Graph Unitary Message Passing

Anonymous authors
Paper under double-blind review

Abstract

Unitarity is a useful principle for stabilizing deep neural networks, but in graph neural networks (GNNs) instability is induced not only by learnable parameters but also by the graph propagation operator. Motivated by this distinction, we propose Graph Unitary Message Passing (GUMP), a message-passing framework that uses a unitary propagation operator on a transformed graph to avoid graph-induced exponential decay under repeated propagation. GUMP combines (i) a graph transformation that maps an input graph to an Eulerian line-graph construction admitting unitary adjacency matrices, and (ii) a practical unitary projection procedure based on Newton-Schulz iteration. Theoretical analysis clarifies that, under standard analysis assumptions, unitary propagation keeps the graph-propagation term depth-stable, while vanilla normalized propagation exhibits exponential decay in its non-trivial spectral components. Across synthetic long-range tasks, TUDataset benchmarks, and LRGB datasets, GUMP consistently improves over vanilla message passing and competitive baselines.

1 Introduction

Unitarity has become an important principle for stabilizing deep neural networks and preventing gradient pathologies. Prior work has applied it through orthogonal or unitary initialization (Saxe et al., 2013; Arjovsky et al., 2016; Orvieto et al., 2023; De et al., 2024), recurrent architectures with unitary transition matrices (Arjovsky et al., 2016; Jing et al., 2017), and optimizers such as Muon (Jordan et al., 2024; Liu et al., 2025). Across these settings, the common benefit is that unitary transformations preserve norms and make deep signal propagation easier to control.

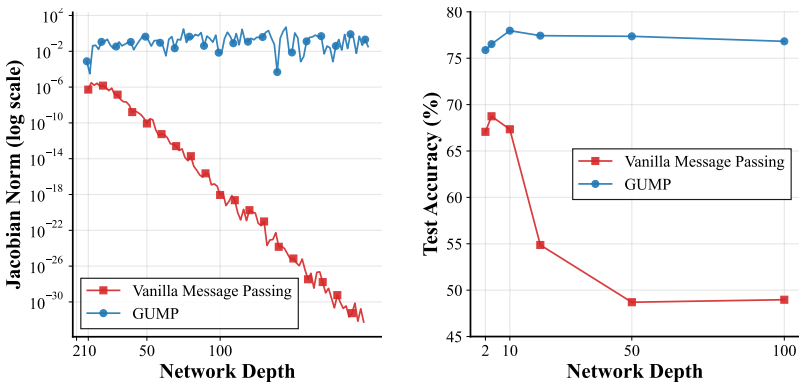


Figure 1: Compared to vanilla message passing (graph convolution), GUMP exhibits a stable Jacobian norm (left), which leads to more stable model performance as the number of layers increases (right).

However, constraining only the learnable parameters is insufficient for graph neural networks (GNNs), because repeated message passing is also governed by the graph structure itself. Nodes update their representations by aggregating information from neighbors (Gilmer et al., 2017), so the propagation dynamics depend on the adjacency operator rather than only on trainable weights. This creates graph-dependent forms of

instability, most notably oversquashing and oversmoothing (Alon & Yahav, 2020; Topping et al., 2022; Chen et al., 2020a), which limit the benefits of depth for long-range reasoning on graphs.

Here, we extend the principle of unitarity from model parameters to the graph propagation operator itself. We propose Graph Unitary Message Passing (GUMP), a message-passing framework that replaces the usual normalized adjacency propagation with a unitary propagation operator on a transformed graph. The key point is not that all components of a GNN become unitary, but that the graph-induced factor in repeated propagation no longer decays exponentially with depth. This targets the graph-structure side of training instability, complementing standard practices that keep learnable weight matrices well conditioned.

To implement this idea, we transform a general graph into an Eulerian line-graph construction that admits unitary adjacency matrices while preserving the admissible edge-to-edge transitions induced by the original graph. We then use Newton-Schulz iteration (Kovarik, 1970; Björck & Bowie, 1971) to compute the unitary propagation matrix efficiently. Experiments across diverse graph learning tasks show that GUMP improves over vanilla message passing and strong baselines.

Notations In this paper, we use bold uppercase letters $\mathbf{X}$ to denote matrices, bold lowercase letters $\mathbf{x}$ to denote vectors, and lowercase letters x to denote scalars. Given a matrix $\mathbf{X}$, the i-th row of matrix $\mathbf{X}$ is denoted as $\mathbf{x}_i$ and the entry of the i-th row and j-th column of matrix $\mathbf{X}$ is denoted as $\mathbf{X}_{ij}$. The transpose and conjugate transpose of matrix $\mathbf{X}$ is denoted as $\mathbf{X}^\top$ and $\mathbf{X}^\dagger$, respectively. A graph with n nodes and e edges is denoted as $G = (V, E, \mathbf{X})$ where $V = \{1, 2, \cdots, n\}$ is the node set, $E \subseteq V \times V$ is the edge set, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the d-dimensional node feature matrix. For convenience, the operator $\mathsf{V}[G]$ and $\mathsf{E}[G]$ are used to denote the node set and edge set of graph G respectively, i.e., $\mathsf{V}[G] = V$ and $\mathsf{E}[G] = E$. The adjacency matrix of graph G is denoted as $\tilde{\mathbf{A}}[G] \in \{0, 1\}^{n \times n}$ where $\tilde{\mathbf{A}}_{ij} = 1$ if $(i, j) \in E$ and $\tilde{\mathbf{A}}_{ij} = 0$ otherwise. The normalized adjacency matrix of graph G is denoted as $\hat{\mathbf{A}}[G] \in \mathbb{R}^{n \times n}$ where $\hat{\mathbf{A}}[G] = \mathbf{D}^{-1/2}\tilde{\mathbf{A}}[G]\mathbf{D}^{-1/2}$ and $\mathbf{D}$ is the degree matrix of graph G. We also use matrix $\mathbf{A}[G] \in \mathbb{R}^{n \times n}$ to represent the general adjacency matrix in graph G, i.e., $\mathbf{A}_{ij} \neq 0$ if $(i, j) \in E$ and $\mathbf{A}_{ij} = 0$ otherwise. Therefore, without specifying the type of adjacency matrix, $\mathbf{A}[G]$ can also represent $\tilde{\mathbf{A}}[G]$ and $\hat{\mathbf{A}}[G]$. For convenience, the adjacency matrices above are denoted as $\tilde{\mathbf{A}}$, $\hat{\mathbf{A}}$, and $\mathbf{A}$ respectively. Some preliminaries used in this paper are provided in appendix. Finally, the GNN representation at layer k is denoted as $\mathbf{H}^{(k)} \in \mathbb{R}^{n \times d}$ with d being the dimension of node features, and the vector $\mathbf{h}_i^{(k)} \in \mathbb{R}^d$ denotes the GNN representation of node i at layer k.

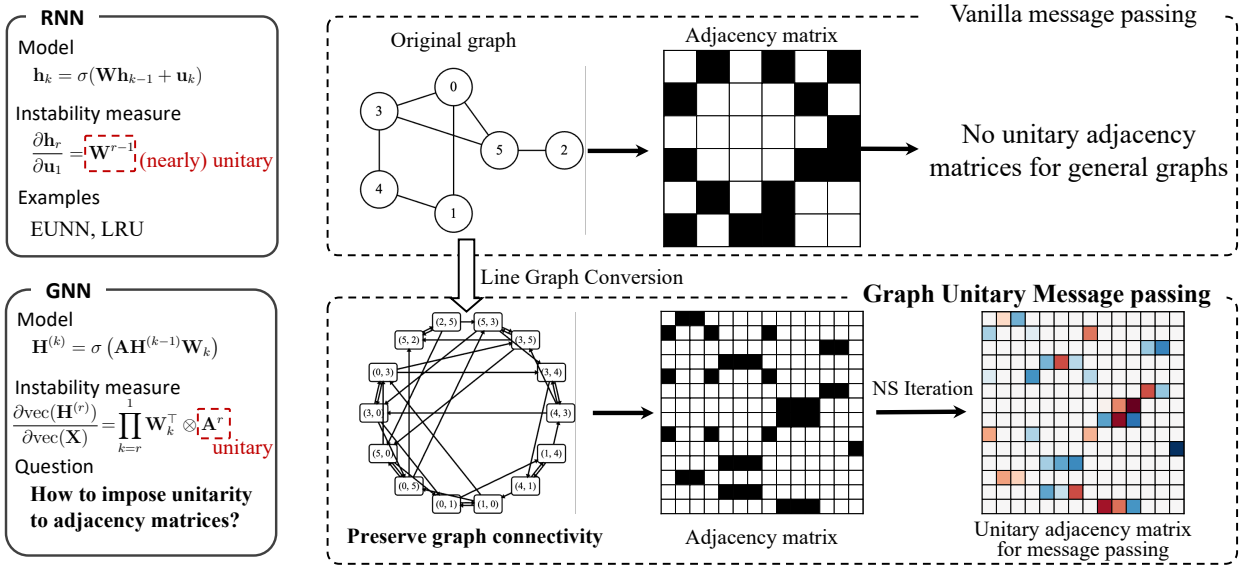
2 Preliminaries on Unitarity and Training Instability

Unitarity in Deep Learning A matrix $\mathbf{U} \in \mathbb{C}^{n \times n}$ is called *unitary* if $\mathbf{U}^\dagger\mathbf{U} = \mathbf{U}\mathbf{U}^\dagger = \mathbf{I}$. For real matrices, this reduces to the orthogonal condition $\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$. A fundamental property of unitary matrices is that they preserve vector norms: $\|\mathbf{U}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for any vector $\mathbf{x}$. Unitarity has emerged as an important principle in deep learning due to its ability to stabilize training and prevent gradient pathologies. Unitarity manifests across multiple domains:

- **Parameter Initialization:** Orthogonal and unitary initialization schemes (Saxe et al., 2013; Orvieto et al., 2023) have been shown to maintain signal propagation through deep networks by preserving the norms of activations. This prevents the vanishing and exploding gradient problems that plague randomly initialized networks, enabling effective training of deeper models.

- **Unitary Constraints:** The breakthrough work of Arjovsky et al. (2016) demonstrated that constraining the recurrent weight matrix to be unitary allows RNNs to capture long-range dependencies without suffering from vanishing gradients. Subsequent work (Jing et al., 2017; Orvieto et al., 2023; De et al., 2024) has built upon this foundation, showing that unitary RNNs can effectively model sequences with temporal dependencies spanning hundreds or thousands of time steps.

- **Optimizers:** Recent advances in optimization have leveraged unitarity for improved convergence. The Muon optimizer (Jordan et al., 2024; Liu et al., 2025) applies unitary transformations to parameter updates, achieving superior performance in training large language models and other deep architectures.



(a) Motivation (b) Method

Figure 2: Overview of GUMP. (a) RNN versus GNN. The measures of the stability are derived with the identity activation function and $\otimes$ denotes the Kronecker product. (b) GUMP aims to impose unitarity on $\mathbf{A}$ in message passing. GUMP achieves this goal with graph transformation (Section 3.2) and unitary adjacency matrix calculation (Section 3.3).

Training Instability in GNNs To understand the training instability in GNNs, we first introduce the instability in RNNs from Figure 2(a). We simply formulate RNN as $\mathbf{h}_k = \sigma(\mathbf{W}\mathbf{h}_{k-1} + \mathbf{u}_k)$ with $\mathbf{h}_k$ being the hidden state at layer k, $\mathbf{W}$ being the transformation matrix, $\mathbf{u}_k$ being the input at time step k, and σ being the activation function. The Jacobian of the hidden state with respect to the input is used to measure the training instability of RNN (Arjovsky et al., 2016). If the Jacobian norm $\|\partial \mathbf{h}_k / \partial \mathbf{u}_0\|_2$ grows exponentially with k, the gradient will explode, and if it decays exponentially, the gradient will vanish. When imposing unitarity on $\mathbf{W}$, the Jacobian norm is bounded, thus improves the training stability of RNNs.

For graphs, the training instability is mainly caused by the message passing mechanism. In Figure 2(a), a one-hop MPNN layer for GNN is formulated as $\mathbf{H}^{(k)} = \sigma(\mathbf{A}\mathbf{H}^{(k-1)}\mathbf{W}_k)$ for simplicity. Similar to RNN (Pascanu et al., 2013), the Jacobian $\partial \mathbf{H}^{(r)} / \partial \mathbf{X}$ (Topping et al., 2022) can measure the training instability of GNN. When the Jacobian is large, the gradient will explode, and when it is small, the gradient will vanish. Therefore, we are motivated to impose unitarity on $\mathbf{A}$ in GNN to improve learning efficiency.

3 Graph Unitary Message Passing

3.1 Challenges of Imposing Unitarity

Imposing unitarity on an adjacency matrix is not as straightforward as that in RNNs. The challenge comes from the sparsity of the adjacency matrix, as almost all unitary matrices are dense (Figure 2(b)). Therefore, the number of graphs with unitary adjacency matrices is limited. Also, the unitary adjacency matrix should be input-dependent and permutation-equivariant because the unitary adjacency matrix depends on input graph, and the order of nodes in a graph should not impact GNN representations.

In this section, the challenge of existence of unitary adjacency matrix is addressed by transforming the original graph to a special graph (Algorithm 1) which is guaranteed to have unitary adjacency matrices while preserving the admissible edge-to-edge transitions induced by the original graph at the same time. The challenge of preserving adjacency matrix property is addressed by calculating the unitary adjacency matrix with a unitary projection algorithm (Algorithm 2), which is implemented by utilizing the Newton-Schulz iteration and allows GUMP to be permutation-equivariant (Proposition 2). As a general one-hop

message-passing mechanism, any convolution operation can be combined with GUMP by setting the edge weights to be the entries of the unitary adjacency matrix for message passing. The overview of GUMP is shown in Figure 2(b).

