Gaussian Process Limit Reveals Structural Benefits of Graph Transformers

Nil Ayday[*1,3] **Lingchu Yang**[*2] **Debarghya Ghoshdastidar**[1,3,4]

[1]Technical University of Munich, TUM School of Computation, Information and Technology
[2]The University of Tokyo, Graduate School of Information Science and Technology
[3]Konrad Zuse School of Excellence in Reliable AI
[4]Munich Center for Machine Learning
nil.ayday@tum.de, yang-lingchu@g.ecc.u-tokyo.ac.jp, ghoshdas@cit.tum.de

Abstract

Graph transformers are the state-of-the-art for learning from graph-structured data and are empirically known to avoid several pitfalls of message-passing architectures. However, there is limited theoretical analysis on why these models perform well in practice. In this work, we prove that attention-based architectures have structural benefits over graph convolutional networks in the context of node-level prediction tasks. Specifically, we study the neural network gaussian process limits of graph transformers (GAT, Graphormer, Specformer) with infinite width and infinite heads, and derive the node-level and edge-level kernels across the layers. Our results characterise how the node features and the graph structure propagate through the graph attention layers. As a specific example, we prove that graph transformers structurally preserve community information and maintain discriminative node representations even in deep layers, thereby preventing oversmoothing. We provide empirical evidence on synthetic and real-world graphs that validate our theoretical insights, such as integrating informative priors and positional encoding can improve performance of deep graph transformers.

1 INTRODUCTION

Graph Neural Networks (GNNs) are currently the de facto models for learning from graph-structured data. They solve complex prediction tasks on graphs by iteratively aggregating feature information across edges. Recent empirical and theoretical work has established that classical message-passing GNN architectures have several limitations, including difficulties in simultaneously handling homophilic and heterophilic graphs [NT and Maehara, 2019, Ayday et al., 2025], oversmoothing by deep GNNs [Li et al., 2018, Keriven, 2022, Wu et al., 2023b], oversquashing [Alon and Yahav, 2021], etc. These challenges have motivated the development of attention-based GNNs, such as Graph Attention Network (GAT) [Velickovic et al., 2018], Graphormer [Ying et al., 2021], and Specformer [Bo et al., 2023], which are the current state of the art models in practice.

Despite the empirical success of graph transformers, there is little theoretical understanding of the benefits of the attention mechanism in graph learning. The rich literature on the expressivity of GNNs shows that transformers are more expressive than standard message-passing GNNs, but less expressive than more general message-passing architectures, like spectrally invariant GNNs [Zhang et al., 2024]. However, this line of work does not explain: *Why do graph transformers perform better with more layers and not suffer from oversmoothing, similar to message-passing GNNs?* Notably, *oversmoothing*—the phenomenon of node representations becoming indistinguishable in deeper layers—remains a topic of debate in recent theoretical works. Wu et al. [2023a] use a dynamical system-based equivalence of GATs to argue that they cannot prevent oversmoothing, while a Markov chain perspective of GATs suggests their ability to mitigate oversmoothing Zhao et al. [2023].

Unlike GAT, which is limited to local neighborhoods, global attention-based models, like Graphormer and Specformer, allow nodes to attend to all others, making them more complex to analyze, with no theoretical work currently available. This lack of a unified theoretical lens for graph transformers leaves a fundamental question open: *Can graph transformers overcome the structural bottlenecks of message-passing, or do they merely delay the onset of oversmoothing?*

We postulate that the above questions can be answered by studying the structural characteristics of graph transformers, specifically, *how node representations evolve across layers*. To this end, we utilize the framework of Neural Network Gaussian Processes (NNGPs), which establishes that randomly initialized neural networks converge in dis-

[*]Equal contribution.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:287–327.

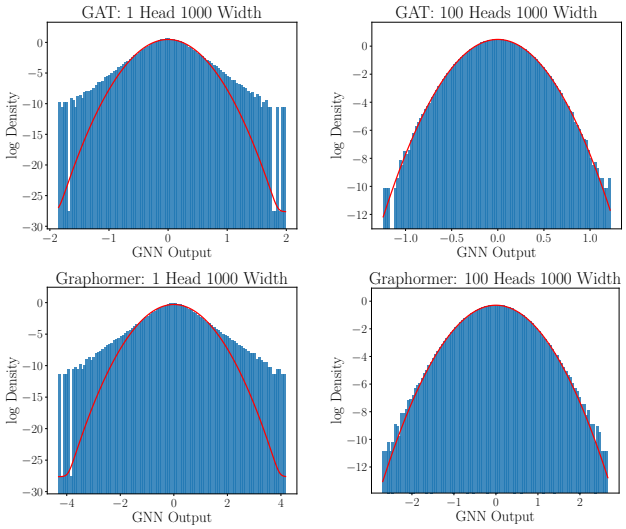

Figure 1: Histogram of the output of GAT and Graphormer for different number of heads. The output distribution converges to a Gaussian (red line) fitted with mean and variance of the empirical distribution when both width and number of heads are large. Plots for Specformer is in Appendix F.

tribution to Gaussian processes in the infinite-width limit [Lee et al., 2018b]. The NNGP limit allows one to simplify GNNs, but still retain the architectural properties related to information aggregation and attention mechanism. In particular, recent works derive NNGP and related neural tangent limits of Graph Convolutional Networks [Niu et al., 2023, Sabanayagam et al., 2023], which has also been used to characterize oversmoothing in these models. However, NNGP limits of graph transformers are not known.

Contributions and significance. The main technical contribution of this work is a derivation of the NNGP kernels for graph transformers, which we use to analytically show when and how these architectures can mitigate oversmoothing. We briefly discuss the significance of the contributions.

Derivation of NNGP kernel for graph transformers. We derive the NNGP limit of the three most relevant graph transformer architectures (GAT, Graphormer, and Specformer) and provide analytical forms of the kernel evolution across layers for GAT-GP (Theorem 7), Graphormer-GP (Theorem 10), and Specformer-GP (Theorem 12). In contrast to previous work on GCN, where NNGP-equivalence holds in the infinite-width limit [Niu et al., 2023], transformer architectures behave as NNGPs only when both the width and the number of heads are large (see Figure 1).

Furthermore, unlike the NNGP limit of standard transformers [Hron et al., 2020], we show that graph transformers induce graph-specific kernels at multiple levels—node-level kernels reflect the node feature interactions, whereas structural kernels capture connectivity patterns, which correspond to edge-level kernels over graph neighborhoods in spatial models (GAT, Graphormer) and spectral kernels in spectral models (Specformer). These kernels interact nontrivially for various architectures (see Tables 1–2).

Theoretical analysis of oversmoothing. To study whether graph transformers mitigate oversmoothing, we simplify the NNGP kernels under the population version of (contextual) stochastic block model (CSBM), that is, graphs with well-defined community structure. Our theoretical results, summarized in Table 3, confirm that while GCN is structurally destined for oversmoothing, all attention-based GNNs can retain community information across layers under specific conditions. In particular, GAT-GP avoids rank collapse under well-separated communities, whereas Specformer-GP and Graphormer-GP counteract representational decay through spectral amplification or informative structural priors. The analysis provides a crucial insight: if the NNGP limit of graph transformers does not oversmooth, then their neural architecture could potentially be trained to avoid oversmoothing noted in prior work [Wu et al., 2023a].

Empirical validation of the impact of depth. We validate our theoretical insights through numerical studies on synthetic and real-world benchmarks. As a practical consequence of our theory, we identify Laplacian-based positional encodings as a structural prior that prevents oversmoothing in Graphormer-GP. Moreover, with a sufficiently informative structural prior, performance can even improve as depth increases, which is unexpected in GCNs.

2 PRELIMINARIES

Notation. $\odot$ denotes the Hadamard product, $\|$ concatenation, and $|\cdot|_F^2$ the squared Frobenius norm, and $\mathrm{tr}(\cdot)$ the trace of a matrix. $\phi(\cdot)$ is the row-wise softmax and $\sigma(\cdot)$ an arbitrary activation function.

Graph Data. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with n nodes, where $\mathcal{V}$ denotes the set of nodes and $\mathcal{E}$ the set of edges. Let $A \in \mathbb{R}^{n\times n}$ denote the adjacency matrix of the graph. The neighborhood of a node $a \in \mathcal{V}$ is defined as $\mathcal{N}_a := \{\, i \in \mathcal{V} \mid (a,i) \in \mathcal{E} \,\}$. The d_{in}-dimensional features for all N nodes are collected in rows of the $n \times d_{\text{in}}$ matrix X, with the a-th row denoted by $x_a^\top$.

Graph Neural Networks (GNNs). GNNs are a class of deep learning models designed to learn node representations from graph-structured data by aggregating information from local neighborhoods. $S_{\text{GNN}}^{(\ell,h)} \in \mathbb{R}^{n\times n}$ denotes the graph convolution operator of head h in layer ℓ, whose form depends on the specific GNN architecture. For message-passing GNNs such as GCN, $d^{\ell,H} = 1$ and the graph convolution is fixed, determined solely by the graph topology, given by $S_{\text{GCN}} = D^{-1/2}AD^{-1/2}$. In comparison, graph transformers employ a data-dependent convolution in which neighborhood weights are learned via an attention mechanism. Given input features $f^0(X) := X$, pre-activation of

the ℓ-th layer is

$$g^{\ell}(X) := \Big(\|_{h=1}^{d^{\ell,H}} S_{\mathrm{GNN}}^{(\ell,h)} f^{\ell-1}(X)\, W^{\ell,h}\Big) W^{\ell,H}, \quad (1)$$

where $\|$ denotes concatenation of the attention heads and $W^{\ell,h} \in \mathbb{R}^{d_{\ell-1}\times d_\ell}$ and $W^{\ell,H} \in \mathbb{R}^{d_{\ell-1} d^{\ell,H}\times d_\ell}$ are learnable weight matrices. The layer output is obtained by applying a nonlinearity. ReLU activation is a common choice and is considered in this work: $f^{\ell}(X) = \mathrm{ReLU}\big(g^{\ell}(X)\big)$.

Gaussian Process Limit of GNN Node Representations. We analyse the node representations learned by GNNs through the lens of *Neural Network Gaussian Processes* (NNGPs). In the infinite-width and head limit, the model's pre-activations $\{g_a^{\ell}(X)\}_{a\in V}$ converge in distribution to a centered Gaussian process $g^{\ell}(X) \sim \mathcal{GP}(0, \Sigma^{(\ell)})$. For ReLU activations ($f^{\ell}(X) = \mathrm{ReLU}\big(g^{\ell}(X)\big)$), the post-activation kernel $K^{(\ell)}$ admits a closed-form recursive update [Bietti and Mairal, 2019]:

$$K_{ab}^{(\ell)} = \frac{1}{2\pi}\sqrt{\Sigma_{aa}^{(\ell)}\,\Sigma_{bb}^{(\ell)}}\Big(\sin\theta_{ab}^{(\ell)} + \big(\pi - \theta_{ab}^{(\ell)}\big)\cos\theta_{ab}^{(\ell)}\Big),$$
$$\theta_{ab}^{(\ell)} = \arccos\left(\frac{\Sigma_{ab}^{(\ell)}}{\sqrt{\Sigma_{aa}^{(\ell)}\,\Sigma_{bb}^{(\ell)}}}\right). \quad (2)$$

Starting from the initial kernel $K_{ab}^{(0)} := \frac{x_a \cdot x_b}{d_{\mathrm{in}}}$, the sequence of kernels $\{\Sigma^{(\ell)}\}_{\ell\geq 0}$ and $\{K^{(\ell)}\}_{\ell\geq 0}$ are defined recursively. Unlike standard NNGPs, in which the kernel is solely based on node features, the GNNs incorporate the graph structure and obtain predictions on unlabeled nodes by conditioning on observed node labels. In this work, we characterise the NNGP induced by graph transformers by deriving the corresponding covariance updates, thereby placing models such as GAT, Graphormer, and Specformer within a unified Gaussian process framework.

