	8.26	0.43	<u>0.037</u>
GraphMamba-PathE	10.89	0.59	0.038
ExPath	**14.77**	**0.84**	**0.036**

Table 4: Biological meaningfulness comparison results: The best-performing results are highlighted in **bold**. The second-best results are highlighted in <u>underline</u>.

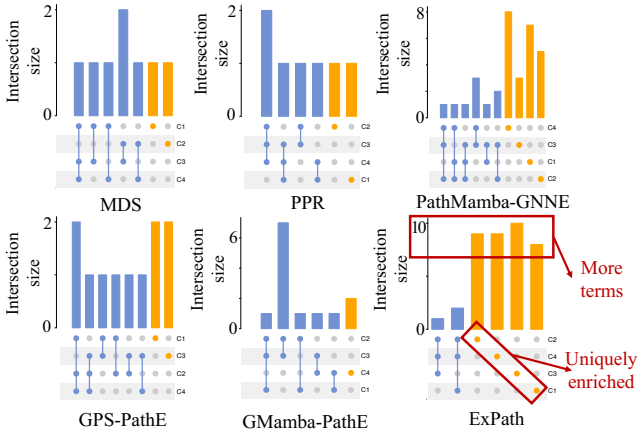

Figure 4: UpSet plot of enriched GO terms across four pathway classes, based on top feature sets from subgraphs for different methods. Orange indicates GO terms uniquely enriched in one class, and blue represents GO terms shared across multiple classes.

form relatively poorly in overall #EBF and ECS, with almost boundary P-values achieved.

Figure 4 evaluates the differences in enriched GO terms across four pathway classes based on top gene sets from subgraphs extracted by different methods. The upset plot reveals that ExPath identifies the most extensive sets of unique GO terms (shown as the orange bars and links) across all four pathway classes while maintaining fewer shared terms (shown as the blue bars and links) among different classes. This suggests that ExPath tends to assign appropriate weights to genes based on their importance within the network and effectively captures the distinct biological roles of top-ranked genes in specific pathway classes.

Experiment IV: PoC of Biological Case Study

Setup. We make a Proof-of-Concept case analysis using the T cell receptor (TCR) signaling pathway, which is a well-characterized human pathway. In this case study, we compare subgraphs extracted by two methods: TCR Subgraph A,

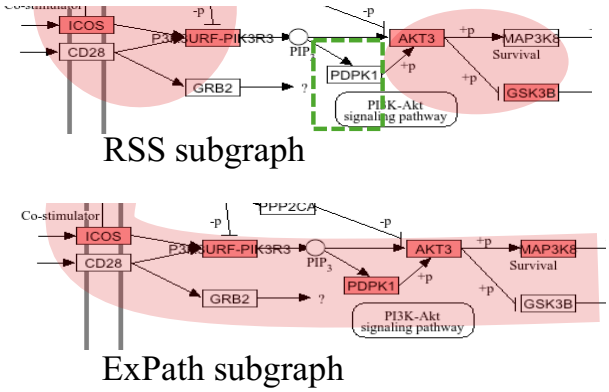

Figure 5: Comparison of subgraphs extracted from the TCR signaling pathway. The subgraph nodes and their signaling modules are colored in red. The disruptions within signaling paths are marked in green boxes.

generated using the RSS method, and Subgraph B, obtained via our proposed method. Each method selects the top 10% highest-ranked nodes and their associated edges to construct a representative subgraph.

Results. In Figure 5, the upper subfigure presents the TCR subgraph extracted by the RSS method, while the lower one corresponds to our ExPath. In subgraph A, generated by the RSS method, high scores are distributed uniformly across a broad range of nodes within the TCR pathway. However, this suggests unnatural, fragmented signal propagation, as evidenced by the numerous isolated red-marked nodes and a green broken connection in PDPK1. In contrast, subgraph B, extracted by our method, exhibits a strong focus on the PI3K-AKT signaling axis (Vara et al. 2004) and the downstream components of the MAP3K8 survival (Dolcet et al. 2005), as highlighted by a coherent red-marked path.