Algorithm 1 Graph transformation

Require: A undirected graph $G = (V, E)$;
  Initialize a new digraph $G' = (V, E')$;
  for $(i, j) \in E$ **do**
    Add (i, j) and (j, i) to E';
  end for
  Convert G' to its line graph $\mathsf{L}(G')$;
  Return: A digraph $\mathsf{L}(G')$.

3.2 Graph Transformation: Convert Graph to Have Unitary Adjacency Matrix

Since unitary matrices are generally non-symmetric, the graph transformation algorithm should convert the original graph to a directed graph. We first formally define the unitary adjacency matrix as in Severini (2003). Given an adjacency matrix $\mathbf{A}$, its support matrix $\mathsf{S}[\mathbf{A}] \in \mathbb{R}^{n \times n}$ is a binary matrix with entries equal to one if the corresponding entry of $\mathbf{A}$ is non-zero and equal to zero otherwise, i.e., $\mathsf{S}[\mathbf{A}]_{ij} = 1$, if $\mathbf{A}_{ij} \neq 0$ and $\mathsf{S}[\mathbf{A}]_{ij} = 0$, if $\mathbf{A}_{ij} = 0$. Then, the unitary adjacency matrix $\mathbf{U}_G$ of graph G is a unitary matrix whose support is equal to the support of its adjacency matrix $\mathbf{A}$, i.e., $\mathsf{S}[\mathbf{U}_G] = \mathsf{S}[\mathbf{A}]$.

We propose the transformation in Algorithm 1 for undirected graph G to make it have unitary adjacency matrices while preserving the admissible edge-to-edge transitions induced by the original graph. Algorithm 1 first transforms the undirected graph into a digraph G' by splitting each undirected edge into two directed edges. For each connected component of G that contains at least one edge, the corresponding component of G' is connected and has equal in-degree and out-degree at every vertex, hence that component is Eulerian. In particular, when G is connected, G' is Eulerian. Then, it converts the resulting digraph to its line graph $\mathsf{L}(G')$ (see definition in appendix). Finally, $\mathsf{L}(G')$ has unitary adjacency matrices because of its specularity and strong quadrangularity properties (Severini, 2003), which is proved in the following proposition.

Proposition 1. *The line graph $\mathsf{L}(G')$ returned by Algorithm 1 has unitary adjacency matrices.*

Proposition 1 is proved in the appendix. The proof treats disconnected inputs componentwise and combines the resulting unitary blocks by a block-diagonal direct sum. Algorithm 1 does not literally preserve node-level adjacency, because the vertices of $\mathsf{L}(G')$ correspond to directed edges of G. Rather, it preserves the admissible edge-to-edge transitions induced by the original graph: two vertices of $\mathsf{L}(G')$ are adjacent only when the corresponding directed edges in G' can be traversed consecutively through a shared endpoint. Thus, the construction does not introduce spurious transitions across disconnected components or unrelated parts of the original graph. Finally, Algorithm 1 takes as input graph G with n nodes and e edges, and outputs a line graph $\mathsf{L}(G')$ with $2e$ nodes. Its edge set contains exactly $\sum_{v \in V} d(v)^2$ directed edges in the undirected-to-Eulerian construction, because each vertex v contributes all admissible transitions from its $d(v)$ incoming directed edges to its $d(v)$ outgoing directed edges. Therefore, constructing $\mathsf{L}(G')$ takes $\mathcal{O}\big(\sum_{v \in V} d(v)^2\big)$ time, which is $\mathcal{O}(|E|\Delta)$ for maximum degree Δ and $\mathcal{O}(|E|^2)$ in the worst case.

3.3 Unitary Adjacency Matrix Calculation: Compute the Edge Weights for Message Passing

According to Proposition 1, the line graph $\mathsf{L}(G')$ has unitary adjacency matrices. In this section, we propose an algorithm to calculate a unitary adjacency matrix for GUMP, because the unitary adjacency matrix depends on the input graph and should be calculated for each graph.

3.3.1 Permutation-equivariant Projection

Permutation equivariance of message passing is a key property for GNN to apply to graphs with varying node orders. To achieve this, the calculation of unitary adjacency matrix has to be permutation equivariant.

GUMP consists of two steps: 1) calculate edge weights to form weighted adjacency matrix; 2) impose unitarity on weighted adjacency matrix.

Firstly, edge weight for $(i, j) \in \mathsf{E}[\mathsf{L}(G')]$ is calculated by

$$\alpha_{ij} = \mathsf{Tanh}\left(\mathbf{w}^\top \cdot \mathsf{LeakyReLU}(\mathbf{W}_s \mathbf{h}_i + \mathbf{W}_t \mathbf{h}_j)\right), \tag{1}$$

where $\mathbf{h}_i$ ($\mathbf{h}_j$) is the representations for node i (j) in $\mathsf{L}(G')$, $\mathbf{W}_s, \mathbf{W}_t \in \mathbb{R}^{d' \times d}$ are transformation matrices for source and target nodes of an edge respectively, and $\mathbf{w} \in \mathbb{R}^{d'}$ is a learnable parameter. Then, the weighted adjacency matrix of $\mathsf{L}(G')$, denoted as $\mathbf{A}_w \in \mathbb{R}^{2e \times 2e}$, is formed from edge weights, i.e., $(\mathbf{A}_w)_{ij} = \alpha_{ij}$.

After calculating $\mathbf{A}_w$, we impose unitarity on $\mathbf{A}_w$ by projection. We use the projection algorithm in Keller (1975), which takes advantage of the fact that the polar transformation yields the closest unitary matrix to a given matrix in terms of the Frobenius norm, i.e., $\mathsf{U}[\mathbf{A}_w] = \arg\min_{\mathbf{U} \text{ is unitary}} \|\mathbf{A}_w - \mathbf{U}\|_F^2 = \mathbf{A}_w \left(\mathbf{A}_w^\dagger \mathbf{A}_w\right)^{-\frac{1}{2}}$. The unitary projection $\mathsf{U}[\mathbf{A}_w]$ is guaranteed to be permutation-equivariant when $\mathbf{A}_w$ is a full-rank matrix with the following proposition.

Proposition 2 (Strong permutation equivariance)**.** *Given two permutation matrices $\mathbf{P}_1$ and $\mathbf{P}_2$, if $\mathbf{A}_w$ is a full-rank matrix, the unitary projection $\mathsf{U}[\mathbf{A}_w]$ is equivariant to both row and column permutations of $\mathbf{A}_w$, i.e., $\mathbf{P}_1 \mathsf{U}[\mathbf{A}_w] \mathbf{P}_2^\top = \mathsf{U}[\mathbf{P}_1 \mathbf{A}_w \mathbf{P}_2^\top]$.*

By Proposition 2 (proved in appendix), the weighted adjacency matrix $\mathbf{A}_w$ should be full-rank to guarantee permutation equivariance of GUMP. Empirically, inspired by GATv2 (Brody et al., 2021), $\mathbf{A}_w$ induced by equation 1 is full-rank in experiments and thus can guarantee the permutation equivariance of GUMP. However, the unitary projection is computationally expensive due to the inverse square root of $\mathbf{A}_w^\dagger \mathbf{A}_w \in \mathbb{R}^{2e \times 2e}$.

Before discussing the efficient implementation, we first clarify why the projection step is compatible with the sparse message-passing structure of $\mathsf{L}(G')$. Although polar factors are dense in general, the Eulerian line-graph construction gives $\mathbf{A}_w$ a special row/column block structure indexed by the intermediate vertex shared by two consecutive directed edges. The next proposition makes this structure explicit and shows that the projection preserves it.

Proposition 3 (Admissible-transition support preservation)**.** *Let $\mathbf{A}_\mathsf{L}$ and $\mathbf{A}_w$ be the binary and weighted adjacency matrices of $\mathsf{L}(G')$, where $G' = (V, E')$ is returned by Algorithm 1. If rows are grouped by incoming directed edges $E_{\text{in}}(v)$ and columns by outgoing directed edges $E_{\text{out}}(v)$, then there exist permutation matrices $\mathbf{P}_{\text{in}}$ and $\mathbf{P}_{\text{out}}$ such that*

$$\mathbf{D} := \mathbf{P}_{\text{in}} \mathbf{A}_w \mathbf{P}_{\text{out}}^\top = \mathsf{diag}\big(\mathbf{B}_v\big)_{v \in V},$$

where $\mathbf{B}_v$ contains the weights for transitions $E_{\text{in}}(v) \to E_{\text{out}}(v)$ and is square because $|E_{\text{in}}(v)| = |E_{\text{out}}(v)|$. If $\mathbf{A}_w$ is full-rank, then

$$\mathbf{P}_{\text{in}} \mathsf{U}[\mathbf{A}_w] \mathbf{P}_{\text{out}}^\top = \mathsf{diag}\big(\mathsf{U}[\mathbf{B}_v]\big)_{v \in V},$$

and consequently

$$\mathsf{S}[\mathsf{U}[\mathbf{A}_w]] \subseteq \mathsf{S}[\mathbf{A}_\mathsf{L}],$$

with equality whenever each block polar factor $\mathsf{U}[\mathbf{B}_v]$ is fully supported, which holds generically for continuous full-rank learned weights.

Proposition 3 shows that GUMP does not rely on masking a dense polar factor after projection: the Eulerian line-graph construction induces the required block structure, and the polar projection preserves it.

3.3.2 Efficient Implementation of Unitarity Projection

Calculating the unitary projection $\mathsf{U}[\mathbf{A}_w] = \mathbf{A}_w (\mathbf{A}_w^\dagger \mathbf{A}_w)^{-\frac{1}{2}}$ directly involves computing the inverse square root of a matrix, which is computationally expensive and numerically unstable for large matrices, typically requiring singular value decomposition (SVD) with cubic complexity $\mathcal{O}((2e)^3)$. To address this, we employ Newton-Schulz iteration (Kovarik, 1970; Björck & Bowie, 1971; Jordan et al., 2024), an iterative method that converges quadratically to the unitary polar factor without expensive decompositions.

The Newton-Schulz iteration (Jordan et al., 2024) for computing the unitary factor $\mathbf{U}$ of a matrix $\mathbf{A}_w$ is given by the recurrence:

$$\mathbf{X}_{t+1} = \frac{15}{8}\mathbf{X}_t - \frac{5}{4}\mathbf{X}_t\left(\mathbf{X}_t^\top \mathbf{X}_t\right) + \frac{3}{8}\mathbf{X}_t\left(\mathbf{X}_t^\top \mathbf{X}_t\right)^2, \quad \mathbf{X}_0 = \frac{\mathbf{A}_w}{\|\mathbf{A}_w\|_F}. \tag{2}$$

Here, $\mathbf{X}_t$ converges to $\mathsf{U}[\mathbf{A}_w]$ as $t \to \infty$. The normalization by the Frobenius norm $\|\mathbf{A}_w\|_F$, which upper bounds the spectral norm, ensures the spectral norm is at most 1 for better convergence. In practice, a small number of iterations (e.g., $K = 5$ to 10) is sufficient to obtain a high-quality approximation of the unitary matrix. Because $\mathbf{A}_w$ can be written as a block diagonal matrix after the row/column permutations in Proposition 3, the Newton-Schulz iteration can be applied efficiently either by sparse matrix multiplications or by applying the iteration separately to each block $\mathbf{B}_v$ of $\mathbf{D}$ after the corresponding row/column permutations, making it significantly more efficient than SVD-based approaches for the sparse but large adjacency matrices encountered in GNNs. This yields an explicit complexity bound for the actual algorithm used in this paper. Because Algorithm 1 replaces each undirected edge by both orientations, we have $d_{\text{in}}(v) = d_{\text{out}}(v) = d(v)$ for every vertex v, so each block $\mathbf{B}_v$ has size $d(v) \times d(v)$. Therefore every Newton-Schulz iterate decomposes as independent block updates rather than a dense multiplication on the full $(2e)\times(2e)$ matrix. Equivalently, the same computation can be carried out directly by sparse matrix multiplication on the original support. Hence one iteration costs $\mathcal{O}\big(\sum_{v\in V} d(v)^3\big)$, and Algorithm 2 costs $\mathcal{O}\big(K\sum_{v\in V} d(v)^3\big)$ over K iterations, rather than $\mathcal{O}(K(2e)^3)$ for a naive dense implementation. For a more general directed-graph variant with rectangular blocks of size $d_{\text{in}}(v) \times d_{\text{out}}(v)$, the dominant sparse matrix products would scale as $\mathcal{O}\big(\sum_v d_{\text{in}}(v) d_{\text{out}}(v)^2\big)$ or $\mathcal{O}\big(\sum_v d_{\text{in}}(v)^2 d_{\text{out}}(v)\big)$ depending on the multiplication order.