In the following, we introduce graph transformers investigated in this paper: GAT, Graphormer, and Specformer. In contrast to these models, which operate on homogeneous graphs, the Graph Transformer Network (GTN) is designed for heterogeneous graphs and is presented in Appendix D, including its model formulation, GP-equivalent theorem, and proof. The design choices underlying GAT, Graphormer, and Specformer are detailed in Appendix A.

Graph Attention Network (GAT). A multi-head GAT consists of $d^{\ell,H}$ attention heads in layer ℓ. For each head h and layer ℓ, attention scores are computed on edges $(a, i) \in \mathcal{E}$:

$$E_{ai}^{(\ell h)}(X) = \big(\vec{v}^{\ell h}\big)^{\top}\big[W^{\ell,h} f_a^{\ell-1}(X) \,\|\, W^{\ell,h} f_i^{\ell-1}(X)\big] \quad (3)$$

where $\vec{v}^{\ell h} \in \mathbb{R}^{2d_\ell\times 1}$ is the attention vector.

The normalized attention coefficients are obtained via

$$\big(S_{\mathrm{GAT}}^{(\ell,h)}(X)\big)_{ai} = \frac{\exp\big(\sigma(E_{ai}^{(\ell h)}(X))\big)}{\sum_{k\in\mathcal{N}_a}\exp\big(\sigma(E_{ak}^{(\ell h)}(X))\big)}, \quad (4)$$

and $\big(S_{\mathrm{GAT}}^{(\ell,h)}(X)\big)_{ai} = 0$ for $(a, i) \notin \mathcal{E}$.

Graphormer. A Graphormer is a multi-head attention-based architecture with $d^{\ell,H}$ heads in layer ℓ. Each layer augments node representations with graph positional encodings: $\tilde{f}_i^{\ell-1}(X) := [f_i^{\ell-1}(X),\, P_i^{\ell}]$ for $i \in V$, where P_i^{ℓ} encodes structural information such as Laplacian eigenvectors, degree features, or random-walk-based statistics. For each head h and layer ℓ, attention scores are computed on all node pairs $(i, j) \in V \times V$ as

$$E_{ij}^{(\ell h)}(X) = \frac{\big(\tilde{f}_i^{\ell-1}(X)W^{Q,\ell h}\big)\,\big(\tilde{f}_j^{\ell-1}(X)W^{K,\ell h}\big)^{\top}}{\sqrt{d_\ell}} + b_{\rho(i,j)}, \quad (5)$$

where $W^{Q,\ell h}, W^{K,\ell h} \in \mathbb{R}^{d_{\ell-1}\times d_\ell}$ are learnable projection matrices, and the scaled dot-product term measures content-based compatibility between node i and node j. The additive bias $b_{\rho(i,j)}$ is a learnable scalar that depends only on the structural relation $\rho(\cdot,\cdot) : V \times V \to \mathcal{R}$ (e.g., shortest-path distance bucket, edge type, or other graph-structural categories).The graph convolution operator is obtained by applying the row-wise softmax, to the logits, i.e., $S_{\mathrm{Graphormer}}^{(\ell,h)}(X) = \phi\big(E^{(\ell h)}(X)\big)$ for head h.

Specformer. Specformer operates on the spectrum of the normalized graph Laplacian: $L := I_n - D^{-1/2}AD^{-1/2} = U\Lambda U^{\top}$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Each eigenvalue λ_i is encoded by a sinusoidal map $\rho(\lambda_i) \in \mathbb{R}^d$ with $\rho(\lambda, 2t) = \sin\Big(\epsilon\lambda/10000^{2t/d}\Big), \rho(\lambda, 2t+1) = \cos\Big(\epsilon\lambda/10000^{2t/d}\Big)$ and the initial spectral tokens are formed as $H^0 = \Big[(\lambda_1\|\rho(\lambda_1))^{\top}, \ldots, (\lambda_n\|\rho(\lambda_n))^{\top}\Big]^{\top} \in \mathbb{R}^{n\times(d+1)}$.

Consider a multi-head Specformer consisting of an *eigenvalue encoding* module with $d^{t,H}$ heads and a *graph convolution* module with $d^{\ell,H}$ heads. For each graph convolution head $h \in \{1, \ldots, d^{\ell,H}\}$, the eigenvalue encoding module applies a $d^{t,H}$-head attention at each eigenvalue encoding layer t, producing head-specific attention scores on token pairs $(i, j) \in V \times V$:

$$E_{ij}^{(t,k,h)} = \frac{\big(H_i^{t-1,k,h}W^{Q,tkh}\big)\big(H_j^{t-1,k,h}W^{K,tkh}\big)^{\top}}{\sqrt{d_t}}. \quad (6)$$

Here, $k \in 1, \ldots, d^{t,H}$ indexes the attention head within token layer t, and $W^{Q,tkh}, W^{K,tkh} \in \mathbb{R}^{d_{t-1}\times d_t}$ are the corresponding projection matrices. The normalized attention coefficients are given by $S_{\mathrm{Spec}}^{(t,k,h)} = \phi\big(E^{(t,k,h)}\big)$. The token-layer update for the h-th graph convolution head is

$$H^{t,h} = \Big(\|_{k=1}^{d^{t,H}} S_{\mathrm{Spec}}^{(t,k,h)}\, H^{t-1,k,h}W^{V,tkh}\Big)W^{th,H}, \quad (7)$$

where $W^{V,tkh} \in \mathbb{R}^{d_{t-1}\times d_t}$ is the value projection for the k-th eigenvalue encoding head and $W^{t,H} \in \mathbb{R}^{d^{t,H}d_t\times d_t}$ is the token-layer output projection. After T eigenvalue layers, the

aggregated eigenvalue encoding output associated with the h-th graph convolution head is denoted by $\bar{H}^{(h)} = H^{T,h}$.

The graph convolution module then constructs a multi-head spectral filtering operator. For each graph convolution head $h \in \{1, \ldots, d^{\ell,H}\}$, a spectral filter is generated by $\bar{\lambda}^{(h)} = \sigma\big(\bar{H}^{(h)} W_\lambda^{(h)}\big) \in \mathbb{R}^n$, where $W_\lambda^{(h)} \in \mathbb{R}^{d_T \times 1}$ is a learnable parameter vector associated with the h-th graph convolution head. This induces the graph convolution operator $S^{(\ell,h)}_{\text{Specformer}} = U \operatorname{diag}\big(\bar{\lambda}^{(h)}\big) U^\top \in \mathbb{R}^{n \times n}$.

3 GAUSSIAN PROCESS LIMITS OF GRAPH TRANSFORMERS

In this section, we analyse the GP limits of Graph Transformers architectures introduced in Section 2. We begin by introducing the kernels that govern the Gaussian process limit and explain their roles, then present their explicit formulas in Table 1, and finally interpret their implications for the corresponding architectures in Section 3.1.

Our analysis is carried out within the Neural Network Gaussian Process (NNGP) framework. Specifically, we characterise the behavior of GAT, Graphormer, and Specformer in the **infinite-width and infinite-head limit**, denoted as $d_\ell, d_{\ell,H} \to \infty$, and under the independent Gaussian priors listed in Table 4. In this regime, the models admit kernel recursions that govern their layerwise propagation. In particular, we focus on the pre-activations g^ℓ with kernel $\Sigma^{(\ell)}$, and the post-activation kernel $K^{(\ell)}$ is obtained via the standard ReLU NNGP transformation (Equation 2). The complete theorems are provided in Appendix B.

Spatial Models: GAT-GP and Graphormer-GP To provide a unified analysis of GAT and Graphormer, we define four kernels that govern the GP limit. These kernels model the propagation of information: we define the structural input (post-activation), model the attention mechanism (edge-level), aggregate these into a graph operator (convolution), and finally produce the node representations (node-level). Together, they form a recursion that characterizes the evolution of graph signals through deep layers.

- **Post-Activation Kernel of the Previous Layer:** The propagation is initialized using the kernel from the previous iteration. For GAT, this is the post-activation kernel $K^{(\ell-1)}$. For Graphormer, this is the **augmented kernel** $\tilde{K}^{(\ell)} = \alpha K^{(\ell-1)} + (1-\alpha) R^{(\ell)}$ for $\alpha \in [0,1]$, which incorporates structural positional encoding $R^{(\ell)}$.
- **Edge-Level Logit Kernel ($K^{(\ell),E}$):** The attention logits $E^{(\ell)}$ converge to a centered Gaussian process $E^{(\ell)} \sim GP(0, K^{(\ell),E})$. This kernel describes the covariance between attention scores for different edges.
- **Convolution Kernel ($C^{(\ell)}$):** This kernel captures the covariance of the graph convolution operator across independent heads and is defined as $C^{(\ell)}_{ai,bj} := \mathbb{E}\Big[\big(S^{(\ell,1)}(X)\big)_{ai} \big(S^{(\ell,1)}(X)\big)_{bj}\Big]$. Its explicit form is determined by the Edge-Level Logit Kernel $K^{(\ell),E}$.
- **Node-Level Kernel ($\Sigma^{(\ell)}$):** The final covariance between node representations a and b after neighborhood aggregation, where $\Sigma^{(\ell)}_{ab} := \mathbb{E}[g^\ell_{a1}(X) g^\ell_{b1}(X)]$.

Spectral Model: Specformer-GP Unlike spatial models, Specformer operates on the spectrum of L. Alongside the layerwise limits, when $d_t, d_{t,H} \to \infty$, the model converges to a Gaussian process characterized by a set of kernels:

- **Spectral Token Kernel:** The representations of the t-th token layer, $H^{(t)}$, converge in distribution to a Gaussian process $H^{(t)} \sim GP(0, K_H^{(t)})$, which captures the evolution of spectral information through the eigenvalue encoding.
- **Learned Spectral Kernel (K_λ):** After T token layers, the learned spectral filter coefficients $\bar{\lambda}$, induces a kernel defined as $K_{\lambda,ab} := \mathbb{E}[\bar{\lambda}_a \bar{\lambda}_b]$.
- **Node-Level Kernel ($\Sigma^{(\ell)}$)** This kernel propagates feature covariance by mapping previous-layer representations into the Laplacian eigenbasis for elementwise modulation by the induced spectral kernel (K_λ) before projecting them back to the node domain.

The specific kernels for each architecture are summarized in Table 1. The independent Gaussian weight priors considered are summarized in Table 4 at the beginning of Appendix B. Formal theorems establishing the existence of the Gaussian process limits are provided in Appendix B, and detailed proofs of these theorems can be found in Appendix C.