Discussion. In summary, the extracted subgraphs by ExPath align with the needs of real-world pathway analysis practices: maintaining signal continuity within regulatory cascades and even accommodating relatively long signaling paths, making them more suitable for focused analyses of bio-network regulatory mechanisms.

Conclusion

We introduced ExPath, a novel framework for understanding targeted pathways within biological knowledge bases. ExPath integrates PathMamba, a hybrid model to capture local and global dependencies; and PathExplainer, a subgraph learning module that identifies key nodes and edges via trainable pathway masks. ExPath seamlessly integrated biological foundation models to encode the experimental molecular data. We also introduced machine-learning-oriented biological evaluations and a new metric. The experiments involving 301 bio-networks evaluations demonstrated that pathways inferred by ExPath maintain biological meaningfulness. Future work will expand ExPath to analyze other types of bio-networks, enabling broader applications in systems biology and medicine.

Acknowledgments

The authors would like to thank the reviewers. This work was supported by JST BOOST, Japan Grant Number JPMJBS2402, "Program for Leading Graduate Schools" of The University of Osaka, JSPS KAKENHI Grant-in-Aid for Scientific Research Number JP24K20778, NSF award SCH-2205289, SCH-2014438, and IIS-2034479, JST CREST JPMJCR23M3, JST START JPMJST2553, JST CREST JPMJCR20C6, JST K Program JPMJKP25Y6, JST COI-NEXT JPMJPF2009, JST COI-NEXT JPMJPF2115, the Future Social Value Co-Creation Project - The University of Osaka.

References

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1): 25–29.

Behrouz, A.; and Hashemi, F. 2024. Graph Mamba: Towards Learning on Graphs with State Space Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, 119–130.

Cheng, Z.; Yan, C.; Wu, F.-X.; and Wang, J. 2021. Drug-target interaction prediction using multi-head self-attention and graph attention network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(4): 2208–2218.

Dolcet, X.; Llobet, D.; Pallares, J.; and Matias-Guiu, X. 2005. NF-kB in development and progression of human cancer. *Virchows archiv*, 446: 475–482.

Gligorijević, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H.; et al. 2021. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1): 3168.

Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*.

Haynes, T. W.; Hedetniemi, S.; and Slater, P. 2013. *Fundamentals of domination in graphs*. CRC press.

Ideker, T.; and Krogan, N. J. 2012. Differential network biology. *Molecular systems biology*, 8(1): 565.

Iván, G.; and Grolmusz, V. 2011. When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks. *Bioinformatics (Oxford, England)*, 405—407.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873): 583–589.

Kanehisa, M.; Furumichi, M.; Sato, Y.; Matsuura, Y.; and Ishiguro-Watanabe, M. 2024. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research*, D672–D677.

Kanehisa, M.; and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1): 27–30.

Kanehisa, M.; and Sato, Y. 2020. KEGG Mapper for inferring cellular functions from protein sequences. *Protein Science*, 29(1): 28–35.

Karger, D. R. 1994. Random sampling in cut, flow, and network design problems. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, 648–657.

Kashtan, N.; Itzkovitz, S.; Milo, R.; and Alon, U. 2004a. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11): 1746–1758.

Kashtan, N.; Itzkovitz, S.; Milo, R.; and Alon, U. 2004b. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 1746–1758.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

Li, M.; Wang, Z.; Liu, L.; Liu, X.; and Zhang, W. 2024. Subgraph-Aware Graph Kernel Neural Network for Link Prediction in Biological Networks. *IEEE Journal of Biomedical and Health Informatics*.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 1123–1130.

Luo, D.; Cheng, W.; Xu, D.; Yu, W.; Zong, B.; Chen, H.; and Zhang, X. 2020. Parameterized Explainer for Graph Neural Network. In *Advances in Neural Information Processing Systems*, 19620–19631.