Complexity and memory. Let $M_L = \sum_{v\in V} d(v)^2$ be the number of admissible directed transitions in the line graph, let F be the line-graph hidden feature dimension, and let $F_a = d'$ be the attention dimension used in equation 1. The graph transformation has one-time time and storage cost $\mathcal{O}(e + M_L)$ per input graph. After precomputing the source and target feature projections for the $2e$ line-graph nodes, calculating the learned edge weights costs $\mathcal{O}(2eFF_a + M_L F_a)$ time and stores $\mathcal{O}(M_L)$ scalar edge weights; this becomes $\mathcal{O}(2eF^2 + M_L F)$ only when $F_a = \mathcal{O}(F)$. The K-step blockwise Newton-Schulz projection costs $\mathcal{O}\big(K\sum_{v\in V} d(v)^3\big)$ time. Its output and forward working storage are $\mathcal{O}(M_L)$ when implemented blockwise with a constant number of temporaries, while standard automatic differentiation through all K iterations may retain $\mathcal{O}(KM_L)$ intermediate values unless checkpointing or a custom backward pass is used. These costs are in addition to storing the line-graph node representations, which require $\mathcal{O}(2eF)$ memory. Once $\mathsf{U}[\mathbf{A}_w]$ is obtained, a GCN-style GUMP propagation layer on the line graph costs $\mathcal{O}(M_L F + 2eF^2)$, whereas a standard GCN layer on the original graph costs $\mathcal{O}(eF + nF^2)$. Thus GUMP has a different cost profile from standard message passing: it is linear in graph size for bounded-degree sparse graphs, but the constants depend on the degree because M_L and the projection are degree-sensitive. For example, on an r-regular graph, $e = nr/2$, $M_L = nr^2 = 2er$, and the K-step projection costs $\mathcal{O}(Knr^3) = \mathcal{O}(Ker^2)$. Therefore high-degree vertices are the main bottleneck, since they dominate $\sum_v d(v)^2$ and $\sum_v d(v)^3$.

Algorithm 2 Blockwise calculation of unitary adjacency matrix via Newton-Schulz iteration

Require: $\mathsf{L}(G')$ outputted by Algorithm 1, iterations K;
1: Calculate $\mathbf{A}_w$ of $\mathsf{L}(G')$ with equation 1;
2: Find $\mathbf{P}_{\text{in}}, \mathbf{P}_{\text{out}}$ such that $\mathbf{D} = \mathbf{P}_{\text{in}}\mathbf{A}_w\mathbf{P}_{\text{out}}^\top = \mathsf{diag}(\mathbf{B}_v)_{v\in V}$ as in Proposition 3;
3: **for** each block $\mathbf{B}_v$ **do**
4: Normalize $\mathbf{X}_{v,0} = \mathbf{B}_v/\|\mathbf{B}_v\|_F$;
5: **for** $t = 0$ to $K-1$ **do**
6: $\mathbf{X}_{v,t+1} = \frac{15}{8}\mathbf{X}_{v,t} - \frac{5}{4}\mathbf{X}_{v,t}\left(\mathbf{X}_{v,t}^\top\mathbf{X}_{v,t}\right) + \frac{3}{8}\mathbf{X}_{v,t}\left(\mathbf{X}_{v,t}^\top\mathbf{X}_{v,t}\right)^2$;
7: **end for**
8: **end for**
9: $\mathsf{U}[\mathbf{D}] = \mathsf{diag}(\mathbf{X}_{v,K})_{v\in V}$;
10: $\mathsf{U}[\mathbf{A}_w] = \mathbf{P}_{\text{in}}^\top \mathsf{U}[\mathbf{D}]\mathbf{P}_{\text{out}}$;
11: **Return:** Approximately unitary adjacency matrix $\mathsf{U}[\mathbf{A}_w]$.

Algorithm 3 GNN with the graph unitary message passing mechanism (GNN-GUMP)

Require: A graph $G = (V, E, \mathbf{X})$;
1: $\mathbf{X}^{(0)} = \mathsf{GNN}(\mathbf{X}, G)$;
2: Transform G to $\mathsf{L}(G')$ with Algorithm 1;
3: Generate initial representation $\mathbf{H}^{(0)}$ for $\mathsf{L}(G')$ with $\mathbf{h}_{(i,j)} = [\mathbf{x}_i^{(0)}; \mathbf{x}_j^{(0)}], \forall (i,j) \in \mathsf{V}[\mathsf{L}(G')]$;
4: Calculate $\mathsf{U}[\mathbf{A}_w]$ with Algorithm 2;
5: **for** $k = 1 \cdots L$ **do**
6: $\quad \mathbf{h}_v^{(k)} = \gamma(\mathbf{h}_v^{(k-1)}, \sum_{u \in N_v} \mathsf{U}[\mathbf{A}_w]_{vu} \phi^{(k)}(\mathbf{h}_v^{(k-1)}, \mathbf{h}_u^{(k-1)}))$, $v \in V$
7: **end for**
8: Scatter $\mathbf{H}^{(L)}$ to nodes of G with $\mathbf{H}_s^{(L)} = \mathsf{Scatter}(\mathbf{H}^{(L)}, G)$;
9: Generate node representations of G with $\mathbf{X}^{(L)} = [\mathbf{X}^{(0)}; \mathbf{H}_s^{(L)}]$.
10: **Return:** Node representations $\mathbf{X}^{(L)}$ of G.

3.4 Apply GUMP to GNN

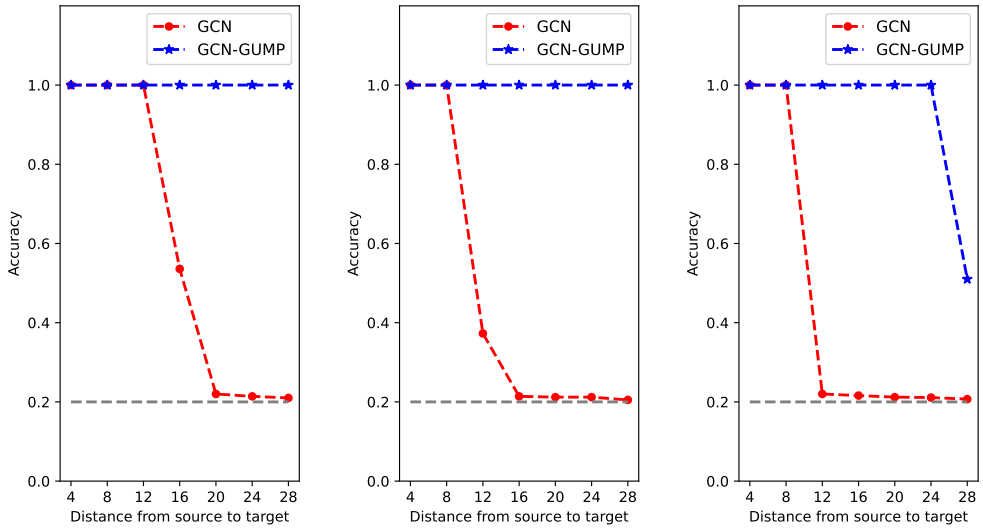
We can apply GUMP to different GNN architectures for graph learning tasks in Algorithm 3. Given a graph G, a base GNN first computes the initial node representations of G, i.e., $\mathbf{X}^{(0)} = \mathsf{GNN}(\mathbf{X}, G)$. Then, Algorithm 1 transforms G to $\mathsf{L}(G')$. The initial node representations $\mathbf{H}^{(0)} \in \mathbb{R}^{2e \times 2d}$ of $\mathsf{L}(G')$ are generated with $\mathbf{h}_{(i,j)} = [\mathbf{x}_i^{(0)}; \mathbf{x}_j^{(0)}]$, $\forall (i,j) \in \mathsf{V}[\mathsf{L}(G')]$ $(i, j \in \mathsf{V}[G])$. Next, the unitary adjacency matrix $\mathsf{U}[\mathbf{A}_w]$ of $\mathsf{L}(G')$ is calculated from Algorithm 2 and applied to propagate messages in graph with $\mathbf{h}_v^{(k)} = \gamma(\mathbf{h}_v^{(k-1)}, \sum_{u \in N_v} \mathsf{U}[\mathbf{A}_w]_{vu} \phi^{(k)}(\mathbf{h}_v^{(k-1)}, \mathbf{h}_u^{(k-1)}))$, $v \in V$ with any graph convolution operator. Here, $\phi^{(k)}(\cdot, \cdot)$ denotes a generic layer-k message function, and $\gamma(\cdot, \cdot)$ denotes the generic combine/update function that merges a node's previous representation with the aggregated message. In the method itself, GUMP constrains only the propagation operator $\mathsf{U}[\mathbf{A}_w]$; the learnable matrices inside the base GNN remain unconstrained. In the theory, we separately assume that the cumulative contribution of the weight matrices remains well conditioned, so that the analysis isolates the effect of repeated graph propagation. After L layers of unitary message passing, we obtain the node representations $\mathbf{H}^{(L)}$, which is later scattered to nodes of G with $\mathbf{H}_s^{(L)} = \mathsf{Scatter}(\mathbf{H}^{(L)}, G) \in \mathbb{R}^{n \times d'}$. Then, $\mathbf{H}_s^{(L)}$ are concatenated with $\mathbf{X}^{(0)}$ to obtain the final node representations $\mathbf{X}^{(L)} = [\mathbf{X}^{(0)}; \mathbf{H}_s^{(L)}] \in \mathbb{R}^{n \times (d+d')}$ of G. Finally, various graph learning tasks, e.g., graph and node classification, link prediction, and graph regression, are performed based on $\mathbf{X}^{(L)}$. In this paper, GUMP is a general one-hop message-passing mechanism for GNN. Therefore, depending on the convolution in line 6 of Algorithm 3, GNN with GUMP is named as [GNN type]-GUMP in Section 4, e.g., GCN-GUMP and GIN-GUMP have graph convolution and isomorphism operators in line 6 of Algorithm 3, respectively.

From the theoretical perspective, GUMP removes the graph-induced exponential instability in a source-averaged Jacobian measure. The theorem below is proved under the analysis approximation stated in Appendix C: ReLU gates are modeled by i.i.d. Bernoulli variables with depth-stable activation rate ρ, and the learnable weight matrices are assumed to remain well-conditioned so that they do not themselves introduce exponential growth or decay with depth. We use the source-averaged influence $\mathbb{E}_s\left[\left|(\mathbf{A}^L)_{is}\right|^2\right]$ because equation 3 isolates a source-target graph factor $(\mathbf{A}^L)_{is}$. Fixing a target node i and averaging $\left|(\mathbf{A}^L)_{is}\right|^2$ over a uniformly sampled source s measures how much graph-induced influence a typical source can exert on that target after L propagation steps. This avoids over-interpreting oscillatory single entries, is invariant to relabeling of source nodes, and coincides with the row energy of $\mathbf{A}^L$, which makes the comparison between unitary and vanilla propagation precise.

Theorem 1. *Let $\mathbf{A} \in \mathbb{C}^{N \times N}$ be the unitary propagation matrix used by GUMP, fix a target node $i \in [N]$, and sample the source node s uniformly from $[N]$. Then for every depth $L \geq 0$,*

$$\mathbb{E}_s\left[\left|(\mathbf{A}^L)_{is}\right|^2\right] = \frac{1}{N}.$$

Equivalently, the source-averaged influence is depth-invariant. Hence, under the expected-Jacobian approximation and the well-conditioned-weight assumption, GUMP does not suffer graph-induced exponential decay or explosion with depth in this aggregate measure.



(a) CrossedRing (b) Ring (c) CliquePath

Figure 3: The performance of GCN and GCN-GUMP on the CrossedRing, Ring, and CliquePath with different distances from source to target.

Theorem 1 does not claim that every entry $(\mathbf{A}^L)_{is}$ is constant in L, uniformly lower bounded, or that the full Jacobian is depth-invariant. For a fixed target-source pair (i, s), the different propagation components contributing to $(\mathbf{A}^L)_{is}$ can reinforce or partially cancel each other, so an individual entry may increase, decrease, or oscillate with depth. What the theorem guarantees is an aggregate invariant: for each target node i, the source-averaged quantity $\mathbb{E}_s[|(\mathbf{A}^L)_{is}|^2]$ remains exactly $1/N$ for all L, equivalently, the squared ℓ_2 norm of the i-th row of $\mathbf{A}^L$ stays constant under unitary propagation. By contrast, for vanilla normalized propagation with $\hat{\mathbf{A}}$, Theorem 5 shows that only the stationary component is preserved, while the non-stationary source-averaged influence decays as $\Theta(c^{2L}/N)$ whenever $\alpha_i > 0$. Thus, the key difference is that GUMP removes the graph-induced exponential decay present in vanilla propagation at the level of this source-averaged aggregate measure.

4 Experiments

In this section, we perform experiments to evaluate GUMP on graph learning tasks. All experiments are implemented by PyTorch Geometric (Fey & Lenssen, 2019) and conducted on NVIDIA RTX 4090 GPUs and AMD EPYC 7763 CPUs.

4.1 Experiments on Synthetic Dataset

Setup In this section, we conduct experiments on synthetic datasets, i.e., CrossedRing, Ring, and CliquePath, in Di Giovanni et al. (2023) to test GUMP. The performance is evaluated on the distances from source to target in the range of 4 to 28. This experiment is designed to test the ability of GUMP to learn long-range interactions when increasing the model layers. In the experiments, we compare GCN-GUMP and GCN. The layer L of GCN-GUMP and GCN is appropriately set up according to the distance d between source and target in the synthetic datasets (i.e., $L = \lfloor d/2 \rfloor + 1$), such that the long-range interactions can be captured by GNN. We set the hidden dimension to be 32 for both GCN-GUMP and GCN. The hyperparameters of GCN-GUMP for synthetic datasets are in appendix.

Results We plot the average results from three random seeds of GCN-GUMP and GCN experiments in Figure 3. For two easier datasets, i.e., CrossedRing and Ring, GUMP achieves 100% accuracy when the distance ranges from 4 to 28. For the challenging CliquePath dataset, GCN-GUMP's performance

Table 1: Graph classification accuracy on the TUDataset. **First** and second best results are bold and underlined, respectively.