Explicit Kernel Expressions. Table 1 summarizes a general kernel recursion, which can be specialized by considering specific choices of the attention and activation nonlinearities to obtain a closed-form expressions. Table 2 presents the closed-form kernels obtained under linear attention (formally in Corollaries 9, and 11). For completeness, and to facilitate the comparison with attention-based architectures in the Section 4, we also include the corresponding GCN-GP kernel from [Niu et al., 2023], which does not incorporate attention mechanisms. These kernels provide a tractable way to study the GP limit under simplified attention and activation choices, while still capturing the interaction between node features and graph structure. Although our explicit kernel derivations rely on linear attention for mathematical tractability, Appendix F.3 demonstrates that linear attention is an effective practical alternative, achieving test accuracy comparable to or better than softmax attention. More general kernels can be obtained by considering RELU as output activation and applying Equation 2. In Appendix B Corollary 8, we also present GAT with RELU attention.

Table 1: Summary of the GP kernels for graph transformers. The weight priors are in Table 4 in Appendix B. The table presents both the edge/structural kernels ($K_{ab}^{(\ell),E}$ for spatial models or $K_{\lambda,ab}$ for Specformer) and the node-level kernels ($\Sigma_{ab}^{(\ell)}$) that describe the propagation of covariance across layers. For GAT-GP and Graphormer-GP, the edge-level kernels capture correlations between attention scores, while the node-level kernels aggregate these correlations. The learned spectral kernel K_λ of Specformer-GP modulates the mapping of node representations in the Laplacian eigenbasis.

Model	**Edge/Structural Kernel** ($K_{ab}^{(\ell),E}$ **or** $K_{\lambda,ab}$)	**Node Kernel** $\Sigma_{ab}^{(\ell)}$
GAT-GP (Theorem 7)	$\sigma_w^2\sigma_v^2\big(K_{ab}^{(\ell-1)} + K_{ij}^{(\ell-1)}\big)$	$\sigma_H^2\sigma_w^2 \sum_{i\in N_a}\sum_{j\in N_b} K_{ij}^{(\ell-1)}\, C_{ai,bj}^{(\ell)}$
Graphormer-GP (Theorem 10)	$\Big(\sigma_Q^2\sigma_K^2\, \tilde{K}_{ab}^{(\ell-1)}\, \tilde{K}_{ij}^{(\ell-1)} + \sigma_b^2\, \mathbf{1}_{\phi(a,i)=\phi(b,j)}\Big)$	$\sigma_H^2\sigma_w^2 \sum_{i,j\in V} \tilde{K}_{ij}^{(\ell-1)} C_{ai,bj}^{(\ell)}$
Specformer-GP (Theorem 12)	$K_{\lambda,ab} = \mathbb{E}[\bar{\lambda}_a\bar{\lambda}_b]$ (Spectral)	$\sigma_H^2\sigma_w^2 U(K_\lambda \odot (U^\top K^{(\ell-1)}U))U^\top$

3.1 REMARKS ON GP KERNELS

The kernels in Table 1 reveal how node features and graph structure interact in the GP limit. In the following, we discuss the interpretation of these kernels in the graph domain and their implications for node representation propagation.

GAT-GP. The edge-level kernel $K^{(\ell),E}$ of GAT-GP characterizes the covariance between attention logits on two edges (a,i) and (b,j), determined by the similarity of their source nodes a and b and their target nodes i and j. Edges whose endpoints are simultaneously similar in representation space exhibit stronger correlation. The node-level kernel $\Sigma^{(\ell)}$ then aggregates these edge contributions across all neighbor pairs (i,j) with $i \in \mathcal{N}_a$ and $j \in \mathcal{N}_b$, combining the previous-layer feature similarity with the alignment induced by attention. This yields a kernel recursion describing how both features and graph structure govern node representation propagation across layers.

Graphormer-GP Graphormer incorporates graph information through positional embeddings. The matrix R can be interpreted as the covariance matrix of the corresponding positional embeddings which can be computed as an inner product, i.e., $R_{ab}^{(\ell)} = \langle P_a^{(\ell)}, P_b^{(\ell)}\rangle$. Alternatively, instead of introducing explicit positional embeddings, the graph structure itself can be directly leveraged to construct a node-level graph-structure covariance matrix. The node-level kernel $\Sigma^{(\ell)}$ of Graphormer-GP indicates that attention correlations are strengthened for node pairs sharing the same structural relation, ensuring that graph information is explicitly leveraged in guiding attention. The edge-level kernel $K^{(\ell),E}$ then propagates the attention correlations to the next layer through an aggregation over all node pairs.

Unlike GAT which is restricted to local neighborhoods, Graphormer induces similarity connections between every pair of nodes in the graph.

Specformer. The convolution in Specformer is composed of *spectral tokens* and *node features*. The spectral tokens are first processed through T layers of encoding, and passed through a decoder to construct a data-dependent spectral filtering that governs the propagation of node features. As detailed in Theorem 12 in Appendix B, (I) the spectral-token representations H^t, which are updated through attention, and (II) the node-feature pre-activations $g(X)$ converge in distribution to a Gaussian process. The learned spectral kernel K_λ induces a kernel over the learned spectral filter coefficients. The node level kernel Σ^l then defines the recursion of the node-feature kernel: the previous-layer node covariance is mapped into the Laplacian eigenbasis, modulated elementwise according to the induced spectral kernel, and projected back to the node domain.

Spatial models aggregate information over local node neighborhoods, whereas Specformer operates in the graph spectral domain.

4 ANALYSIS OF GNN-GPS UNDER CONTEXTUAL STOCHASTIC BLOCK MODEL (CSBM)

The GP formulation of graph transformers allows us to compare structural benefits and limitations of different architectures. Having introduced kernel recursion of graph transformers, we now focus on a key phenomenon in deep graph models: **Oversmoothing**. In the kernel perspective, this corresponds to convergence of the node-level covariance toward a constant matrix. To understand whether graph transformers can structurally maintain discriminative signals across layers, it is useful to study their GP kernels on a controlled setting with known community structure.

In this section, we analyse how the resulting kernels behave when the underlying graph is governed by a **CSBM**. We consider population version of 2-CSBM with n nodes par-

Table 2: Closed-form expressions for the node-level Gaussian process (GP) kernels under linear attention and linear activation. These kernels provide an explicit characterization of kernel propagation for each architecture. For comparison, the GCN-GP kernel from [Niu et al., 2023] is included, which does not involve attention.

Model	**Node Kernel** ($K^{(\ell)}$)
GCN-GP [Niu et al., 2023]	$K^{(\ell)} = \sigma_w^2 A K^{\ell-1} A^T$
GAT-GP (Corollary 9)	$K^{(\ell)} = \sigma_H^2 \sigma_w^4 \sigma_v^2 \Big(K^{(\ell-1)} \odot (A K^{(\ell-1)} A^\top) + A(K^{(\ell-1)} \odot K^{(\ell-1)}) A^\top \Big)$
Graphormer-GP (Corollary 11)	$K_{ab}^{(\ell)} = \sigma_H^2 \sigma_w^2 \left[\sigma_Q^2 \sigma_K^2 \tilde{K}_{ab}^{(\ell)} \sum_{i,j \in V} \tilde{K}_{ij}^{(\ell-1)} \tilde{K}_{ij}^{(\ell-1)} + \sigma_b^2 \sum_{i,j \in V} \tilde{K}_{ij}^{(\ell-1)} \mathbf{1}_{\rho(a,i)=\rho(b,j)} \right]$
Specformer-GP (Theorem 12)	$K^{(\ell)} = \sigma_H^2 \sigma_w^2 U (K_\lambda \odot (U^\top K^{(\ell-1)} U)) U^\top$

titioned into two equally sized communities, $\mathcal{C}_1$ and $\mathcal{C}_2$. In this setting, the probability of an edge existing between two nodes is p if they belong to the same community and q if they belong to different communities. Consequently, CSBM accommodates both homophilic graphs (where $p > q$, favoring intra-community connections) and heterophilic graphs (where $p < q$, favoring cross-community edges). The expected adjacency matrix takes the following block structure: $A = \begin{pmatrix} p\mathbf{1}\mathbf{1}^\top & q\mathbf{1}\mathbf{1}^\top \\ q\mathbf{1}\mathbf{1}^\top & p\mathbf{1}\mathbf{1}^\top \end{pmatrix}$, where $\mathbf{1}\mathbf{1}^\top$ denotes the all-ones matrix of size $\frac{n}{2} \times \frac{n}{2}$ [1]. A kernel that captures the community structure should mirror this block structure, i.e. $K^{(\ell)} = \begin{pmatrix} x_\ell \mathbf{1}\mathbf{1}^\top & y_\ell \mathbf{1}\mathbf{1}^\top \\ y_\ell \mathbf{1}\mathbf{1}^\top & x_\ell \mathbf{1}\mathbf{1}^\top \end{pmatrix}$, where x_ℓ represents the kernel value for intra-community node pairs and y_ℓ for inter-community pairs at layer ℓ, initialized by the feature covariance values x_0 and y_0 at the input layer ($\ell = 0$). While we analyse the population version, our experiments use random CSBM realizations, which exhibit similar behavior. We summarise the evolution of the kernels (formulas for x_ℓ and y_ℓ) and their asymptotic behaviour in Table 3, highlighting how each model propagates structural information under the CSBM. The full corollaries and proofs for the closed-form expressions of the kernels are provided in Appendix E.

Oversmoothing can be formalized by examining the normalized kernel $\frac{K^{(\ell)}}{\frac{1}{n} tr(K^{(\ell)})}$. Since the normalisation term corresponds to x_ℓ, the discriminability of the model is captured by the ratio y_ℓ / x_ℓ: if $\lim_{\ell \to \infty} y_\ell / x_\ell = 1$, the kernel converges to a rank-one matrix of ones, meaning the representations of the two communities become indistinguishable.

GCN. The collapse into a rank-one kernel is clearly visible in the Gaussian Process limit of standard GCN.

Corollary 1 (Rank collapse in GCN-GP indicates oversmoothing)**.** *For the GCN kernel presented in Corollary 37 (and in Table 3). The normalized kernel converges to the all-ones matrix:*

$$\lim_{\ell \to \infty} \frac{K^{(\ell)}}{\frac{1}{n} tr(K^{(\ell)})} = \begin{pmatrix} \mathbf{1}\mathbf{1}^\top & \mathbf{1}\mathbf{1}^\top \\ \mathbf{1}\mathbf{1}^\top & \mathbf{1}\mathbf{1}^\top \end{pmatrix}.$$

Consequently, rank $\left(\lim_{\ell \to \infty} \frac{K^{(\ell)}}{\frac{1}{n} tr(K^{(\ell)})} \right) = 1$*, indicating that GCNs suffer from complete oversmoothing.*

The detailed derivations for the closed-form expressions of the GCN kernel and the spectral analysis leading to rank collapse are provided in Appendix E.1.