Maehara, T.; Akiba, T.; Iwata, Y.; and Kawarabayashi, K.-i. 2014. Computing personalized pagerank quickly by exploiting graph structures. *Proceedings of the VLDB Endowment*, 7(12): 1023–1034.

Mahendran, A.; and Vedaldi, A. 2016. Salient deconvolutional networks. In *ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, 120–135.

Muzio, G.; O'Bray, L.; and Borgwardt, K. 2021. Biological network analysis with deep learning. *Briefings in bioinformatics*, 22(2): 1515–1530.

Nacher, J. C.; and Akutsu, T. 2016. Minimum dominating set-based methods for analyzing biological networks. *Methods*, 102: 57–63.

Rampášek, L.; Galkin, M.; Dwivedi, V. P.; Luu, A. T.; Wolf, G.; and Beaini, D. 2022. Recipe for a General, Powerful, Scalable Graph Transformer. In *Advances in Neural Information Processing Systems*, 14501–14515.

Ravindra, N.; Sehanobish, A.; Pappalardo, J. L.; Hafler, D. A.; and van Dijk, D. 2020. Disease state prediction from single-cell data using graph attention networks. In *Proceedings of the ACM conference on health, inference, and learning*, 121–130.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, ICML'17, 3145–3153.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Strumbelj, E.; and Kononenko, I. 2010. An Efficient Explanation of Individual Classifications using Game Theory. *J. Mach. Learn. Res.*, 1–18.

Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; Nastou, K.; Mehryary, F.; Hachilif, R.; Gable, A. L.; Fang, T.; Doncheva, N. T.; Pyysalo, S.; et al. 2023. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1): D638–D646.

Tonshoff, J.; Ritzert, M.; Wolf, H.; and Grohe, M. 2023. Walking Out of the Weisfeiler Leman Hierarchy: Graph Learning Beyond Message Passing. *Transactions on Machine Learning Research*.

Vara, J. Á. F.; Casado, E.; de Castro, J.; Cejas, P.; Belda-Iniesta, C.; and González-Barón, M. 2004. PI3K/Akt signalling pathway and cancer. *Cancer treatment reviews*, 30(2): 193–204.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

Wang, C.; Tsepa, O.; Ma, J.; and Wang, B. 2024. Graph-Mamba: Towards Long-Range Graph Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2402.00789*.

Wang, H.; Zheng, H.; Browne, F.; and Wang, C. 2014. Minimum dominating sets in cell cycle specific protein interaction networks. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 25–30. IEEE.

Wang, J.; Ma, A.; Chang, Y.; Gong, J.; Jiang, Y.; Qi, R.; Wang, C.; Fu, H.; Ma, Q.; and Xu, D. 2021. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nature communications*, 12(1): 1882.

Wuchty, S. 2014. Controllability in protein interaction networks. *Proceedings of the National Academy of Sciences*, 7156–7160.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations*.

Yang, Z.; Chen, Z.; Liu, X.; Kotoge, R.; Chen, P.; Matsubara, Y.; Sakurai, Y.; and Sun, J. 2025. GeSubNet: Gene Interaction Inference for Disease Subtype Network Generation. In *The Thirteenth International Conference on Learning Representations*.

Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; and Leskovec, J. 2019. GNNExplainer: Generating Explanations for Graph Neural Networks. In *Advances in Neural Information Processing Systems*.

Yue, X.; Wang, Z.; Huang, J.; Parthasarathy, S.; Moosavinasab, S.; Huang, Y.; Lin, S. M.; Zhang, W.; Zhang, P.; and Sun, H. 2020. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics*, 36(4): 1241–1251.

Zhang, X.-M.; Liang, L.; Liu, L.; and Tang, M.-J. 2021. Graph neural networks and their current applications in bioinformatics. *Frontiers in genetics*, 12: 690049.