Classes	Methods	Mutag	Proteins	Enzymes	NCI1	NCI109
–	GIN	77.70±3.60	70.80±0.83	33.80±1.12	75.65±0.49	74.93±0.46
	GIN (+layer)	69.80±2.75	68.71±0.96	25.92±1.07	73.49±0.46	72.47±0.53
Rewiring (GIN)	DIGL	79.80±2.08	70.71±0.67	35.74±1.20	79.37±0.43	76.88±0.39
	SDRF	78.40±2.80	69.81±0.79	35.82±1.09	74.55±0.54	73.89±0.43
	FoSR	78.00±2.22	75.11±0.82	29.20±1.38	70.15±0.47	69.93±0.45
	GTR	77.60±2.84	73.13±0.69	30.57±1.42	75.45±0.44	75.28±0.42
	ELDANADD	82.16±0.03	70.53±0.86	26.36±0.01	–	–
Diffusion	ADGN	81.39±1.81	73.81±0.80	28.78±1.25	76.15±0.42	74.31±0.44
	GRAND	77.94±1.73	73.24±0.94	24.13±1.05	68.51±0.48	67.26±0.46
Transformer	GPS	70.12±1.68	69.32±0.86	30.43±1.76	61.32±0.51	61.01±0.22
–	Ortho-GConv	71.78±2.52	63.80±0.98	18.30±1.13	69.92±0.60	68.91±0.50
–	Unitary GNN	78.00±1.96	71.57±0.60	43.67±1.45	78.71±0.42	78.10±0.34
K-hop MP	GRIT	80.76±2.18	73.71±0.89	35.22±1.17	72.21±0.46	71.68±0.44
Ours	GIN-GUMP	**86.72±1.53**	**75.43±0.70**	**48.43±1.24**	**81.25±0.37**	**78.45±0.44**

deteriorates to random guessing at a distance of 28. The results show that GUMP can help capture the long-range interactions in graph learning tasks. We compare with more baselines in appendix.

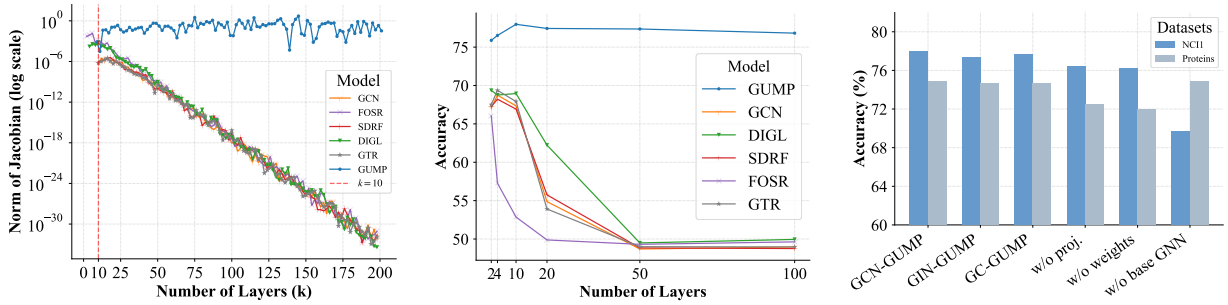
4.2 Experiments on the TUDataset

Datasets We select five datasets, i.e., Mutag, Proteins, Enzymes, NCI1, and NCI109 from the TUDataset (Morris et al., 2020). We chose these datasets because they consist of chemistry or biological graphs with complex or long-range interactions, e.g., the atoms far apart may be closer in space. The statistics of these datasets are in appendix.

Baselines Baselines include various rewiring methods, i.e., DIGL (Gasteiger et al., 2019), SDRF (Topping et al., 2022), FoSR (Karhadkar et al., 2023), GTR (Black et al., 2023), and ELDANADD (Jamadandi et al., 2024), diffusion methods, i.e., ADGN (Tönshoff et al., 2023) and GRAND (Chamberlain et al., 2021), transformers method GPS (Rampášek et al., 2022), GNN with orthogonal parameters Ortho-GConv (Guo et al., 2022), k-hop message passing GRIT (Di Giovanni et al., 2023), and Unitary GNN (Kiani et al., 2024). The baselines' settings follow Karhadkar et al. (2023). For the Unitary GNN entries on these five TU datasets, we reproduce the results by running the released code of Kiani et al. (2024) under the TU evaluation protocol.

Experimental Details To evaluate each method, we initially designate a test set comprising 10% of the graphs and a development set encompassing the remaining 90% of the graphs. The accuracies of each configuration are determined through 100 random train/validation splits of the development set, with 80% for training and 10% for validation. During the training phase, a stopping patience of 100 epochs is employed based on validation loss. Subsequently, for the test results, we report 95% confidence intervals for the best validation accuracy observed across the 100 runs. The number of layers for GUMP is manually tuned because the long-range interactions in a graph can only be captured by increasing its layers. The detailed hyperparameters of GUMP for different datasets are presented in appendix.

Results The results of GUMP on TUDataset are shown in Table 1. GUMP achieves superior performance across all five datasets, with GIN-GUMP obtaining the best results, demonstrating that the unitary message passing mechanism effectively learns complex interactions in graphs. Moreover, since GUMP usually has more layers than baselines in these datasets, experiments show that GCNs with more layers exhibit degraded



(a) Jacobian norm versus layers on NCI1 (b) Accuracy versus layers on NCI1 (c) Ablation on different components

Figure 4: Model analysis. GCN, GIN, and GUMP in (c) represent the convolution of GUMP. The base GNN of GUMP is GCN. "w/o proj" removes unitary projection in GUMP. "w/o weights" removes weighted adjacency matrix and unitary projection in GUMP. "w/o base GNN" removes base GNN in GUMP.

performance on all datasets, showing that improvement of GUMP does not come from increasing expressivity with more GNN layers.

4.3 Experiments on LRGB

Here, we conduct experiments on the Long Range Graph Benchmark (LRGB) (Dwivedi et al., 2022), which is a set of GNN benchmarks involving long-range interactions. Two datasets are selected from LRGB for comparison, i.e., Peptides-func and Peptides-struct. The statistics of these datasets are shown in appendix.

Table 2: Results of Peptides-func and Peptides-struct. **Bold** and underline denote best and second-best results.

Methods	Peptides-func Test AP ↑	Peptides-struct Test MAE ↓
GCN	0.6860±0.0050	0.2460±0.0007
SDRF	0.6874±0.0032	0.2453±0.0018
FoSR	0.6878±0.0030	0.2461±0.0016
GTR	0.6740±0.0033	0.2509±0.0013
LASER	0.6440±0.0010	0.3043±0.0019
GRAND	0.5789±0.0062	0.3418±0.0015
ADGN	0.5975±0.0044	0.2874±0.0021
GPS	0.6534±0.0091	0.2509±0.0014
Exphormer	0.6527±0.0043	0.2481±0.0007
GRIT	0.6988±0.0082	0.2460±0.0012
Graph ViT	0.6942±0.0075	0.2449±0.0016
G-MLPMixer	0.6921±0.0054	0.2475±0.0015
Drew-GCN	0.6996±0.0076	0.2781±0.0028
ProxyAdd+GCN	0.6789±0.0002	0.2465±0.0004
k-GCN-SSM	0.6902±0.0138	0.2898±0.0324
UniGCN	0.7072±0.0035	**0.2425±0.0009**
GUMP	**0.7088±0.0026**	0.2438±0.0014

The LRGB experiments follow the protocol of Tönshoff et al. (2023). We compare GUMP with reproduced long-range graph baselines, including SDRF, FoSR, GTR, LASER (Barbero et al., 2023), GRAND, ADGN, GPS, Exphormer (Shirzad et al., 2023), GRIT, Graph ViT (He et al., 2023), G-MLPMixer (He et al., 2023), Drew-GCN (Gutteridge et al., 2023), ProxyAdd+GCN (Jamadandi et al., 2024), and k-GCN-SSM (Arroyo et al., 2025). For GUMP, the base one-hop message-passing operator is GCN; the appendix reports the tuned number of GUMP layers, base-GCN layers, hidden dimension, optimizer, learning rates, weight decay, dropout, batch size, scheduler, and training epochs for each Peptides dataset. We additionally include the UniGCN results reported by Kiani et al. (2024) to compare against a recent unitary-convolution baseline, while noting that those numbers are reported under that paper's setup rather than reproduced in our protocol. The hyperparameters of GUMP and more results for LRGB are presented in appendix.

The results of LRGB are shown in Table 2. GUMP achieves the best result on Peptides-func and the second-best result on Peptides-struct, indicating that its unitary propagation operator improves one-hop message passing on long-range graph tasks while not being uniformly best across both datasets. On Peptides-func, GUMP obtains a test AP of 0.7088, surpassing Drew-GCN at 0.6996, GRIT at 0.6988, and the reported UniGCN result at 0.7072. On Peptides-struct, GUMP obtains a test MAE of 0.2438, improving over Graph ViT at 0.2449 and GCN at 0.2460, while the reported UniGCN result from Kiani et al. (2024) is slightly better with a test MAE of 0.2425. Overall, these results show that GUMP is competitive with recent

unitary-convolution methods on LRGB while using a different sparse, input-dependent unitary message-passing construction.

4.4 Model Analysis

In this section, we perform more experiments to analyze GUMP from three aspects, i.e., Jacobian, number of layers, and ablation studies.

Jacobian We visualize the Jacobian norm of GUMP and several baselines involving graph-structured propagation. We choose a pair of nodes with a distance of ten from NCI1 and calculate the spectral norm of the Jacobian for GUMP and baselines with base GNN as GCN. Figure 4(a) shows the Jacobian norm in log scale. As depth increases, the Jacobian norm of GUMP remains stable while those of the baselines decay rapidly. While Theorems 1 and 5 are stated for a source-averaged propagation quantity, the operator-norm trend in Figure 4(a) is directionally consistent with that theory: unitary propagation avoids graph-induced exponential decay, whereas normalized propagation suppresses non-stationary components with depth. Secondly, the norm of the Jacobian varies for different baselines. For example, DIGL has a large norm of Jacobian when the number of layers is smaller than 50. Therefore, DIGL performs better than other rewiring methods in Table 1.

Deep GNN The number of GNN layers indicates the ability of GNNs to capture complex and long-range interactions. We increase the number of layers of GUMP and baselines to see how their performances change. This experiment is conducted in NCI1 with base GNN as GCN and the number of layers in the range of 2 to 100. The results in Figure 4(b) demonstrate that the performance of GUMP increases from 75.88% to 77.97% when increasing the number of layers from 2 to 10. However, the performance of baselines decreases a lot when increasing the number of layers. For example, the performance of FoSR decreases from 66.06% to 52.86% when the number of layers increases from 2 to 10. The performance of baselines changes drastically as the layers increase, while the performance of GUMP is more stable. The results show GUMP can be deeper than previous methods, thus learning more complex interactions in graphs.

Ablation Studies Lastly, we conduct ablation studies to analyze GUMP in Figure 4(c). We first replace the convolution of GUMP with GIN and GraphConv (Morris et al., 2019), showing that the choice of convolution can impact the performance of GUMP. Then, we remove unitary projection (i.e., message passing with $\mathbf{A}_w$) and weighted adjacency matrix (i.e., message passing with $\tilde{\mathbf{A}}[\mathsf{L}(G')]$) in GUMP and the results show that their performances decrease, indicating the importance of GUMP. Finally, we remove the base GNN in GUMP and the results show that the performance of GUMP varies on different datasets, i.e., the performance on NCI1 decreases, while the performance on Proteins does not decrease. This phenomenon is expected because the quality of the unitary adjacency matrix depends on the representations of nodes in the line graph (see equation 1).

In model analysis, GUMP demonstrates the most depth-stable Jacobian behavior among the compared methods, and greater stability with increased layers than prior methods, highlighting the role of its propagation design in achieving superior performance. Appendix D also contains the dataset statistics, hyperparameters, additional comparisons on other datasets, Jacobian visualizations, and training-time measurements that support the main experimental claims.

5 Discussions and Limitations

In this paper, we propose Graph Unitary Message Passing (GUMP), which propagates messages using unitary adjacency matrices to improve graph learning efficiency, which shows superior performance across various graph learning tasks. We discuss below the limitations and future works of GUMP:

Information Loss Here we discuss the potential information loss in applying GUMP. For undirected, unweighted graphs, GUMP's transformation is lossless. For weighted graphs, edge weights are encoded as features following R-GNNs (Battaglia et al., 2018). For directed graphs, directionality is preserved by encod-

ing it as edge features and filtering invalid edges in post-processing (Algorithm 3, step 8). These adaptations ensure GUMP maintains the essential properties of various graphs, enabling its broad application.

Implications for Large Graph Models GUMP represents a promising step toward applying unitary constraints to graph learning. Our work demonstrates that unitary message passing significantly improves training stability and model performance on moderate-scale graphs. The graph transformation we propose could serve as foundational components for scaling up graph models. Future large graph models could leverage our transformation to convert arbitrary graphs into forms amenable to unitary processing, while employing initialization schemes that enable approximate unitarity rather than exact projection. This would enable large graph models that maintain high learning efficiency. The largest remaining theoretical gap is the finite-step Newton-Schulz approximation. Our main stability analysis is stated for the exact polar factor $\mathsf{U}[\mathbf{A}_w]$, which is unitary, whereas Algorithm 2 uses only K Newton-Schulz iterations and therefore returns an approximately unitary operator. Still, finite-step Newton-Schulz is expected to help because its polynomial update regulates the singular values of each block and pushes them toward one, improving conditioning and approximate norm preservation even before exact convergence. The block-support argument still explains why the finite iterates do not create inadmissible transitions, but a complete depth-stability theory should quantify how the residual non-unitarity after finite K propagates through many GNN layers and how this error depends on the block condition numbers and graph degrees.