GAT. Unlike the GCN kernel, which collapses to a rank-one matrix, we next show that GAT-GP avoids oversmoothing. Notably, while the GCN kernel follows a simple linear recurrence, the GAT kernel involves a coupled, non-linear two-variable system. The derivation of the GAT kernel under SBM (Corollary 38) and its limit (Corollary 2) are provided in Appendix E.2. The GAT kernel is governed by the **global growth factor** $G = (x_0 + y_0)(p + q)^2 (\frac{n}{2})^2$ and the **structural preservation factor** $F = 2\frac{p^2 - pq + q^2}{(p+q)^2}$. Here, G controls the overall magnitude of the kernel, while F determines whether the community structure is preserved across layers. In particular, when the communities are sufficiently well-separated ($\frac{p}{q}$ is bounded away from 1), F ensures that the kernel maintains its ability to distinguish between them even in the infinite-depth limit. This property is formalised in the following corollary:

Corollary 2 (GAT-GP avoids rank collapse indicating the preservation of discriminative community structure)**.** *For the GAT kernel defined in Corollary 38, the normalized kernel converges to:* $\lim_{\ell \to \infty} \frac{K^{(\ell)}}{\frac{1}{n} tr(K^{(\ell)})} = \begin{pmatrix} \mathbf{1}\mathbf{1}^\top & \gamma \mathbf{1}\mathbf{1}^\top \\ \gamma \mathbf{1}\mathbf{1}^\top & \mathbf{1}\mathbf{1}^\top \end{pmatrix}$, *where* $\gamma = \lim_{\ell \to \infty} \frac{1 - \left(\frac{x_0 - y_0}{x_0 + y_0}\right) F^\ell}{1 + \left(\frac{x_0 - y_0}{x_0 + y_0}\right) F^\ell}$. *Consequently, if* $F \geq 1$*, which occurs when* $\frac{p}{q}$ *is bounded away from* 1*, the kernel maintains a rank of* 2 *and preserves community separation as* $\ell \to \infty$*, avoiding oversmoothing.*

[1] Throughout this section, SBM refers to the two-community stochastic block model introduced above.

Table 3: **Evolution of the GP kernel $K^{(\ell)} = \begin{pmatrix} x_\ell \mathbf{1}\mathbf{1}^\top & y_\ell \mathbf{1}\mathbf{1}^\top \\ y_\ell \mathbf{1}\mathbf{1}^\top & x_\ell \mathbf{1}\mathbf{1}^\top \end{pmatrix}$ under CSBM and conditions for oversmoothing.** To preserve community structure, the kernel must maintain a gap between the intra-community (x_ℓ) and inter-community (y_ℓ) values; this discriminative signal is captured by the second term in each expression.The second column indicates whether oversmoothing occurs as $\ell \to \infty$, i.e., whether the kernel loses its ability to distinguish the two communities. For graph transformers oversmoothing can be avoided, maintaining the discriminative signal across layers.

Model	**Intra/Inter-Community Formulas** (x_ℓ, y_ℓ)	**Oversmoothing** $(\ell \to \infty)$
GCN-GP	$\frac{1}{2}(\frac{n}{2})^{2\ell}\left[(x_0+y_0)(p+q)^{2\ell} \pm (x_0-y_0)(p-q)^{2\ell}\right]$ (Corollary 37)	Yes (Corollary 1)
GAT-GP	$\frac{1}{2}\frac{G^{2\ell}}{(\frac{n}{2})^2(p+q)^2}\left[1 \pm \left(\frac{x_0-y_0}{x_0+y_0}\right) F^\ell\right]$ (Corollary 38)	No, if communities are separated ($F \geq 1$) (Corollary 2)
Graphormer-GP	$\tilde{x}_\ell = \alpha^\ell x_0 + (1-\alpha^\ell)p, \quad \tilde{y}_\ell = \alpha^\ell y_0 + (1-\alpha^\ell)q$ (Corollary 39)	No, if the prior (R) captures communities (Remark 3)
Specformer-GP	$\frac{1}{2}\left[(x_0+y_0)\bar{\lambda}_1^{2\ell} \pm (x_0-y_0)\bar{\lambda}_2^{2\ell}\right]$ (Corollary 40)	No, if $\lvert\bar{\lambda}_2\rvert \geq \lvert\bar{\lambda}_1\rvert$ (Remark 5)

Graphormer. Consider the Graphormer kernel from Corollary 11 in Appendix B without the bias term, and assume that the positional encodings capture all structural information of the graph. In the SBM setting, this implies that the spatial relation matrix coincides with the adjacency matrix, i.e., $R = A$. Under SBM, the block entries $\tilde{x}_\ell$ and $\tilde{y}_\ell$ of $\tilde{K}^{(\ell)}$ are determined by the recurrence given in Table 3. The normalized kernel from Table 2 without the bias term can be written as $K^{(\ell)} = \tilde{K}^{(\ell)} \cdot Z_{\ell-1}$, with $Z_{\ell-1} = tr(A^\top(\tilde{K}^{(\ell-1)} \odot \tilde{K}^{(\ell-1)}))$. Since our analysis focuses on whether the kernel preserves the block structure induced by the SBM, the trace term does not play a role. Consequently, it suffices to study the evolution of $\tilde{K}^{(\ell)}$.

Remark 3 (Graphormer Convergence)**.** *As the number of layers $\ell \to \infty$, the unnormalized Graphormer kernel $\tilde{K}^{(\ell)}$ converges to the spatial relation matrix R. In the SBM setting where $R = A$, the kernel effectively recovers the ground truth, improving its ability to distinguish communities with increasing depth. This benefit, however, relies entirely on the quality of the positional encodings; in a real-world dataset if R is uninformative, the kernel may collapse toward a non-discriminative prior, resulting in oversmoothing.*

Specformer. The evolution of the Specformer kernel under SBM is presented in Table 3. Unlike the GCN and Graphormer kernels, the Specformer uses a learnable spectral filter, which is characterised by the parameters $\bar{\lambda}_1$ and $\bar{\lambda}_2$ that determine the cross-covariance K_λ (Equation 33). These eigenvalues control how different spectral components are amplified across layers, and therefore determine whether the kernel preserves the block structure induced by the SBM. Depending on the choice of $\bar{\lambda}_1$ and $\bar{\lambda}_2$, Specformer can either collapse to the GCN behaviour or maintain community separation, as detailed in the following remarks.

Remark 4 (Specformer Collapses to GCN)**.** *The Specformer-GP kernel collapses to the GCN-GP kernel if the learned spectral weights correspond to the eigenvalues of the adjacency matrix, i.e., $\bar{\lambda}_1 = \frac{n}{2}(p+q)$ and $\bar{\lambda}_2 = \frac{n}{2}(p-q)$. In this case, the Specformer kernel exhibits the same oversmoothing behaviour (Corollary 1).*

Remark 5. *Specformer can preserve community structure, even in the limit $\ell \to \infty$, when the learned spectral filter satisfies $\lvert\bar{\lambda}_2\rvert \geq \lvert\bar{\lambda}_1\rvert$, as this amplifies the spectral direction that carries the community information.*

Remark 6. *Graphormer-GP is the only architecture among the analysed models that can remain informative of communities when the initial kernel is uninformative (i.e., $x_0 = y_0$).*

Summary of Oversmoothing Behavior

While GCN-GP is structurally destined for rank collapse and oversmoothing, attention-based models provide mechanisms to preserve structural information. GAT-GP avoids oversmoothing under a data-dependent condition: communities need to be sufficiently separated, which is a standard expectation for graph learning tasks. Graphormer-GP relies on a prior that encodes meaningful structural information to prevent oversmoothing. In contrast, Specformer-GP offers a learnable solution, where avoiding oversmoothing is determined by the spectral weights optimized during training rather than by the data or prior. Together, these findings highlight how architectural design influence oversmoothing.

5 EXPERIMENTS

While graph transformers are notoriously difficult to analyse due to complex attention mechanisms and training dynamics, our Gaussian Process (GP) framework provides a mathematically rigorous closed-form equivalent. We provide the first kernel equivalent for graph transformers, thereby offering a tool for studying their structural properties. A structural phenomenon we investigate is **oversmoothing**. In Section 4, we formalise this behaviour by analysing GP kernels under CSBM, revealing a fundamental distinction: while **GCN-GP is destined to oversmooth** (Corollary 1), attention-based architectures can preserve discriminative community representations as depth increases. Consistent with this theory, evaluating GP kernels on synthetic and benchmark datasets (homophilic Pubmed [Sen et al., 2008] and heterophilic Chameleon [Rozemberczki et al., 2019]) shows that **GCN-GP performance deteriorates with depth**, whereas **GAT-GP remains resilient to oversmoothing** (Figure 3, left; Corollary 2). Details of the experiments are provided in Appendix F, and the code to reproduce the experiments is available at `https://github.com/MDK231/Graph-Transformer-GP`.

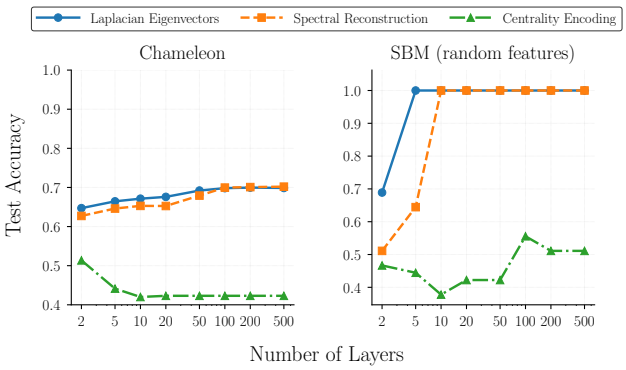

Figure 2: **Oversmoothing behaviour and the impact of positional encodings.** Test accuracy of **Graphormer-GP** on Chameleon (left) and SBM with random features (right) as a function of the number of layers. While accuracy in graph models typically suffers from performance degradation with an increasing number of layers, **Graphormer-GP exhibits increasing accuracy with depth** when utilizing informative positional encodings, such as Laplacian Eigenvectors (blue) and Spectral Reconstruction (orange). This empirical trend aligns with the theoretical results in Corollary 3.

Graphormer-GP is sensitive to structural prior. Remark 3, argues that when an informative structural prior is not explicitly known for a real-world dataset, Graphormer may still suffer from oversmoothing as the kernel collapses toward a **non-discriminative prior**. Since Graphormer can incorporate different structural priors, we ran additional experiments on SBM with random features and Chameleon. We evaluate three positional encodings (without bias) at large depth (Figure 2): **Laplacian Eigenvectors**, **Spectral Reconstruction** (low-rank Laplacian approximation), and **Centrality Encoding** (from Ying et al. [2021]). Across both datasets, Laplacian-based encodings outperform centrality encoding and resist oversmoothing. Motivated by these results, we incorporate Laplacian Eigenvectors into the other GP architectures (Figure 3, right) and find that they can mitigate the onset of oversmoothing.