Another practical direction is to explore faster or more accurate polar-factor approximations than the current Newton-Schulz iteration. In the current undirected formulation, the projection cost is $\mathcal{O}\big(K \sum_v d(v)^3\big)$, so reducing the dependence on high-degree vertices is an important direction for scaling GUMP.

References

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. *arXiv preprint arXiv:2006.05205*, 2020.

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International conference on machine learning*, pp. 1120–1128. PMLR, 2016.

Adrián Arnaiz-Rodríguez, Ahmed Begga, Francisco Escolano, and Nuria M Oliver. Diffwire: Inductive graph rewiring via the lovász bound. In *The First Learning on Graphs Conference*, 2022. URL `https://openreview.net/forum?id=IXvfIex0mX6f`.

Álvaro Arroyo, Alessio Gravina, Benjamin Gutteridge, Federico Barbero, Claudio Gallicchio, Xiaowen Dong, Michael Bronstein, and Pierre Vandergheynst. On vanishing gradients, over-smoothing, and over-squashing in gnns: Bridging recurrent and graph learning. *arXiv preprint arXiv:2502.10818*, 2025.

David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *International conference on machine learning*, pp. 342–350. PMLR, 2017.

Pradeep Kr Banerjee, Kedar Karhadkar, Yu Guang Wang, Uri Alon, and Guido Montúfar. Oversquashing in gnns through the lens of information contraction and graph expansion. In *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–8. IEEE, 2022.

Jørgen Bang-Jensen and Gregory Z Gutin. *Digraphs: theory, algorithms and applications*. Springer Science & Business Media, 2008.

Federico Barbero, Ameya Velingker, Amin Saberi, Michael Bronstein, and Francesco Di Giovanni. Locality-aware graph-rewiring in gnns. *arXiv preprint arXiv:2310.01668*, 2023.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Åke Björck and Clazett Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971.

Mitchell Black, Zhengchao Wan, Amir Nayyeri, and Yusu Wang. Understanding oversquashing in gnns through the lens of effective resistance. In *International Conference on Machine Learning*, pp. 2528–2547. PMLR, 2023.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2021.

Lei Cai, Jundong Li, Jie Wang, and Shuiwang Ji. Line graph neural networks for link prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5103–5113, 2022. doi: 10.1109/TPAMI.2021.3080635.

Ben Chamberlain, James Rowbottom, Maria I Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. Grand: Graph neural diffusion. In *International Conference on Machine Learning*, pp. 1407–1418. PMLR, 2021.

Benjamin Paul Chamberlain, Sergey Shirobokov, Emanuele Rossi, Fabrizio Frasca, Thomas Markovich, Nils Yannick Hammerla, Michael M. Bronstein, and Max Hansmire. Graph neural networks for link prediction with subgraph sketching. In *International Conference on Learning Representations*, 2023.

Jianwen Chen, Jun He, Changyou Chen, and Shuiwang Wang. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 12492–12499, 2020a.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International conference on machine learning*, pp. 1725–1735. PMLR, 2020b.

Soham De, Samuel L Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.

Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Lio, and Michael M Bronstein. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In *International Conference on Machine Learning*, pp. 7865–7885. PMLR, 2023.

Vijay Prakash Dwivedi, Ladislav Rampášek, Michael Galkin, Ali Parviz, Guy Wolf, Anh Tuan Luu, and Dominique Beaini. Long range graph benchmark. *Advances in Neural Information Processing Systems*, 35:22326–22340, 2022.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Hongyang Gao and Shuiwang Ji. Graph u-nets. In *international conference on machine learning*, pp. 2083–2092. PMLR, 2019.

Johannes Gasteiger, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. *Advances in neural information processing systems*, 32, 2019.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Kai Guo, Kaixiong Zhou, Xia Hu, Yu Li, Yi Chang, and Xin Wang. Orthogonal graph neural networks. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 3996–4004, 2022.

Benjamin Gutteridge, Xiaowen Dong, Michael M Bronstein, and Francesco Di Giovanni. Drew: Dynamically rewired message passing with delay. In *International Conference on Machine Learning*, pp. 12252–12267. PMLR, 2023.

Xiaoxin He, Bryan Hooi, Thomas Laurent, Adam Perold, Yann LeCun, and Xavier Bresson. A generalization of vit/mlp-mixer to graphs. In *International conference on machine learning*, pp. 12724–12745. PMLR, 2023.

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled cayley transform. In *International Conference on Machine Learning*, pp. 1969–1978. PMLR, 2018.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22118–22133, 2020.

Adarsh Jamadandi, Celia Rubio-Madrigal, and Rebekka Burkholz. Spectral graph pruning against over-squashing and over-smoothing. *Advances in Neural Information Processing Systems*, 37:10348–10379, 2024.

Xiaodong Jiang, Pengsheng Ji, and Sheng Li. Censnet: Convolution with edge-node switching in graph neural networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 2656–2662, 2019. doi: 10.24963/ijcai.2019/369.

Li Jing, Yichen Shen, Tena Dubcek, John Peurifoy, Scott Skirlo, Yann LeCun, Max Tegmark, and Marin Soljačić. Tunable efficient unitary neural networks (eunn) and their application to rnns. In *International Conference on Machine Learning*, pp. 1733–1741. PMLR, 2017.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.

Kedar Karhadkar, Pradeep Banerjee, and Guido Montufar. Fosr: First-order spectral rewiring for addressing oversquashing in gnns. In *International Conference on Learning Representations*, 2023.

Joseph B Keller. Closest unitary, orthogonal and hermitian operators to a given operator. *Mathematics Magazine*, 48(4):192–197, 1975.

Bobak Kiani, Randall Balestriero, Yann LeCun, and Seth Lloyd. projunn: Efficient method for training deep networks with unitary matrices. *Advances in Neural Information Processing Systems*, 35:14448–14463, 2022.

Bobak T. Kiani, Lukas Fesser, and Melanie Weber. Unitary convolutions for learning on graphs and groups. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136922–136961, 2024. doi: 10.52202/079017-4351.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970.

Mario Lezcano-Casado and David Martınez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *International Conference on Machine Learning*, pp. 3794–3803. PMLR, 2019.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, volume 32, pp. 3538–3545, 2018.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deepgcns: Can gcns go as deep as cnns? In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9267–9276, 2019a.

Qiyang Li, Saminul Haque, Cem Anil, James Lucas, Roger B Grosse, and Jörn-Henrik Jacobsen. Preventing gradient attenuation in lipschitz constrained convolutional networks. *Advances in neural information processing systems*, 32, 2019b.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Federico Monti, Oleksandr Shchur, Aleksandar Bojchevski, Or Litany, Stephan Günnemann, and Michael M. Bronstein. Dual-primal graph convolutional networks. *arXiv preprint arXiv:1806.00770*, 2018.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *AAAI conference on artificial intelligence*, volume 33, pp. 4602–4609, 2019.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*, 2020. URL `www.graphlearning.io`.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947*, 2019.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*, pp. 26670–26698. PMLR, 2023.

Liming Pan, Cheng Shi, and Ivan Dokmanić. Neural link prediction with walk pooling. In *International Conference on Learning Representations*, 2022.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. Pmlr, 2013.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Hanie Sedghi, Vineet Gupta, and Philip M Long. The singular values of convolutional layers. *arXiv preprint arXiv:1805.10408*, 2018.

Simone Severini. On the digraph of a unitary matrix. *SIAM Journal on Matrix Analysis and Applications*, 25(1):295–300, 2003.

Hamed Shirzad, Ameya Velingker, Balaji Venkatachalam, Danica J Sutherland, and Ali Kemal Sinop. Exphormer: Sparse transformers for graphs. In *International Conference on Machine Learning*, pp. 31613–31632. PMLR, 2023.

Sahil Singla and Soheil Feizi. Skew orthogonal convolutions. In *International Conference on Machine Learning*, pp. 9756–9766. PMLR, 2021.

Jan Tönshoff, Martin Ritzert, Eran Rosenbluth, and Martin Grohe. Where did the gap go? reassessing the long-range graph benchmark. *arXiv preprint arXiv:2309.00367*, 2023.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*, 2022.

Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. *arXiv preprint arXiv:2104.07167*, 2021.

Ziming Wang, Jun Chen, and Haopeng Chen. Egat: Edge-featured graph attention network. In *Artificial Neural Networks and Machine Learning – ICANN 2021*, Lecture Notes in Computer Science, pp. 253–264. Springer, 2021. doi: 10.1007/978-3-030-86362-3_21.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *International conference on machine learning*, pp. 5453–5462. PMLR, 2018.

Yulei Yang and Dongsheng Li. Nenn: Incorporate node and edge features in graph neural networks. In *Proceedings of The 12th Asian Conference on Machine Learning*, volume 129 of *Proceedings of Machine Learning Research*, pp. 593–608. PMLR, 2020.

Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yuyang Zhang, Xu Shen, Yu Xie, Ka-Chun Wong, Weidun Xie, and Chengbin Peng. Directed link prediction using gnn with local and global feature fusion. *arXiv preprint arXiv:2506.20235*, 2025.

Zehua Zhang, Shilin Sun, Guixiang Ma, and Caiming Zhong. Line graph contrastive learning for link prediction. *Pattern Recognition*, 140:109537, 2023. doi: 10.1016/j.patcog.2023.109537.

Lingxiao Zhao and Leman Akoglu. Pairnorm: Tackling oversmoothing in gnns. *arXiv preprint arXiv:1909.12223*, 2019.

A Related work

A.1 GNN and its Training Instability

Graph neural networks (GNNs) (Kipf & Welling, 2016; Gilmer et al., 2017) learn node representations through message passing, but deep propagation is often limited by oversquashing and oversmoothing. Oversquashing was highlighted by Alon & Yahav (2020), and Topping et al. (2022) connected it to Jacobian-based sensitivity and proposed curvature-inspired rewiring. Subsequent work has alleviated oversquashing by improving spectral gap or connectivity properties, including expander-style rewiring (Banerjee et al., 2022), FoSR (Karhadkar et al., 2023), DiffWire (Arnaiz-Rodríguez et al., 2022), total-resistance-based rewiring (Black et al., 2023), commute-time analysis (Di Giovanni et al., 2023), locality-preserving rewiring (Barbero et al., 2023), and multi-hop architectures such as Drew (Gutteridge et al., 2023).

Oversmoothing refers to the tendency of node representations to become indistinguishable as depth increases (Li et al., 2018; Oono & Suzuki, 2019). This effect has been studied empirically and theoretically (Kipf & Welling, 2016; Wu et al., 2020), with remedies including structure modification such as DropEdge and diffusion-based sparsification (Rong et al., 2019; Gasteiger et al., 2019), normalization methods such as PairNorm (Zhao & Akoglu, 2019), and architectural changes such as residual connections, Jumping Knowledge, and hierarchical pooling (Li et al., 2019a; Chen et al., 2020b; Xu et al., 2018; Gao & Ji, 2019). Recent work also studies oversquashing and oversmoothing jointly (Jamadandi et al., 2024; Arroyo et al., 2025).

A.2 Line graphs and edge-centric GNNs.

Graph neural networks on line graphs and related edge-centric formulations have been studied to better align representation learning with edge-level tasks such as link prediction. In a line graph, each edge in the original graph is represented as a node, allowing message passing to operate directly over links rather than first learning endpoint embeddings and then applying a separate decoder. Early dual or edge-node co-embedding architectures, such as Dual-Primal Graph Convolutional Networks (Monti et al., 2018) and CensNet (Jiang et al., 2019), propagate information over both vertex- and edge-level structures, with CensNet explicitly using the line graph to switch the roles of nodes and edges. Edge-aware architectures such as NENN (Yang & Li, 2020) and EGAT (Wang et al., 2021) further model interactions between node and edge features, although they are not primarily designed as line-graph link-prediction methods. For link prediction, SEAL (Zhang & Chen, 2018) established the effectiveness of extracting enclosing subgraphs around candidate links and learning structural heuristics with a GNN, but it treats each candidate as a graph-classification instance. LGLP (Cai et al., 2022) instead directly transforms the enclosing subgraph of a candidate link into a line graph, reducing link prediction to node classification on the corresponding edge-node. LGCL (Zhang et al., 2023) extends this direction by introducing contrastive learning between the original subgraph view and the line-graph view, demonstrating that line-graph representations capture complementary edge-neighborhood information useful for predicting missing links. Recent preprint work has also explored directed link prediction by constructing directed line graphs and fusing local and global features (Zhang et al., 2025), suggesting a possible extension of line-graph GNNs beyond undirected settings. In parallel, strong link-prediction models such as WalkPooling (Pan et al., 2022) and ELPH/BUDDY (Chamberlain et al., 2023) learn subgraph-aware or edge-centric representations without explicitly materializing the full line graph, offering more scalable alternatives when line-graph construction becomes expensive around high-degree nodes. Overall, existing work suggests that line-graph GNNs are attractive for link prediction because they make edges the primary objects of message passing, but their practical use must balance task alignment against scalability, directionality, multigraph handling, and the cost of constructing dense edge-to-edge neighborhoods.