6 BROADER PERSPECTIVE

The GP equivalence for graph transformers established in this work provides a foundation for a broader theoretical analysis. Analysing the model in the infinite-width regime, independent of training dynamics, yields corresponding kernel equivalents that enable the study of structural properties. Among the central questions in such analyses is expressivity, as the model's representational power can be characterised by the expressivity of the corresponding kernel. While our current analysis demonstrates community separation at large depth in the CSBM setting, it can be extended to study other critical structural phenomena, including *oversquashing* (the bottleneck of information propagation) and the model's *sensitivity to heterophily* under more complex random graph models. Furthermore, although our primary focus is on node-level kernels, our framework gives rise to edge-level kernels, opening the door to analysing edge-level tasks such as link prediction or classification. GP equivalence has already successfully enabled the study of generalisation [Seeger, 2003] and inductive biases [Li et al., 2021] in deep learning, and our results provide a basis for extending these analyses to graph transformers. From a practical perspective, an important remaining challenge is scaling these kernels to massive graph datasets, which remains non-trivial due to the non-linear, input-dependent pairwise interactions induced by the attention mechanism. Addressing this challenge would broaden the applicability of GP-based methods while preserving the theoretical insights they offer. This work provides the first unified theoretical lens based on GP for graph transformers, laying the foundation for a deeper understanding of graph attention mechanisms.

Acknowledgements

This work has been supported by the German Research Foundation (DFG) through the Priority Program SPP 2298 (project GH 257/2-2) and the research grant GH 257/4-1, and by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Research, Technology and Space.

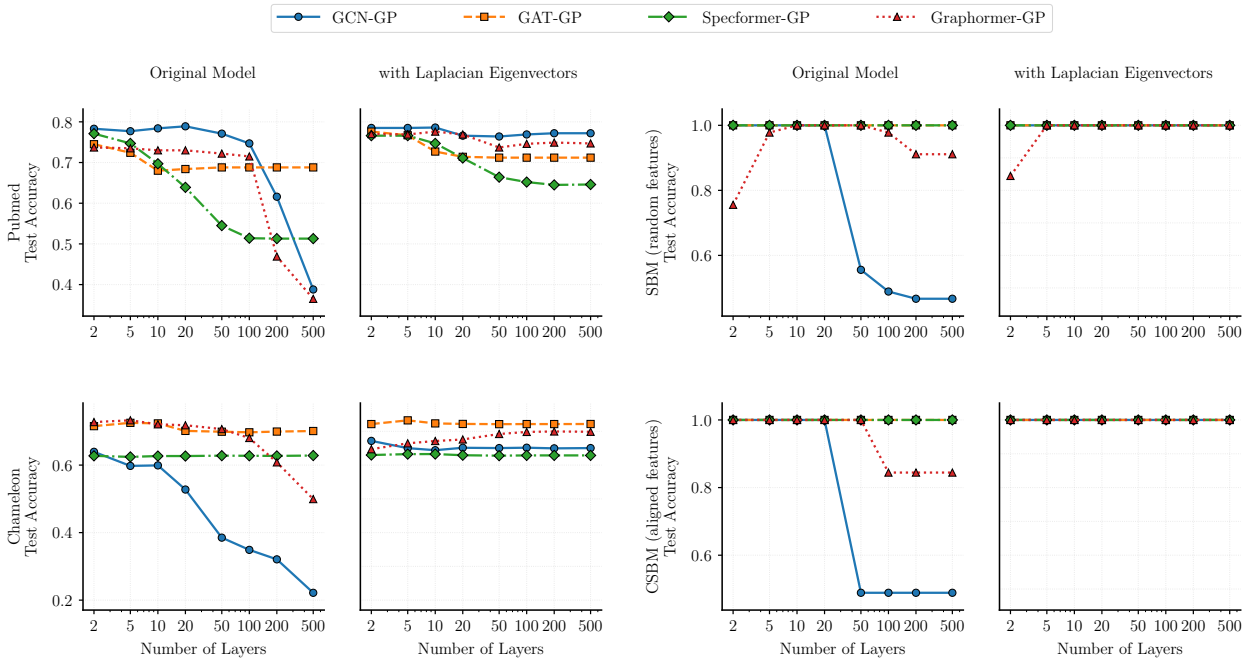

Figure 3: **Oversmoothing behaviour across various GNN-GP architectures.** Test accuracy is shown as a function of depth for each dataset, comparing the original configurations against models augmented with Laplacian Eigenvectors. In the original configurations, **GCN-GP (blue) consistently suffers from a performance drop-off as the number of layers increases**. In contrast, **GAT-GP (orange) demonstrates resilience to depth**, maintaining stable performance as argued in Corollary 2. The behavior of Specformer-GP and Graphormer-GP depends on the dataset. Notably, upon incorporating Laplacian Eigenvectors, the tendency to oversmooth is significantly mitigated across all architectures.

References

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *9th International Conference on Learning Representations, 2021*, 2021.

Nil Ayday, Mahalakshmi Sabanayagam, and Debarghya Ghoshdastidar. Why does your graph neural network fail on some graphs? insights from exact generalisation error. *CoRR*, abs/2509.10337, 2025.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, 2019.

Patrick Billingsley. *Probability and Measure*. Wiley, 2 edition, 1986.

J. R. Blum, J. Kiefer, and M. Rosenblatt. Distribution free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics*, 29, 1958.

Deyu Bo, Chuan Shi, Lele Wang, and Renjie Liao. Specformer: Spectral graph neural networks meet transformers. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023.

Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations (ICLR)*, 2019.

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: NNGP and NTK for deep attention networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research (PMLR), 2020.

Nicolas Keriven. Not too little, not too much: a theoretical analysis of graph (over)smoothing. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jascha Sohl-Dickstein, and Jeffrey Pennington. Deep neural networks as gaussian processes. In *International Conference on Learning Representations (ICLR)*, 2018a.

Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep

neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018b.

Michael Y. Li, Erin Grant, and Thomas L. Griffiths. Meta-learning inductive biases of learning systems with gaussian processes. In *Fifth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems*, 2021.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, 2018.

Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Zehao Niu, Mihai Anitescu, and Jie Chen. Graph Neural Network-Inspired Kernels for Gaussian Processes in Semi-Supervised Learning. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023.

Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations (ICLR)*, 2019.

Hoang NT and Takanori Maehara. Revisiting Graph Neural Networks: All we have is low-pass filters. *CoRR*, 2019.

Benedek Rozemberczki, Ryan Davies, Rik Sarkar, and Charles Sutton. GEMSEC: Graph embedding with self clustering. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2019.

Mahalakshmi Sabanayagam, Pascal Mattia Esser, and Debarghya Ghoshdastidar. Analysis of convolutions, non-linearity and depth in graph neural networks using neural tangent kernel. *Trans. Mach. Learn. Res.*, 2023, 2023.

Matthias W. Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *J. Mach. Learn. Res.*, 2003.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3), 2008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations*, 2018.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous Graph Attention Network. In *The world wide web conference*, 2019.

Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, 2023a.

Xinyi Wu, Zhengdao Chen, William Wei Wang, and Ali Jadbabaie. A non-asymptotic analysis of oversmoothing in graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023b.

Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel, 2019.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Bohang Zhang, Lingxiao Zhao, and Haggai Maron. On the expressive power of spectral invariant Graph Neural Networks. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Weichen Zhao, Chenguang Wang, Congying Han, and Tiande Guo. Exploring over-smoothing in graph attention networks from the markov chain perspective. In *Proceedings of the International Conference on Frontiers of Artificial Intelligence and Machine Learning, FAIML*, 2023.

Gaussian Process Limit Reveals Structural Benefits of Graph Transformers (Supplementary Material)

Nil Ayday[*1,3] **Lingchu Yang**[*2] **Debarghya Ghoshdastidar**[1,3,4]

[1]Technical University of Munich, TUM School of Computation, Information and Technology
[2]The University of Tokyo, Graduate School of Information Science and Technology
[3]Konrad Zuse School of Excellence in Reliable AI
[4]Munich Center for Machine Learning

nil.ayday@tum.de, yang-lingchu@g.ecc.u-tokyo.ac.jp, ghoshdas@cit.tum.de

A DESIGN CHOICES

A.1 GNN WITH MULTI-HEAD AGGREGATION

A common finite-head formulation aggregates $H = d^{\ell,H}$ head outputs by concatenation:

$$f^{\ell}(X) = \Big\|_{h=1}^{d^{\ell,H}} \sigma\Big(\mathbf{S}_{\mathrm{GNN}}^{(\ell,h)} f^{\ell-1}(\mathbf{X})\, \mathbf{W}^{\ell,h}\Big), \tag{8}$$

where $\|$ denotes concatenation. This formulation is not suitable for taking the limit $d^{\ell,H} \to \infty$, since concatenation leads to an ever-growing output dimension and prevents a CLT-based infinite-head analysis.

To obtain a well-defined infinite-head limit in a fixed feature dimension, we modify the aggregation mechanism by introducing a shared projection that aggregates all head-wise outputs into the next-layer representation:

$$f^{\ell}(\mathbf{X}) = \sigma\left(\left(\Big\|_{h=1}^{d^{\ell,H}} \mathbf{S}_{\mathrm{GNN}}^{(\ell,h)} f^{\ell-1}(\mathbf{X})\, \mathbf{W}^{\ell,h}\right) \mathbf{W}^{\ell,H}\right), \tag{9}$$

where $\mathbf{W}^{\ell,h}$ are head-specific weight matrices and $\mathbf{W}^{\ell,H}$ is a shared projection mapping the concatenated head outputs to the fixed-dimensional feature space of layer ℓ.

This modification plays a crucial role in the infinite-head analysis: it replaces dimension growth by a fixed-dimensional linear aggregation of head-wise random features, which enables a CLT-type argument despite the $n \times n$ stochasticity of $\mathbf{S}_{\mathrm{GNN}}^{(\ell,h)}(\mathbf{X})$ and the induced dependence across coordinates. As a result, the infinite-head GNN in (9) admits a principled Gaussian-process limit while retaining the expressive stochastic message-passing structure encoded by $\mathbf{S}_{\mathrm{GNN}}^{(\ell,h)}$.

A.2 GRAPHORMER

In the original Graphormer architecture, structural information is injected through a *centrality encoding*, where each node representation is augmented by learnable embeddings indexed by its in-degree and out-degree. Using our notation, the input node representation can be written as

$$f_i^{(0)}(X) = x_i + z_{\deg^-}(v_i) + z_{\deg^+}(v_i), \tag{10}$$

where x_i denotes the raw node feature and $z_{\deg^-}, z_{\deg^+} \in \mathbb{R}^d$ are learnable degree-based embeddings. For undirected graphs, the two degree terms are unified into a single degree encoding.

[*]Equal contribution.

While such a design effectively injects node importance signals into the attention mechanism and has demonstrated strong empirical performance on graph-level tasks, it represents a specific instance of structural augmentation based solely on degree statistics. This degree-centric formulation becomes restrictive when considering node-level prediction tasks or when aiming to incorporate richer and more general notions of graph structure.

To generalize this mechanism, we reinterpret centrality encoding as a special case of feature-level graph encoding insertion. Instead of restricting the structural signal to degree-based embeddings, we introduce a general graph encoding term P_i^ℓ at layer ℓ, and define the augmented node representation as

$$\tilde{f}_i^\ell(X) := \big[f_i^\ell(X),\ P_i^\ell\big], \qquad i \in V, \tag{11}$$

where $f_i^\ell(X)$ denotes the node representation produced by the backbone network at layer ℓ, and $P_i^\ell \in \mathbb{R}^{d_p}$ encodes structural or positional information associated with node v_i. More importantly, (11) allows P_i^ℓ to represent arbitrary graph encodings, including but not limited to Laplacian positional encodings, random walk-based encodings, or other task-specific structural descriptors.