A.3 Unitarity in Deep Learning

Unitarity has been widely used to stabilize deep models. In recurrent networks, unitary and orthogonal parameterizations such as uRNN, EUNN, scoRNN, expRNN, projUNN, and LRU improve long-range signal propagation by controlling the spectrum of the transition operator (Arjovsky et al., 2016; Jing et al., 2017; Helfrich et al., 2018; Lezcano-Casado & Martınez-Rubio, 2019; Kiani et al., 2022; Orvieto et al., 2023).

Related recurrent architectures, including RWKV and Mamba, share the broader goal of maintaining stable long-range dynamics (Peng et al., 2023; Gu & Dao, 2023).

Beyond RNNs, orthogonality or unitarity has also been explored in convolutional and graph models. Unitary CNNs constrain convolutional filters through structured parameterizations or Cayley/Lie-algebra methods (Sedghi et al., 2018; Li et al., 2019b; Singla & Feizi, 2021; Trockman & Kolter, 2021), while Ortho-GConv imposes orthogonality on the feature transformation matrix in GNNs (Guo et al., 2022). Closely related to our work, Kiani et al. (2024) construct unitary graph and group convolutions through matrix exponentials such as $\exp(i\mathbf{A})$, yielding norm-preserving propagation operators with strong empirical performance. GUMP is complementary in both construction and scope: rather than applying a matrix-exponential unitary convolution on the original graph, we transform the input graph into an Eulerian line-graph construction and compute an input-dependent unitary propagation matrix by polar projection while preserving the admissible sparse edge-to-edge transition structure. Orthogonal initialization and optimizers such as Muon further illustrate the usefulness of unitarity at the parameter level (Saxe et al., 2013; Balduzzi et al., 2017; Jordan et al., 2024). More broadly, our setting differs from parameter-level unitarity because the propagation operator is graph-dependent rather than a learned parameter, so enforcing unitarity requires graph transformation and projection rather than only parameter-level constraints.

B Preliminaries

Definition 1 (Line graph)**.** *Given a graph G, its line graph $\mathsf{L}(G)$ is a graph such that*

- *each vertex of $\mathsf{L}(G)$ represents an edge of G, i.e., $\mathsf{V}[\mathsf{L}(G)] = \mathsf{E}[G]$;*
- *two vertices of $\mathsf{L}(G)$ are adjacent if and only if their corresponding edges share a common endpoint in G, i.e., $\mathsf{E}[\mathsf{L}(G)] = \{((i,j),(j,k)) \in \mathsf{V}[\mathsf{L}(G)] \times \mathsf{V}[\mathsf{L}(G)] \mid (i,j),(j,k) \in \mathsf{E}[G]\}$.*

Definition 2 (Eulerian graph)**.** *An Eulerian graph G is a graph containing an Eulerian cycle, i.e., there is a trail in G that starts and ends on the same vertex and visits every edge exactly once.*

Definition 3 (Permutation matrix)**.** *A permutation matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ is a square binary matrix that has exactly one entry of 1 in each row and each column with all other entries 0.*

Theorem 2. *Every permutation matrix is orthogonal, i.e., if $\mathbf{P}$ is a permutation matrix, $\mathbf{P}^\top\mathbf{P} = \mathbf{P}\mathbf{P}^\top = \mathbf{I}$.*

C Proof

C.1 Proof of Proposition 1

Proposition 1 is proved based on the following theorem and lemma. Theorem 3 is a direct result from Theorem 3 in Severini (2003). Lemma 1 is a well-known result in graph theory, which can be found in Theorem 1.7.2 of Bang-Jensen & Gutin (2008).

Theorem 3 (Existence of unitary adjacency matrix)**.** *Let G be a single-connected digraph. Its line graph $\mathsf{L}(G)$ (Definition 1) is the digraph of a unitary matrix if and only if G is Eulerian (Definition 2).*

Lemma 1 (A special Eulerian graph)**.** *A digraph graph is Eulerian if and only if it is connected and the in-degree and out-degree are equal at each vertex.*

Proof of Proposition 1. In Algorithm 1, the undirected edges in G are split into two directed edges in G'. Therefore, the in-degree and out-degree of each vertex in G' are equal. We first consider the case where G is connected. Replacing each undirected edge by the pair of directed edges (i,j) and (j,i) preserves connectivity of the underlying graph, so G' is connected as well. By Lemma 1, G' is Eulerian, and Theorem 3 yields a unitary adjacency matrix $\mathbf{U}$ such that $\mathsf{S}[\mathbf{U}] = \mathbf{A}[\mathsf{L}(G')]$.

Now suppose that G is disconnected. Let $G_1, \ldots, G_m$ be the connected components of G that contain at least one edge, and let $G'_1, \ldots, G'_m$ be the corresponding connected components of G' after the edge-splitting step. The same argument as above shows that each G'_r is connected and satisfies $d_{\text{in}}(v) = d_{\text{out}}(v)$ at every

vertex, so Lemma 1 implies that each G'_r is Eulerian. Theorem 3 then gives a unitary adjacency matrix $\mathbf{U}_r$ for each $\mathsf{L}(G'_r)$. Since the line graph of a disjoint union is the disjoint union of the corresponding line graphs, $\mathsf{L}(G')$ is block diagonal up to vertex ordering with blocks $\mathsf{L}(G'_1), \ldots, \mathsf{L}(G'_m)$; isolated vertices of G contribute no vertices to $\mathsf{L}(G')$, but their node features remain available through GUMP's skip/concatenation path. Hence $\mathbf{U} := \mathsf{diag}(\mathbf{U}_1, \ldots, \mathbf{U}_m)$ is unitary and satisfies $\mathsf{S}[\mathbf{U}] = \mathbf{A}[\mathsf{L}(G')]$. □

C.2 Proof of Proposition 2

Lemma 2. *For any unitary matrix* $\mathbf{U}$*, given two permutation matrices* $\mathbf{P}_1$ *and* $\mathbf{P}_2$*,* $\mathbf{P}_1\mathbf{U}\mathbf{P}_2^\top$ *is also unitary.*

Proof. Let $\hat{\mathbf{U}} = \mathbf{P}_1\mathbf{U}\mathbf{P}_2^\top$. Then, we have

$$\hat{\mathbf{U}}\hat{\mathbf{U}}^\dagger = \mathbf{P}_1\mathbf{U}\mathbf{P}_2^\top\mathbf{P}_2\mathbf{U}^\dagger\mathbf{P}_1^\top = \mathbf{P}_1\mathbf{U}\mathbf{U}^\dagger\mathbf{P}_1^\top = \mathbf{P}_1\mathbf{P}_1^\top = \mathbf{I},$$
$$\hat{\mathbf{U}}^\dagger\hat{\mathbf{U}} = \mathbf{P}_2\mathbf{U}^\dagger\mathbf{P}_1^\top\mathbf{P}_1\mathbf{U}\mathbf{P}_2^\top = \mathbf{P}_2\mathbf{U}^\dagger\mathbf{U}\mathbf{P}_2^\top = \mathbf{P}_2\mathbf{P}_2^\top = \mathbf{I}$$

which proves $\mathbf{P}_1\mathbf{U}\mathbf{P}_2^\top$ is unitary. □

Proof of Proposition 2. Since $\mathbf{A}_w$ is full-rank, the polar factor is given by $\mathsf{U}[\mathbf{A}_w] = \mathbf{A}_w(\mathbf{A}_w^\dagger\mathbf{A}_w)^{-1/2}$. Let $\hat{\mathbf{A}}_w = \mathbf{P}_1\mathbf{A}_w\mathbf{P}_2^\top$. Then

$$\begin{aligned}\hat{\mathbf{A}}_w^\dagger\hat{\mathbf{A}}_w &= \mathbf{P}_2\mathbf{A}_w^\dagger\mathbf{P}_1^\top\mathbf{P}_1\mathbf{A}_w\mathbf{P}_2^\top \\ &= \mathbf{P}_2(\mathbf{A}_w^\dagger\mathbf{A}_w)\mathbf{P}_2^\top.\end{aligned}$$

Because permutation similarity preserves matrix functions,

$$(\hat{\mathbf{A}}_w^\dagger\hat{\mathbf{A}}_w)^{-1/2} = \mathbf{P}_2(\mathbf{A}_w^\dagger\mathbf{A}_w)^{-1/2}\mathbf{P}_2^\top.$$

Therefore,

$$\begin{aligned}\mathsf{U}[\hat{\mathbf{A}}_w] &= \hat{\mathbf{A}}_w(\hat{\mathbf{A}}_w^\dagger\hat{\mathbf{A}}_w)^{-1/2} \\ &= \mathbf{P}_1\mathbf{A}_w\mathbf{P}_2^\top\mathbf{P}_2(\mathbf{A}_w^\dagger\mathbf{A}_w)^{-1/2}\mathbf{P}_2^\top \\ &= \mathbf{P}_1\mathbf{A}_w(\mathbf{A}_w^\dagger\mathbf{A}_w)^{-1/2}\mathbf{P}_2^\top \\ &= \mathbf{P}_1\mathsf{U}[\mathbf{A}_w]\mathbf{P}_2^\top,\end{aligned}$$

which proves the proposition. □

C.3 Proof of Proposition 3

Proof. Let $G' = (V, E')$. Index the rows and columns of $\mathbf{A}_w$ and $\mathbf{A}_\mathsf{L}$ by directed edges of E'. By Definition 1, for two directed edges $e = (u, v)$ and $f = (x, y)$, the entry $(\mathbf{A}_\mathsf{L})_{ef}$ is one exactly when $v = x$. Thus $(\mathbf{A}_w)_{ef}$ can be nonzero only when $e \in E_{\text{in}}(v)$ and $f \in E_{\text{out}}(v)$ for the same vertex $v \in V$.

Let $\mathbf{P}_{\text{in}}$ be the permutation that groups rows by the sets $E_{\text{in}}(v)$ and let $\mathbf{P}_{\text{out}}$ be the permutation that groups columns by the sets $E_{\text{out}}(v)$. Under these permutations, rows associated with $E_{\text{in}}(v)$ interact only with columns associated with $E_{\text{out}}(v)$, so

$$\mathbf{D} := \mathbf{P}_{\text{in}}\mathbf{A}_w\mathbf{P}_{\text{out}}^\top = \mathsf{diag}\big(\mathbf{B}_v\big)_{v\in V}.$$

Each block $\mathbf{B}_v$ contains all transitions $(u, v) \to (v, w)$ in the line graph. Because Algorithm 1 inserts both orientations of every undirected edge, $|E_{\text{in}}(v)| = |E_{\text{out}}(v)|$, hence $\mathbf{B}_v$ is square.

Since $\mathbf{D}$ is block diagonal,

$$\mathbf{D}^\dagger\mathbf{D} = \mathsf{diag}\big(\mathbf{B}_v^\dagger\mathbf{B}_v\big)_{v\in V}.$$

Matrix functions preserve block diagonality, so under the full-rank assumption,

$$(\mathbf{D}^\dagger\mathbf{D})^{-1/2} = \mathsf{diag}\big((\mathbf{B}_v^\dagger\mathbf{B}_v)^{-1/2}\big)_{v\in V}.$$

Therefore,

$$\begin{aligned}\mathsf{U}[\mathbf{D}] &= \mathbf{D}(\mathbf{D}^\dagger\mathbf{D})^{-1/2}\\ &= \mathsf{diag}\big(\mathbf{B}_v(\mathbf{B}_v^\dagger\mathbf{B}_v)^{-1/2}\big)_{v\in V}\\ &= \mathsf{diag}\big(\mathsf{U}[\mathbf{B}_v]\big)_{v\in V}.\end{aligned}$$

By Proposition 2,

$$\mathsf{U}[\mathbf{D}] = \mathsf{U}[\mathbf{P}_{\text{in}}\mathbf{A}_w\mathbf{P}_{\text{out}}^\top] = \mathbf{P}_{\text{in}}\mathsf{U}[\mathbf{A}_w]\mathbf{P}_{\text{out}}^\top.$$

Hence $\mathsf{U}[\mathbf{A}_w] = \mathbf{P}_{\text{in}}^\top\mathsf{U}[\mathbf{D}]\mathbf{P}_{\text{out}}$ has the same row/column block structure as $\mathbf{A}_w$. Therefore, entries that are structurally zero because they lie outside the admissible line-graph blocks remain zero after projection, so

$$\mathsf{S}[\mathsf{U}[\mathbf{A}_w]] \subseteq \mathsf{S}[\mathbf{A}_\mathsf{L}].$$

If each block polar factor $\mathsf{U}[\mathbf{B}_v]$ is fully supported, this inclusion becomes equality. This condition is generic for full-rank blocks whose entries are produced by continuous learned weights: each entry of $\mathsf{U}[\mathbf{B}_v]$ is a real-analytic function of the entries of $\mathbf{B}_v$ on the full-rank domain, and it is not identically zero because one can choose a fully supported unitary block as $\mathbf{B}_v$. Hence the weights that make a specific projected entry exactly zero form a measure-zero set.

For the Newton-Schulz implementation, let $\mathbf{X}_{v,t}$ denote the t-th iterate for block $\mathbf{B}_v$, initialized as $\mathbf{X}_{v,0} = \mathbf{B}_v/\|\mathbf{B}_v\|_F$. If $\mathbf{X}_{v,t}$ is blockwise supported, then $\mathbf{X}_{v,t}^\top\mathbf{X}_{v,t}$ and the polynomial update in Algorithm 2 remain within the same block. By induction, every iterate is block diagonal under the same permutations, so no off-support fill-in is introduced during the iteration either. □

C.4 Proof of Theorem 1

Our proof is based on the GNN model from Figure 2(a) with the activation function being ReLU. GUMP is analyzed with $\mathbf{A}$ being unitary and vanilla message passing is analyzed with $\mathbf{A}$ being the normalized adjacency matrix.

Motivated by Xu et al. (2018), we first introduce the expected Jacobian. We stress that the common activation rate ρ is an analysis-only approximation: the ReLU gates along computation paths are modeled as i.i.d. Bernoulli variables with a depth-stable success probability ρ in order to isolate the effect of repeated graph propagation.

Theorem 4. *Given a L-layer GNN with ReLU as activation function, i.e., $\mathbf{H}^{(k)} = \mathsf{ReLU}(\mathbf{A}\mathbf{H}^{(k-1)}\mathbf{W}_k), \mathbf{H}^{(0)} = \mathbf{X}, k = 1\cdots L$, and under the i.i.d.-Bernoulli gate approximation with common success probability ρ, the expected Jacobian is*

$$\mathbb{E}\left[\frac{\partial\mathbf{h}_i^{(L)}}{\partial\mathbf{x}_s}\right] = \rho\prod_{l=L}^{1}\mathbf{W}_l^\top\left(\mathbf{A}^L\right)_{is}, \tag{3}$$

Proof. Denote by $\mathbf{f}_i^{(l)}$ the pre-activated feature of $\mathbf{h}_i^{(l)}$, i.e., $\mathbf{f}_i^{(l)} = \sum_{z\in\mathcal{N}(i)}\mathbf{A}_{iz}\mathbf{h}_z^{(l-1)}\mathbf{W}_l$, for any $l = 1\cdots L$, we have

$$\frac{\partial\mathbf{h}_i^{(l)}}{\partial\mathbf{h}_s^{(0)}} = \mathsf{diag}\left(1_{\mathbf{f}_i^{(l)}>0}\right)\cdot\left(\sum_{z\in\mathcal{N}(i)}\mathbf{A}_{iz}\frac{\partial\mathbf{h}_z^{(l-1)}}{\partial\mathbf{h}_s^{(0)}}\right)\cdot\mathbf{W}_l^\top.$$

By the chain rule, we get

$$\begin{aligned}\frac{\partial \mathbf{h}_i^{(L)}}{\partial \mathbf{h}_s^{(0)}} &= \sum_{p=1}^{\Psi} \left[\frac{\partial \mathbf{h}_i^{(L)}}{\partial \mathbf{h}_s^{(0)}}\right]_p \\ &= \sum_{p=1}^{\Psi} \prod_{l=L}^{1} \mathsf{diag}\left(1_{\mathbf{f}_{v_p^l}^{(l)}>0}\right) \mathbf{A}_{v_p^l v_p^{l-1}} \mathbf{W}_l^\top.\end{aligned}$$

Here, Ψ is the total number of paths $v_p^L v_p^{L-1} \cdots v_p^1 v_p^0$ of length $L+1$ from $v_p^0 = s$ to $v_p^L = i$. For $l = 1 \cdots L-1$, $v_p^{l-1} \in \mathcal{N}(v_p^l)$.

For each path p, the derivative $[\partial \mathbf{h}_i^{(L)} / \partial \mathbf{h}_s^{(0)}]_p$ represents a directed acyclic computation graph. At a layer l, we can express an entry of the derivative as

$$\left[\frac{\partial \mathbf{h}_i^{(L)}}{\partial \mathbf{h}_s^{(0)}}\right]_p^{(m,n)} = \prod_{l=L}^{1} \mathbf{A}_{v_p^l v_p^{l-1}} \sum_{q=1}^{\Phi} Z_q \prod_{l=L}^{1} w_q^{(l)},$$

where Φ is the number of paths q from the input neurons to the output neuron (m,n), in the computation graph of $[\partial \mathbf{h}_i^{(L)} / \partial \mathbf{h}_s^{(0)}]_p$. For each layer l, $w_q^{(l)}$ is the entry of $\mathbf{W}_l^\top$ that is used in the q-th path. Finally, $Z_q \in \{0,1\}$ represents whether the q-th path is active ($Z_q = 1$) or not ($Z_q = 0$) as a result of ReLU activation of the entries of $\mathbf{f}_{v_p^l}^{(l)}$'s on the q-th path.

Under the assumption that Z_q is a Bernoulli random variable with success probability ρ. Because of $\mathbb{P}[Z_q = 1] = \rho, \forall q$, we have

$$\mathbb{E}\left[\left[\frac{\partial \mathbf{h}_i^{(L)}}{\partial \mathbf{h}_s^{(0)}}\right]_p^{(m,n)}\right] = \rho \prod_{l=L}^{1} \mathbf{A}_{v_p^l v_p^{l-1}} \sum_{q=1}^{\Phi} \prod_{l=L}^{1} w_q^{(l)}.$$

Then, the expected Jacobian is

$$\mathbb{E}\left[\frac{\partial \mathbf{h}_i^{(L)}}{\partial \mathbf{x}_s}\right] = \sum_{p=1}^{\Psi} \mathbb{E}\left[\left[\frac{\partial \mathbf{h}_i^{(L)}}{\partial \mathbf{h}_s^{(0)}}\right]_p\right] = \rho \prod_{l=L}^{1} \mathbf{W}_l^\top \left(\mathbf{A}^L\right)_{is}.$$

□

Average-over-sources identity. For the following two theorems, let N denote the size of the propagation matrix, fix a target node $i \in [N]$, and sample the source node $s \sim \mathrm{Unif}([N])$. This metric is natural for our analysis because equation 3 factorizes the expected Jacobian into the cumulative weight term and the scalar graph-propagation term $(\mathbf{A}^L)_{is}$. Averaging over s therefore measures the influence of a typical source on the fixed target node i, while the squared magnitude removes sign or phase cancellations that can make individual entries uninformative. The resulting quantity is exactly the row energy of the propagation matrix at node i.

Lemma 3. *For any $\mathbf{P} \in \mathbb{C}^{N \times N}$ and any $L \geq 0$,*

$$\mathbb{E}_s\left[\left|(\mathbf{P}^L)_{is}\right|^2\right] = \frac{1}{N}\sum_{s=1}^{N}\left|(\mathbf{P}^L)_{is}\right|^2 = \frac{1}{N}\left\|\mathbf{e}_i^\top \mathbf{P}^L\right\|_2^2 = \frac{1}{N}\left(\mathbf{P}^L(\mathbf{P}^L)^\dagger\right)_{ii}.$$

Proof. By definition,

$$\mathbb{E}_s\left[\left|(\mathbf{P}^L)_{is}\right|^2\right] = \frac{1}{N}\sum_{s=1}^{N}\left|\mathbf{e}_i^\top \mathbf{P}^L \mathbf{e}_s\right|^2 = \frac{1}{N}\left\|\mathbf{e}_i^\top \mathbf{P}^L\right\|_2^2.$$

Moreover,

$$\left\|\mathbf{e}_i^\top \mathbf{P}^L\right\|_2^2 = \mathbf{e}_i^\top \mathbf{P}^L (\mathbf{P}^L)^\dagger \mathbf{e}_i = \left(\mathbf{P}^L (\mathbf{P}^L)^\dagger\right)_{ii},$$

which proves the claim. □

Proof of Theorem 1. From equation 3, the depth dependence of the expected Jacobian comes from the weight factor $\prod_{l=L}^{1} \mathbf{W}_l^\top$ and the graph-propagation factor $(\mathbf{A}^L)_{is}$.

Since $\mathbf{A}$ is unitary, $\mathbf{A}^L$ is unitary for every L, and hence

$$\mathbf{A}^L (\mathbf{A}^L)^\dagger = \mathbf{I}.$$

Applying Lemma 3 with $\mathbf{P} = \mathbf{A}$ gives

$$\mathbb{E}_s\left[\left|(\mathbf{A}^L)_{is}\right|^2\right] = \frac{1}{N}\left(\mathbf{A}^L (\mathbf{A}^L)^\dagger\right)_{ii} = \frac{1}{N}.$$

Therefore, the source-averaged influence is exactly depth-invariant.

Under the standard assumption that the cumulative weight factor $\prod_{l=L}^{1} \mathbf{W}_l^\top$ is kept well-conditioned and does not itself scale exponentially with L, any exponential decay or explosion in the expected Jacobian cannot come from the graph term. This is the intended claim of Theorem 1. □

C.5 Theory of Vanilla Message Passing

The next theorem analyzes the expected Jacobian of vanilla message passing.

Theorem 5. *For vanilla message passing with $\mathbf{A} = \hat{\mathbf{A}} \in \mathbb{R}^{N\times N}$, assume $\hat{\mathbf{A}}$ is real symmetric with spectral radius 1 and eigenvalues*

$$1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_N|, \qquad c := \max_{j\geq 2} |\lambda_j(\hat{\mathbf{A}})| < 1.$$

Let $\{v_j\}_{j=1}^N$ be an orthonormal eigenbasis and define $\mathbf{\Pi} := v_1 v_1^\top$. Then, for every depth $L \geq 0$,

$$\mathbb{E}_s\left[\left|(\hat{\mathbf{A}}^L)_{is}\right|^2\right] = \frac{1}{N} v_{1,i}^2 + \frac{1}{N}\sum_{j=2}^{N} v_{j,i}^2 \lambda_j^{2L},$$

and therefore

$$0 \leq \mathbb{E}_s\left[\left|(\hat{\mathbf{A}}^L)_{is}\right|^2\right] - \frac{1}{N} v_{1,i}^2 \leq \frac{1}{N} c^{2L}.$$

Equivalently, if $\delta_L(i,s) := (\hat{\mathbf{A}}^L - \mathbf{\Pi})_{is}$, then

$$\mathbb{E}_s\left[|\delta_L(i,s)|^2\right] = \frac{1}{N}\sum_{j=2}^{N} v_{j,i}^2 \lambda_j^{2L} \leq \frac{1}{N} c^{2L}.$$

Moreover, with $\alpha_i := \sum_{j:\,|\lambda_j|=c} v_{j,i}^2$, we also have

$$\mathbb{E}_s\left[|\delta_L(i,s)|^2\right] \geq \frac{1}{N}\alpha_i c^{2L}.$$

Thus, whenever $\alpha_i > 0$, the non-stationary source-averaged influence decays as $\Theta(c^{2L}/N)$.

Theorem 5 shows that vanilla message passing preserves only the stationary component while the non-stationary source-averaged influence decays exponentially with depth. Therefore, Theorems 1 and 5 indicate that GUMP removes the graph-induced exponential decay present in vanilla normalized propagation. The Jacobian trends in Section 4.4 are directionally consistent with Theorems 1 and 5.

proof of Theorem 5. Because $\hat{\mathbf{A}}$ is real symmetric, it admits the orthonormal eigendecomposition

$$\hat{\mathbf{A}} = \sum_{j=1}^{N} \lambda_j v_j v_j^\top, \qquad \hat{\mathbf{A}}^L = \sum_{j=1}^{N} \lambda_j^L v_j v_j^\top.$$

Since $\mathbf{\Pi} = v_1 v_1^\top$, we also have

$$\hat{\mathbf{A}}^L - \mathbf{\Pi} = \sum_{j=2}^{N} \lambda_j^L v_j v_j^\top.$$

Applying Lemma 3 with $\mathbf{P} = \hat{\mathbf{A}}$ and using symmetry gives

$$\mathbb{E}_s\left[\left|(\hat{\mathbf{A}}^L)_{is}\right|^2\right] = \frac{1}{N}\left(\hat{\mathbf{A}}^L(\hat{\mathbf{A}}^L)^\top\right)_{ii} = \frac{1}{N}\left(\hat{\mathbf{A}}^{2L}\right)_{ii}.$$

Expanding $\hat{\mathbf{A}}^{2L}$ in the eigenbasis yields

$$\left(\hat{\mathbf{A}}^{2L}\right)_{ii} = \sum_{j=1}^{N} \lambda_j^{2L} v_{j,i}^2 = v_{1,i}^2 + \sum_{j=2}^{N} \lambda_j^{2L} v_{j,i}^2,$$

which proves the exact decomposition. Since $|\lambda_j| \le c$ for $j \ge 2$ and $\sum_{j=2}^{N} v_{j,i}^2 \le 1$,

$$0 \le \frac{1}{N}\sum_{j=2}^{N} v_{j,i}^2 \lambda_j^{2L} \le \frac{1}{N} c^{2L} \sum_{j=2}^{N} v_{j,i}^2 \le \frac{1}{N} c^{2L}.$$

For the centered quantity, apply Lemma 3 to $\mathbf{P} = \hat{\mathbf{A}}^L - \mathbf{\Pi}$:

$$\mathbb{E}_s\left[|\delta_L(i,s)|^2\right] = \frac{1}{N}\left((\hat{\mathbf{A}}^L - \mathbf{\Pi})(\hat{\mathbf{A}}^L - \mathbf{\Pi})^\top\right)_{ii}.$$

Using the eigendecomposition above and orthonormality of $\{v_j\}$, we obtain

$$\mathbb{E}_s\left[|\delta_L(i,s)|^2\right] = \frac{1}{N}\sum_{j=2}^{N} v_{j,i}^2 \lambda_j^{2L} \le \frac{1}{N} c^{2L}.$$

For the lower bound, keep only the terms with $|\lambda_j| = c$:

$$\mathbb{E}_s\left[|\delta_L(i,s)|^2\right] \ge \frac{1}{N} \sum_{j:\,|\lambda_j|=c} v_{j,i}^2 c^{2L} = \frac{1}{N} \alpha_i c^{2L}.$$

Hence the non-stationary source-averaged influence decays at rate $\Theta(c^{2L}/N)$ whenever $\alpha_i > 0$. □

D More Experimental Results

Statistics of datasets The statistics of datasets used in experiments are shown in Table 3.