By explicitly separating semantic features $f_i^\ell(X)$ from structural encodings P_i^ℓ, the proposed formulation provides a flexible and general mechanism for injecting graph structure into the model. This is particularly advantageous for node-level tasks, where the prediction target depends on fine-grained local and mesoscopic structural patterns rather than global graph summaries. Furthermore, the concatenation-based design preserves the modularity of the architecture and facilitates theoretical analysis, as different choices of P_i^ℓ can be studied independently of the backbone representation.

In summary, we treat centrality encoding as a special case of a more general graph encoding framework. The unified representation (11) enables the incorporation of diverse structural signals in a principled manner, while maintaining compatibility with Transformer-based architectures and node-level prediction settings.

A.3 SPECFORMER

Spectral Graph Construction in Specformer. In the original Specformer model, the graph structure is learned in the spectral domain through a set of spectral tokens. Let $H^t \in \mathbb{R}^{n \times d^{\ell,H}}$ denote the spectral tokens at layer t in eigenvalue encoding, where each column corresponds to the output of one attention head. At layer t, each attention head $h = 1, \ldots, d^{\ell,H}$ produces a separate spectral representation via

$$Z_h = \mathrm{Attention}\big(H^t W^{Q,th},\ H^t W^{K,th},\ H^t W^{V,th}\big), \tag{12}$$

where $W^{Q,th}$, $W^{K,th}$, and $W^{V,th}$ are learnable projection matrices. The head-specific representation Z_h is then mapped to a vector of filtered eigenvalues,

$$\lambda_h = \sigma(Z_h W_\lambda) \in \mathbb{R}^N, \tag{13}$$

where $\sigma(\cdot)$ denotes an activation function. Each eigenvalue vector λ_h defines a spectral graph operator

$$S_h = U\,\mathrm{diag}(\lambda_h)\,U^\top, \tag{14}$$

with U denoting the eigenvector matrix of the graph Laplacian.

As a result, the $d^{\ell,H}$ attention heads produce $d^{\ell,H}$ distinct spectral graph operators $\{S_h\}_{h=1}^{d^{t,H}}$. These operators are subsequently used as intermediate bases to construct feature-wise graph structures. Specifically, the operators are concatenated along the channel dimension and mapped to the node feature dimension d through a feed-forward network (FFN),

$$\hat{S} = \mathrm{FFN}\big([\,I_N \,\|\, S_1 \,\|\, \cdots \,\|\, S_{d^{t,H}}\,]\big) \ \in\ \mathbb{R}^{N \times N \times d}. \tag{15}$$

The output $\hat{S}$ thus consists of d graph operators, where each slice $\hat{S}_{:,:,i}$ corresponds to a distinct adjacency (or Laplacian) matrix associated with the i-th node feature dimension.

Based on these feature-wise graph operators, graph convolution is performed independently for each feature channel. Let $f(X)^{(\ell-1)} \in \mathbb{R}^{N \times d}$ denote the node representations at layer $\ell - 1$. For the i-th feature dimension, the propagation is given by

$$\hat{f}(X)_{:,i}^{(\ell-1)} = \hat{S}_{:,:,i}\, f(X)_{:,i}^{(\ell-1)}, \tag{16}$$

Table 4: Summary of the Gaussian weight priors used for each graph transformer in the GP analysis. These priors define the variance of the initial weights for all linear transformations and attention parameters.

Model	Weight Priors (Variance)
GAT	$W_{ij}^{\ell,H} \sim \mathcal{N}\Big(0, \frac{\sigma_H^2}{d^{\ell,H} d_{\ell-1}}\Big), W_{ij}^{\ell,h} \sim \mathcal{N}\Big(0, \frac{\sigma_w^2}{d_{\ell-1}}\Big), \vec{v}_k^{\ell h} \sim \mathcal{N}\Big(0, \frac{\sigma_v^2}{d_{\ell-1}}\Big)$
Graphormer	$W_{ij}^{Q,\ell h} \sim \mathcal{N}\Big(0, \frac{\sigma_Q^2}{d_{\ell-1}}\Big), W_{ij}^{K,\ell h} \sim \mathcal{N}\Big(0, \frac{\sigma_K^2}{d_{\ell-1}}\Big),$
	$W_{ij}^{\ell,H} \sim \mathcal{N}\Big(0, \frac{\sigma_H^2}{d^{\ell,H} d_{\ell-1}}\Big), W_{ij}^{V,\ell h} \sim \mathcal{N}\Big(0, \frac{\sigma_w^2}{d_{\ell-1}}\Big), b_\rho(i,j) \sim \mathcal{N}(0, \sigma_b^2)$
Specformer	$W_{ij}^{Q,tkh} \sim \mathcal{N}(0, \sigma_Q^2/d_{t-1}), \quad W_{ij}^{K,tkh} \sim \mathcal{N}(0, \sigma_K^2/d_{t-1}),$
	$W_{ij}^{V,tkh} \sim \mathcal{N}(0, \sigma_V^2/d_{t-1}), W_{ij}^{th,H} \sim \mathcal{N}\Big(0, \frac{\sigma_O^2}{d^{t,H} d_{t-1}}\Big),$

and the node representations are updated as

$$f(X)^{(\ell)} = \sigma\Big(\hat{f}(X)^{(\ell-1)} W^{(\ell-1)}\Big). \tag{17}$$

While this construction is expressive, it introduces a fundamental statistical issue. Since the d graph operators $\{\hat{S}_{:,:,i}\}_{i=1}^{d}$ are jointly generated through a shared FFN, they are statistically dependent. Consequently, the feature dimensions of $\hat{f}^{(\ell)}(X)$ are no longer independent, violating the independence assumptions required by the Central Limit Theorem (CLT). As a result, the original Specformer formulation does not satisfy the conditions necessary for CLT-based infinite-width analysis.

Multi-head Aggregated Spectral Construction. To address this issue, for each head of graph convolution layer heads $d^{\ell,H}$, we modify the spectral construction by employing multi-head attention to generate a single set of spectral coefficients, rather than one per feature dimension. Specifically, at each layer t, the outputs of the $d^{t,H}$ spectral heads are aggregated prior to spectral filtering. The spectral tokens are updated as

$$H^{t+1} = \Big(\|_{h=1}^{d^{t,H}} S_{\text{Spec}}^{(t,h)}(H^t)\ H^t W^{V,th}\Big) W^{H,t}, \tag{18}$$

where $W^{H,t}$ is a learnable projection matrix.

After T eigenvalue encoding layers, we obtain the final spectral tokens $\bar{H} = H^T$. Based on $\bar{H}$, a single set of spectral coefficients is generated as

$$\bar{\lambda} = \sigma\big(\bar{H} W_\lambda\big) \in \mathbb{R}^n, \tag{19}$$

which defines the final graph convolution operator

$$S_{\text{Specformer}} = U\,\text{diag}(\bar{\lambda})\,U^\top \in \mathbb{R}^{n\times n}. \tag{20}$$

By constructing a shared spectral operator across feature dimensions, this formulation preserves feature-wise independence and restores the conditions required for CLT-based infinite-width analysis.

B FORMAL THEOREMS ON GAUSSIAN PROCESS LIMITS OF GRAPH TRANSFORMERS

In this Appendix, we present the formal theorems and kernel recursions establishing the Gaussian process limits for graph transformers, derived under the independent Gaussian weight priors summarized in Table 4.

Graph Attention Network (GAT) In GAT, attention is computed locally over graph neighborhoods, and in the infinite-width and infinite-head limit, both the attention logits and node pre-activations converge to Gaussian processes. The resulting kernel recursion reveals how features and neighborhood structure jointly shape representation propagation.

Theorem 7. *(Infinite-width and infinite-head limit of a GAT layer) Assume that the parameters of the ℓ-th GAT layer satisfy $W_{ij}^{\ell,H} \sim \mathcal{N}\Big(0, \frac{\sigma_H^2}{d^{\ell,H} d_{\ell-1}}\Big), W_{ij}^{\ell,h} \sim \mathcal{N}\Big(0, \frac{\sigma_w^2}{d_{\ell-1}}\Big), \vec{v}_k^{\ell h} \sim \mathcal{N}\Big(0, \frac{\sigma_v^2}{d_{\ell-1}}\Big)$, independently for all i, j, k and heads h, then, as $d_\ell, d^{\ell,H} \to \infty$, the following holds:*

(I) $E^{\ell}(X) = \{E^{(\ell h)}(X) : h \in \mathbb{N}\}$ *converges in distribution to* $E^{\ell}(X) \sim \mathcal{GP}(0, K^{(\ell),E})$ *with*

$$\begin{aligned} K^{(\ell h),E}_{ai,bj} &= \mathbb{E}\big[E^{(\ell h)}_{ai}(X)\, E^{(\ell h')}_{bj}(X)\big] \\ &= \delta_{h=h'}\, \sigma_w^2 \sigma_v^2 \big(K^{(\ell-1)}_{ab} + K^{(\ell-1)}_{ij}\big), \end{aligned} \tag{21}$$

(II) $g^{\ell}(X)$ *converges in distribution to* $g^{\ell}(X) \sim \mathcal{GP}(0, \Sigma^{(\ell)})$ *with*

$$\begin{aligned} C^{(\ell)}_{ai,bj} &:= \mathbb{E}\Big[\big(S^{(\ell,1)}_{\mathrm{GAT}}(X)\big)_{ai}\, \big(S^{(\ell,1)}_{\mathrm{GAT}}(X)\big)_{bj}\Big] \\ &= \mathbb{E}\Big[\phi\big(\sigma(E^{(\ell 1)}_{ai}(X))\big)\, \phi\big(\sigma(E^{(\ell 1)}_{bj}(X))\big)\Big], \end{aligned} \tag{22}$$

$$\begin{aligned} \Sigma^{(\ell)}_{ab} &:= \mathbb{E}\big[g^{\ell}_{a1}(X)\, g^{\ell}_{b1}(X)\big] \\ &= \sigma_H^2 \sigma_w^2 \sum_{i \in N_a} \sum_{j \in N_b} K^{(\ell-1)}_{ij}\, C^{(\ell)}_{ai,bj}, \end{aligned} \tag{23}$$

where $a, b \in V$ *and* $(a, i), (b, j) \in E$.

Theorem 7 shows that, for each attention head, the collection of attention logits over edges converges in distribution to a centered Gaussian process, with different heads becoming independent in the limit. Equation (21) characterizes the covariance between attention logits on two edges (a, i) and (b, j), which is driven by the similarity between the two source nodes a, b together with the similarity between the two target nodes i, j, so edges whose endpoints are simultaneously close in representation space exhibit stronger correlation. Using Equation (21), the theorem then derives the Gaussian process limit of the pre-activations. Equation (23) describes how node-level dependence propagates to the next layer, where the covariance between the pre-activations of nodes a and b at layer ℓ is obtained by aggregating contributions from all neighbor pairs (i, j) with $i \in \mathcal{N}_a$ and $j \in \mathcal{N}_b$, combining similarity among neighboring node representations from the previous layer with the alignment induced by attention on the associated edges. Together, these results yield a kernel recursion that describes how feature and graph govern representation propagation across layers. Theorem 7 establishes a general kernel recursion, and to obtain a closed-form recursion, we next consider specific choices of the attention and activation nonlinearities.

Corollary 8. *(Closed-form GAT kernel recursion) Under the assumptions of Theorem 7, for* $\phi(x) = x$ *and ReLU activation* $\sigma(\cdot)$*, the node-level covariance kernel at layer* ℓ *admits the closed-form expression*

$$\begin{aligned} \Sigma^{(\ell)}_{ab} = \frac{\sigma_H^2 \sigma_w^4 \sigma_v^2}{2\pi} &\sum_{i \in N_a} \sum_{j \in N_b} K^{(\ell-1)}_{ij} \\ &\times \sqrt{\big(K^{(\ell-1)}_{aa} + K^{(\ell-1)}_{ii}\big)\big(K^{(\ell-1)}_{bb} + K^{(\ell-1)}_{jj}\big)} \\ &\times \Big(\sin\theta^{(\ell-1)}_{ai,bj} + \big(\pi - \theta^{(\ell-1)}_{ai,bj}\big)\cos\theta^{(\ell-1)}_{ai,bj}\Big). \end{aligned} \tag{24}$$

$$\theta^{(\ell-1)}_{ai,bj} = \arccos\left(\frac{K^{(\ell-1)}_{ab} + K^{(\ell-1)}_{ij}}{\sqrt{\big(K^{(\ell-1)}_{aa} + K^{(\ell-1)}_{ii}\big)\big(K^{(\ell-1)}_{bb} + K^{(\ell-1)}_{jj}\big)}}\right). \tag{25}$$

Kernel in Corollary 8 induces a quadratic computational cost over neighborhood pairs. To further simplify the recursion, we consider a linear activation.

Corollary 9 (Linear GAT kernel)**.** *Under the assumptions of Corollary 8, and for* $\sigma(x) = x$*, the node-level kernel admits the following simplified closed-form recursion:*

$$\begin{aligned} \Sigma^{(\ell)} = \sigma_H^2\, \sigma_w^4\, \sigma_v^2 \Big(&K^{(\ell-1)} \odot \big(A K^{(\ell-1)} A^{\top}\big) \\ &+ A\big(K^{(\ell-1)} \odot K^{(\ell-1)}\big) A^{\top}\Big). \end{aligned}$$

Graphormer As introduced in Section 2, Graphormer incorporates graph information through a positional encoding mechanism. In the infinite-width limit, the output converges in distribution to a Gaussian process whose kernel satisfies

$$\tilde{K}^{(\ell-1)} \longmapsto \alpha\, K^{(\ell-1)} + (1-\alpha)\, R^{(\ell)} \quad , \alpha \in (0,1). \tag{26}$$

Where the matrix R can be interpreted as the covariance matrix of the corresponding positional encoding which can be computed as an inner product of positional embeddings, i.e., $R_{ab}^{(\ell)} = \langle P_a^{(\ell)}, P_b^{(\ell)} \rangle$. Alternatively, instead of introducing explicit positional embeddings, the graph structure itself can be directly leveraged to construct a node-level graph-structure covariance matrix. Theorem 10 characterizes the Gaussian process limit of a Graphormer layer by establishing joint limits for both the attention logits and the node pre-activations. Although the proof strategy is analogous to GAT, the incorporation of positional encodings and how the attention scores are computed modifies the kernel.

Centrality Encoding (CE) differs from other positional encoding that it constructs a learnable embedding based on node degrees. Specifically, a CE parameter matrix has size $\mathbf{P}_{\mathrm{CE}} \in \mathbb{R}^{(\texttt{max\_degree}+1)\times d}$, and each entry is initialized as $(\mathbf{P}_{\mathrm{CE}})_{ij} \sim \mathcal{N}\Big(0, \frac{\sigma_{\mathrm{CE}}^2}{d}\Big)$. Since we only consider undirected graphs in this paper, nodes with the same degree share the same embedding vector, i.e., $P_v = \mathbf{P}_{\mathrm{CE}}[\deg(v)] \in \mathbb{R}^d$. Under this construction, the embedding correlation satisfies

$$\mathbb{E}\big[P_a^\top P_b\big] = \begin{cases} \sigma_{\mathrm{CE}}^2, & \deg(a) = \deg(b), \\ 0, & \deg(a) \neq \deg(b). \end{cases}$$

Therefore, we obtain

$$R_{ab} = \begin{cases} \sigma_{\mathrm{CE}}^2, & \deg(a) = \deg(b), \\ 0, & \deg(a) \neq \deg(b). \end{cases}$$

Theorem 10. *(Infinite-width and infinite-head limit of a Graphormer layer) Assume the parameters of the ℓ-th Graphormer layer satisfy $W_{ij}^{Q,\ell h} \sim \mathcal{N}\Big(0, \frac{\sigma_Q^2}{d_{\ell-1}}\Big), W_{ij}^{K,\ell h} \sim \mathcal{N}\Big(0, \frac{\sigma_K^2}{d_{\ell-1}}\Big), W_{ij}^{\ell,H} \sim \mathcal{N}\Big(0, \frac{\sigma_H^2}{d^{\ell,H} d_{\ell-1}}\Big), W_{ij}^{V,\ell h} \sim \mathcal{N}\Big(0, \frac{\sigma_w^2}{d_{\ell-1}}\Big), b_\phi(i,j) \sim \mathcal{N}(0, \sigma_b^2)$, independently for all i, j, heads h, and structural indices $\phi(i,j)$, then, as $d_\ell, d^{\ell,H} \to \infty$, the following holds:*

(I) $E^\ell(X) = \big\{E^{(\ell h)}(X) : h \in \mathbb{N}\big\}$ *converges in distribution to* $E^{\ell-1}(X) \sim \mathcal{GP}(0, K^{(\ell),E})$ *with*

$$\begin{aligned} K_{ai,bj}^{(\ell h),E} &= \mathbb{E}\big[E_{ai}^{(\ell h)}(X)\, E_{bj}^{(\ell h')}(X)\big] \\ &= \delta_{h=h'}\Big(\sigma_Q^2 \sigma_K^2\, \tilde{K}_{ab}^{(\ell-1)}\, \tilde{K}_{ij}^{(\ell-1)} + \sigma_b^2\, \mathbf{1}_{\phi(a,i)=\phi(b,j)}\Big). \end{aligned} \tag{27}$$

(II) $g^\ell(X)$ *converges in distribution to* $g^\ell(X) \sim \mathcal{GP}(0, \Sigma^{(\ell)})$ *with*

$$\begin{aligned} C_{ai,bj}^{(\ell)} &:= \mathbb{E}\Big[\big(S_{\mathrm{Graphormer}}^{(\ell,1)}(X)\big)_{ai}\, \big(S_{\mathrm{Graphormer}}^{(\ell,1)}(X)\big)_{bj}\Big] \\ &= \mathbb{E}\Big[\phi\big(E_{ai}^{(\ell 1)}(X)\big)\, \phi\big(E_{bj}^{(\ell 1)}(X)\big)\Big], \end{aligned} \tag{28}$$

$$\Sigma_{ab}^{(\ell)} := \mathbb{E}\big[g_{a1}^\ell(X)\, g_{b1}^\ell(X)\big] = \sigma_H^2 \sigma_w^2 \sum_{i\in V} \sum_{j\in V} \tilde{K}_{ij}^{(\ell-1)}\, C_{ai,bj}^{(\ell)}, \tag{29}$$

where $a, b, i, j \in V$.

Equation (26) shows that the effective similarity between nodes depends on both the similarity of node features and the similarity of graph structural features encoded by the positional information. Equation (27) further indicates that attention correlations are strengthened for node pairs sharing the same structural relation, ensuring that graph information is explicitly leveraged in guiding attention. Equation (29) then propagates the attention correlations to the next layer through an aggregation over all node pairs, reflecting that, unlike GAT which is restricted to local neighborhoods, Graphormer induces similarity connections between every pair of nodes in the graph. The following corollary specializes Theorem 10 to linear attention.

Corollary 11 (Closed-form Graphormer kernel under linear attention)**.** *Under the assumptions of Theorem 10, and for $\phi(x) = x$, Consequently, the node-level covariance kernel at layer ℓ admits the closed-form expression*

$$\Sigma^{(\ell)}_{ab} = \sigma_H^2\sigma_w^2\Bigg[\sigma_Q^2\sigma_K^2\,\tilde{K}^{(\ell-1)}_{ab}\sum_{i,j\in V}\tilde{K}^{(\ell-1)}_{ij}\,\tilde{K}^{(\ell-1)}_{ij}+$$
$$\sigma_b^2\sum_{i,j\in V}\tilde{K}^{(\ell-1)}_{ij}\,\mathbf{1}_{\phi(a,i)=\phi(b,j)}\Bigg], \text{where } a,b,i,j\in V.$$

Specformer. As detailed in Section 2, the convolution in Specformer is composed of *spectral tokens* and *node features*. The spectral tokens are first processed through T layers of encoding, and are subsequently passed through a decoder to construct a data-dependent spectral filtering that governs the propagation of node features. The following theorem characterizes the behavior of Specformer in the infinite-width and infinite-head limit. In this regime, (I) the spectral-token representations H^t, which are updated through attention, and (II) the node-feature pre-activations $g(X)$ converge in distribution to a Gaussian process.

Theorem 12. *(Infinite-width and infinite-head limit of Specformer) If the parameters of the t-st token layer satisfy $W^{Q,tkh}_{ij}\sim\mathcal{N}(0,\sigma_Q^2/d_{t-1})$, $W^{K,tkh}_{ij}\sim\mathcal{N}(0,\sigma_K^2/d_{t-1})$, $W^{V,tkh}_{ij}\sim\mathcal{N}(0,\sigma_V^2/d_{t-1})$ and $W^{th,H}_{ij}\sim\mathcal{N}\Big(0,\frac{\sigma_O^2}{d^{t,H}d_{t-1}}\Big)$, independently for all indices, token heads k, and spectral heads h, then, as $d_t, d^{t,H}\to\infty$, the following holds.*

(I) H^t converges in distribution to $H^t\sim\mathcal{GP}(0,K_H^{(t)})$ with

$$\begin{aligned}K^{(t,k,h),E}_{ik,jl} &= \mathbb{E}\Big[E^{(t,k,h)}_{ij}(H)\,E^{(t,k',h')}_{kl}(H')\Big]\\ &= \delta_{k=k'}\,\delta_{h=h'}\,\sigma_Q^2\sigma_K^2\,K^{(t-1)}_{H,ik}\,K^{(t-1)}_{H,jl},\end{aligned}\tag{30}$$

$$\begin{aligned}C^{(t,h)}_{ik,jl}(H,H') &:= \delta_{h=h'}\,\mathbb{E}\Big[\big(S^{(t,1,h)}_{\mathrm{Spec}}(H)\big)_{ik}\,\big(S^{(t,1,h')}_{\mathrm{Spec}}(H')\big)_{jl}\Big]\\ &= \delta_{h=h'}\,\mathbb{E}\Big[\phi\big(E^{(t,1,h)}_{ik}(H)\big)\,\phi\big(E^{(t,1,hi)}_{jl}(H')\big)\Big],\end{aligned}\tag{31}$$

$$\begin{aligned}K^{(t,h)}_{H,ij} &:= \mathbb{E}\big[(H^{t,h})_i(H^{t,h'})_j\big]\\ &= \delta_{h=h'}\,\big(\sigma_O^2\sigma_V^2\sum_{k,l\in V}K^{(t-1)}_{H,kl}\,C^{(t,h)}_{ik,jl}(H,H')\big),\end{aligned}\tag{32}$$

where $i,k,j,l\in V$ and $K^{(0,h)}_{H,ab}:=\frac{H^0_a\cdot H^0_b}{d}$ with d as the spectral-token embedding dimension.