Table 3: Statistics of datasets.

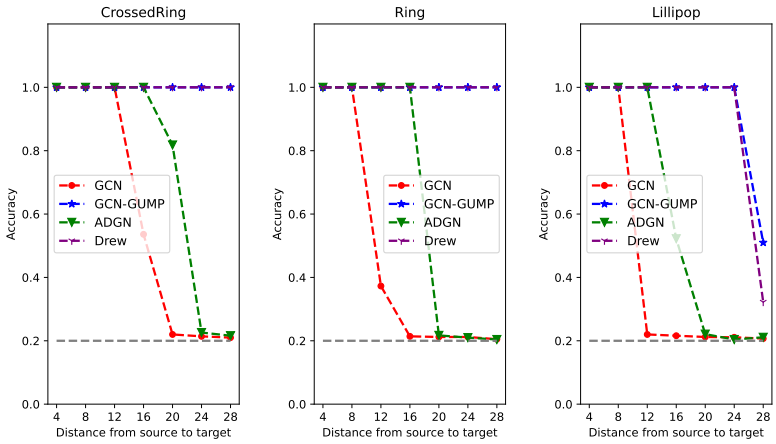
	#graphs	Avg. nodes	Avg. edges	Task type
Mutag	188	17.9	39.6	Graph Classification
Proteins	1,113	39.1	145.6	Graph Classification
Enzymes	600	32.6	124.3	Graph Classification
NC1	4110	29.87	32.30	Graph Classification
NC109	4127	29.68	32.13	Graph Classification
Peptides-func	15,535	150.94	307.30	Graph Classification
Peptides-struct	15,535	150.94	307.30	Graph Regression

Hyperparameters of GUMP The hyperparameters of GUMP for both synthetic and real datasets are shown in Table 4.

Table 4: Hyperparameters of GUMP for datasets in experiments. $\text{layer}_{\text{GUMP}}$, lr_{base}, wd_{base}, lr_{GUMP}, wd_{GUMP}, drop., d', d, batch size, $\text{layer}_{\text{base}}$, opt., sched., and epoch denotes the number of layers of GUMP, the learning rate of base GNN, weight decay of base GNN, the learning rate of GUMP, weight decay of GUMP, dropout rate, dimension of calculating (1), hidden dimension of GNN, batch size, number of layers of base GNN, optimizer, scheduler, and number of epochs, respectively.

	$\text{layer}_{\text{GUMP}}$	lr_{base}	wd_{base}	lr_{GUMP}	wd_{GUMP}	drop.	d'	d	batch size	$\text{layer}_{\text{base}}$	opt.	sched.	epoch
CrossedRing	-	10^{-4}	10^{-6}	10^{-4}	0	0	32	32	20	0	adam	none	200
Ring	-	10^{-4}	10^{-6}	10^{-4}	0	0	32	32	20	0	adam	none	200
CliquePath	-	10^{-4}	10^{-6}	10^{-4}	0	0	32	32	20	0	adam	none	200
Mutag	16	10^{-2}	10^{-4}	10^{-4}	0	0	32	64	16	5	adam	none	100
Proteins	20	10^{-2}	10^{-2}	10^{-4}	10^{-2}	0	32	64	64	3	adam	none	100
Enzymes	10	10^{-2}	10^{-4}	10^{-4}	0	0	32	64	16	1	adam	none	100
NC1	10	10^{-2}	10^{-4}	10^{-4}	0	0	32	64	16	1	adam	none	100
NC109	10	10^{-2}	10^{-4}	10^{-4}	0	0	32	64	16	1	adam	none	100
Peptides-func	12	0.005	0.1	0.1	0.1	0.2	32	256	200	3	adam	cos.	250
Peptides-struct	12	0.005	0.1	0.005	0.1	0.2	32	256	200	3	adam	cos.	250

Comparison GUMP with more methods For the synthetic datasets, we compare GUMP with Drew and ADGN on synthetic datasets in Figure 5. The results show that the performances of GUMP and Drew are close, while ADGN performs worse.



(a) CrossedRing (b) Ring (c) CliquePath

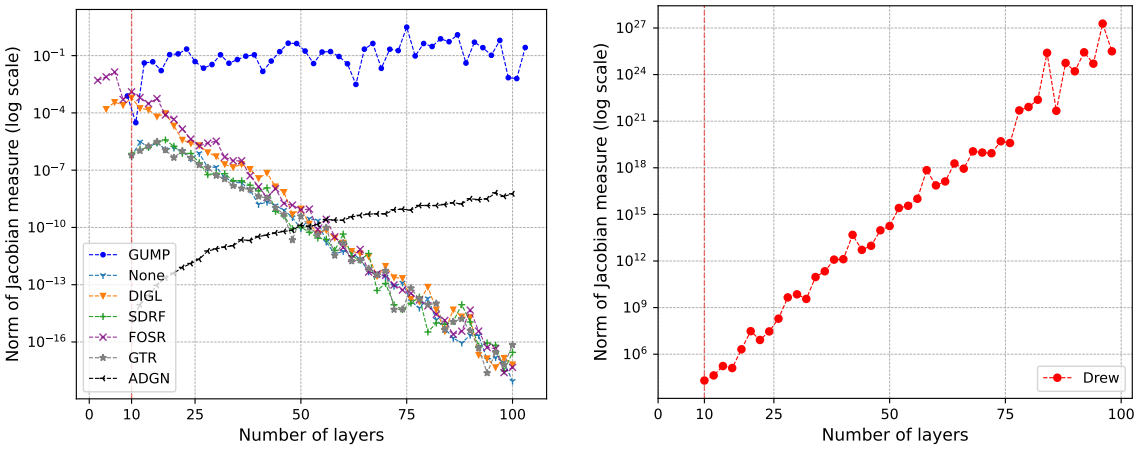
Figure 5: The performance of GCN and GCN-GUMP on the CrossedRing, Ring, and CliquePath with different distances from source to target.

We further compare the performances of GUMP by applying different convolution operator. The results are shown in Table 5. The results show that GUMP can improve the performance of both GCN and GIN, while GIN-GUMP achieves better performance than GCN-GUMP.

Table 5: Graph classification accuracy on the TUDataset. **First**, second, and *third* best results are bold, underlined, and underwaved, respectively.

Base GNN	Methods	Mutag	Proteins	Enzymes	NCI1	NCI109	Rank
GCN	None	72.15±2.44	70.98±0.74	27.67±1.16	68.74±0.45	67.90±0.50	4.2
	None (+layer)	70.05±1.83	69.80±0.99	23.63±1.07	63.94±1.34	55.92±1.26	6.8
	DIGL	*79.70±2.15*	70.76±0.77	35.72±1.12	69.76±0.42	69.37±0.43	3.0
	SDRF	71.05±1.87	70.92±0.79	*28.37±1.17*	68.21±0.43	66.78±0.44	4.8
	FoSR	80.00±1.57	73.42±0.81	25.07±0.99	57.27±0.54	56.82±0.60	4.6
	GTR	79.10±1.86	*72.59±2.48*	27.52±0.99	*69.37±0.38*	*67.97±0.47*	3.6
	GCN-GUMP	**84.89±1.63**	**74.88±0.87**	**36.02±1.43**	**77.97±0.42**	**75.85±0.44**	1.0
GIN	None	77.70±3.60	70.80±0.83	33.80±1.12	*75.65±0.49*	74.93±0.46	4.0
	None (+layer)	69.80±2.75	68.71±0.96	25.92±1.07	73.49±0.46	72.47±0.53	6.6
	DIGL	79.80±2.08	70.71±0.67	*35.74±1.20*	79.37±0.43	76.88±0.39	2.8
	SDRF	*78.40±2.80*	69.81±0.79	35.82±1.09	74.55±0.54	73.89±0.43	4.2
	FoSR	78.00±2.22	75.11±0.82	29.20±1.38	70.15±0.47	69.93±0.45	5.2
	GTR	77.60±2.84	*73.13±0.69*	30.57±1.42	75.45±0.44	*75.28±0.42*	4.2
	GIN-GUMP	**86.72±1.53**	**75.43±0.70**	**48.43±1.24**	**81.25±0.37**	**78.45±0.44**	1.0

We also compare the Jacobian of GUMP with Drew and ADGN in Figure 6. Jacobian of Drew exponentially increases, suggesting its potential numerical instability when training Drew with more layers. The Jacobian of ADGN is small when the ADGN layer is small and steadily increases to 10^{-8} as the ADGN layer reaches 100. Although the Jacobian of ADGN does not exhibit exponential decay, the correlation between distant nodes is significantly weaker compared to GUMP.



(a) (b)

Figure 6: Jacobian norm versus layers on NCI1

D.1 OGB MOLHIV Results

Setup We further evaluate graph classification on `ogbg-molhiv` from the Open Graph Benchmark (OGB) (Hu et al., 2020). The dataset contains molecular graphs and uses scaffold splitting to test generalization to structurally different molecules. Following the official OGB protocol, we report test ROC-AUC.

Results Table 6 reports the OGB MOLHIV results with the same comparison methods as Table 1. GIN-GUMP achieves a test ROC-AUC of 77.89, outperforming GRIT at 77.07, GPS at 76.62, GTR at 76.07, and the reproduced Unitary GNN baseline at 75.48. These results provide an additional standardized graph-classification evaluation beyond the TU datasets and show that GUMP remains effective under the official scaffold split.

Table 6: Experimental results on `ogbg-molhiv`. Test ROC-AUC (%) is reported under the official scaffold split and evaluator.

Classes	Method	Test ROC-AUC (%) ↑
–	GCN	74.47±1.51
Rewiring	DIGL	75.16±1.81
	SDRF	75.00±1.08
	FoSR	73.42±1.35
	GTR	76.07±0.70
Diffusion	ADGN	73.21±0.21
	GRAND	73.29±1.12
Transformer	GPS	76.62±0.82
–	Ortho-GConv	70.96±0.64
–	Unitary GNN	75.48±0.42
K-hop MP	GRIT	77.07±0.70
Ours	GIN-GUMP	77.89±0.54

D.2 Node Classification Results

We also apply GUMP to node classification tasks on Cora and Citeseer datasets. The results are shown in Table 7.

Table 7: Accuracy of node classification datasets: Cora and Citeseer

Layers		2	4	8	16	64
Cora	GCN	81.1	80.4	69.5	60.3	28.7
	GCNII	82.2	82.6	84.2	84.6	85.5
	GCN-GUMP	84.6	86.2	84.8	85.4	87.4
Citeseer	GCN	70.8	67.6	30.2	18.3	20.0
	GCNII	68.2	68.9	70.6	72.9	73.4
	GCN-GUMP	73.0	73.0	72.8	72.4	75.8

Training time of GUMP The time of training GCN and GCN-GUMP for 100 epochs on the TUDataset is shown in Table 8. Using the same graph-level batch size as GCN is a practical setting for graph classification, but it is not a node-budget-normalized throughput comparison. Since GUMP propagates on the directed line graph, a graph-level mini-batch contains $\sum_g 2|E_g|$ line-graph nodes rather than $\sum_g |V_g|$ original graph nodes. Therefore, we also report a node-budget-matched GUMP setting with

$$B_{\mathrm{GUMP}}^{\mathrm{match}} = \max\left(1, \left\lfloor B_{\mathrm{GCN}} \frac{\bar{n}}{2\bar{m}} \right\rfloor\right),$$

where $\bar{n}$ and $\bar{m}$ are the average numbers of nodes and edges per graph. This setting controls the expected number of propagated nodes per mini-batch, making the comparison fairer in terms of the amount of graph state propagated in each update.

Table 8: Training seconds on TUDataset for 100 epochs. The same-batch setting uses the shared GCN/GUMP graph-level batch sizes from the accuracy experiments, while the matched setting controls the expected propagated-node budget per mini-batch.

Dataset	GCN time (s)	GCN/GUMP graph batch	Same-batch GUMP time (s)	Matched GUMP batch	Matched GUMP time (s)
MUTAG	4.39	16	7.21	3	37.9
Proteins	20.57	64	26.19	8	203.7
Enzymes	11.26	16	20.29	2	160.2
NCI1	71.79	16	82.02	7	187.7
NCI109	74.56	16	93.38	7	213.5

Under this matched propagated-node budget, GUMP is more expensive than GCN: the matched GUMP times on MUTAG/PROTEINS/ENZYMES/NCI1/NCI109 are 37.9/203.7/160.2/187.7/213.5 seconds, respectively, which are larger than the corresponding GCN times. Thus, Table 8 should be interpreted as showing the practical same-batch runtime and the additional compute cost incurred by GUMP under a fairer propagated-node budget, rather than as evidence of comparable per-node cost to GCN.

E Apply GUMP to Directed Graphs

GUMP can also be applied to directed graphs in Table 4. The transformation of directed graphs is also based on Lemma 1.

Algorithm 4 Graph transformation for directed graphs

Require: A directed graph $G = (V, E)$;
1: Initialize a new digraph $G' = (V, E')$;
2: **for** $(i, j) \in E$ **do**
3: Add (i, j) and (j, i) to E';
4: **end for**
5: Remove duplicated edges in E';
6: Convert G' to its line graph $\mathsf{L}(G')$;
7: **Return:** A digraph $\mathsf{L}(G')$.