(II) If the parameters of the ℓ-st graph convolution layer satisfy $W_{\ell,ij}\sim\mathcal{N}(0,\sigma_w^2/d_{\ell-1}), W^{\ell,H}_{ij}\sim\mathcal{N}\Big(0,\frac{\sigma_H^2}{d^{\ell,H}d_{\ell-1}}\Big)$ and the eigenvalue decoder layers satisfy $W^{(\ell,h)}_{\lambda,ij}\sim\mathcal{N}(0,\sigma_\lambda^2/d_T)$ independently for all indices and spectral heads h, then, as $d_\ell, d_T, d^{\ell,H}\to\infty$, $g^\ell(X)$ converges in distribution to $\mathcal{GP}(0,\Sigma^{(\ell)})$ with

$$\begin{aligned}K_{\lambda,ab} &:= \delta_{h=h'}\,\mathbb{E}\big[\bar{\lambda}^{(h)}_a(H)\,\bar{\lambda}^{(h')}_b(H')\big]\\ &= \delta_{h=h'}\,\mathbb{E}\Big[\sigma\big((\bar{H}^{(h)}W^{(\ell,h)}_\lambda)_a\big)\,\sigma\big((\bar{H}^{(h')}W^{(\ell,h)}_\lambda)_b\big)\Big],\end{aligned}\tag{33}$$

$$\Sigma^{(\ell)}_{ab} = \sigma_H^2\sigma_w^2 U\Big(K_\lambda\odot\big(U^\top K^{(\ell-1)}U\big)\Big)U^\top,\tag{34}$$

where $a,b\in V$.

Equations (30)–(32) describe how the covariance induced by the previous token layer is recursively transformed through a standard transformer. At the terminal token layer, Equation (33) transfers the resulting token-level covariance to the spectral domain by inducing a kernel over the learned spectral filter coefficients. Equation (34) then defines the recursion of the node-feature kernel: the previous-layer node covariance is mapped into the Laplacian eigenbasis, modulated elementwise according to the induced spectral kernel, and projected back to the node domain. Together, these equations specify a coupled kernel recursion in which attention governs spectral reweighting, and graph convolution propagates the resulting covariance across layers.

C PROOFS OF THEOREMS 7, 10, 12, 35

Proof technique. Our analysis follows the standard induction-on-layers framework for establishing Gaussian process (GP) limits of randomly initialized deep neural architectures (Matthews et al. [2018], Lee et al. [2018a], Novak et al. [2019], Garriga-Alonso et al. [2019], Yang [2019], Hron et al. [2020]). When the layer-$(\ell-1)$ representations converge in distribution to a Gaussian process, we show that the layer-ℓ representations also converge to a Gaussian process under suitable initialization and moment conditions.

A key methodological step of this work is to develop a unified convergence framework for graph transformers in the infinite-head regime. At a high level, our strategy is to establish a single infinite-head convergence theorem that yields a Gaussian-process limit for the model outputs, while making explicit the analytic conditions required for interchanging limits and expectations.

Concretely, Appendix C.1 proves a general infinite-head convergence result for multi-head attention aggregation layers, covering all four architectures studied in this paper within one theorem. Importantly, this output-level convergence in the infinite-head limit is not obtained in isolation: the proof requires distributional convergence of the corresponding graph convolution operators, together with uniform integrability, which ensures convergence of expectations and justifies interchanging limits and $E[\cdot]$. Therefore, we separately establish the convergence of the graph convolution operators for each model Appendix C.2, C.3, C.4, and D. These operator-level results, together with uniform integrability (Lemma 28), yield convergence of expectations via Theorem 29. Combining these ingredients completes a single coherent route to the final, uniform convergence statements for all models under infinitely many heads.

Technically, the core convergence arguments proceed by reducing finite-dimensional distributional claims to one-dimensional linear projections via the Cramér–Wold device (Lemma 27), followed by a central limit theorem for exchangeable triangular arrays (Lemma 13) applied at the multi-head aggregation level. Slutsky's lemma (Lemma 30) and the continuous mapping theorem are then used to combine the convergent random components. Finally, passage of limits through expectations is justified by uniform integrability (Lemma 28) together with the convergence-of-expectations theorem (Theorem 29), which is enabled precisely by the integrability properties established through the graph convolution operator convergence in Appendix C.2–D.

C.1 CONVERGENCE OF $g^{\ell}(X)$

Since all this four models require the number of attention heads to grow to infinity in order to admit a well-defined Gaussian Process (GP) limit, we adopt a unified and general analytical framework to establish their GP convergence.

We analyze the multi-head aggregation at a fixed layer index ℓ, mapping the GP input $f^{\ell-1}$ to the GP output f^{ℓ}. For each head $h \in \{1, \ldots, d^{\ell,H}\}$, node $a \in V$ and input $X \in \mathcal{X}_0$, define the single-head pre-activation output at layer ℓ by

$$g^{\ell h}(X) := S_{\mathrm{GNN}}^{(\ell,h)} f^{\ell-1}(X)\, W_{\ell,h}, \tag{35}$$

The multi-head output at layer ℓ is given by

$$g^{\ell}(X) = \left[g^{\ell 1}(X), \ldots, g^{\ell d^{\ell,H}}(X)\right] W^{\ell,H}, \tag{36}$$

where $W^{\ell,H}$ is the output projection matrix, whose columns we write as

$$W^{\ell,H} = \left[\, w_1^{\ell,H}, \ldots, w_{d^{\ell,H}}^{\ell,H} \,\right],$$

with $w_h^{\ell,H}$ corresponding to head h. The variance of $W^{\ell,H}$ is chosen such that the overall scaling corresponds to a properly rescaled weighted sum over the $d^{\ell,H}$ attention heads.

Lemma 13 (Adaptation of Theorem 2 from Blum et al., 1958)**.** *For each $m \in \mathbb{N}$, let $\{X_{m,i} : i = 1, 2, \ldots\}$ be an infinitely exchangeable sequence with $\mathbb{E}[X_{n,1}] = 0$ and $\mathbb{E}[X_{m,1}^2] = \sigma_n^2$, such that $\lim_{m\to\infty} \sigma_m^2 = \sigma^2$ for some $\sigma^2 \geq 0$. Let*

$$S_n = \frac{1}{\sqrt{d_m}} \sum_{i=1}^{d_m} X_{m,i}, \tag{37}$$

for some sequence $(d_m)_{m\geq 1} \subset \mathbb{N}$ *such that* $\lim_{m\to\infty} d_m = \infty$*. Assume:*

$$(a)\ \mathbb{E}[X_{m,1}X_{m,2}] = 0,$$
$$(b)\ \lim_{n\to\infty} \mathbb{E}[X_{m,1}^2 X_{m,2}^2] = \sigma^4,$$
$$(c)\ \mathbb{E}|X_{m,1}|^3 = o(\sqrt{d_m}).$$

Then $S_n \xrightarrow{d} Z$*, where* $Z = 0$ *if* $\sigma^2 = 0$*, and* $Z \sim \mathcal{N}(0, \sigma^2)$ *otherwise.*

Lemma 14 (Node-level GP limit)**.** *Let the assumptions of Theorem 7 , Theorem 10, Theorem 12 and Theorem 35 hold. The collection of layer-*ℓ *pre-activation outputs*

$$g^\ell(X) := \left\{ g_a^\ell(X) : X \in \mathcal{X}_0,\ a \in V \right\}$$

converges in distribution to a centred Gaussian process whose finite-dimensional marginals are Gaussian with covariance given by the node-level kernel $K^{(\ell)}$*.*

Proof. Following the Cramér–Wold device [Billingsley, 1986, p. 383], it suffices to establish convergence of one-dimensional projections of finite-dimensional marginals of $g^\ell(X)$.

Fix a finite index set $\mathcal{L} \subset \mathcal{X}_0 \times V$ and test vectors $\{\alpha^{X,a} \in \mathbb{R}^N : (X,a) \in \mathcal{L}\}$. Consider the linear statistic

$$T^{(\ell)} := \sum_{(X,a)\in\mathcal{L}} (\alpha^{X,a})^\top g_a^\ell(X). \tag{38}$$

Using the concatenation form (36) and writing $g_a^{(\ell+1)h}(X)$ for the a-th row of the h-th head, we obtain

$$\begin{aligned} T^{(\ell)} &= \sum_{(X,a)\in\mathcal{L}} (\alpha^{X,a})^\top \Big(\sum_{h=1}^{d^{\ell,H}} g_a^{\ell h}(X)\, w_h^{\ell,H} \Big) \\ &= \sum_{h=1}^{d^{\ell,H}} \underbrace{\sum_{(X,a)\in\mathcal{L}} (\alpha^{X,a})^\top g_a^{\ell h}(X)\, w_h^{\ell,H}}_{=:\,\zeta_h}. \end{aligned}$$

Define the rescaled head-level variables

$$\psi_h := \sqrt{d^{\ell,H}}\,\zeta_h, \qquad h = 1, \dots, d^{\ell,H},$$

so that

$$T^{(\ell)} = \frac{1}{\sqrt{d^{\ell,H}}} \sum_{h=1}^{d^{\ell,H}} \psi_h. \tag{39}$$

For each head width $d^{\ell,H}$, this defines a triangular array

$$\{\psi_{d^{\ell,H},h}\}_{h=1}^{d^{\ell,H}},$$

where, suppressing the explicit dependence on $d^{\ell,H}$ for notational simplicity, each ψ_h admits the representation

$$\psi_h = \sqrt{d^{\ell,H}} \sum_{(X,a)\in\mathcal{L}} (\alpha^{X,a})^\top g_a^{\ell h}(X)\, w_h^{\ell,H}. \tag{40}$$

Thus $T^{(\ell)}$ is of the form (37) in Lemma 13, with $X_{n,i}$ identified with ψ_h and d_n with $d^{\ell,H}$. To invoke Lemma 13, it suffices to verify conditions (a)–(c) for the array $\{\psi_h\}_{h=1}^{d^{\ell,H}}$, which is achieved by the following lemmas:

- Exchangeability of $\{\psi_h\}$ in the head index h is established in Lemma 15.

- Zero mean and vanishing cross-covariance, $\mathbb{E}[\psi_1] = 0$ and $\mathbb{E}[\psi_1\psi_2] = 0$, follow from Lemma 16.
- Convergence of the variance, $\mathbb{E}[\psi_1^2] \to \tau^2$, is proved in Lemma 17.
- Convergence of the mixed fourth moment $\mathbb{E}[\psi_1^2\psi_2^2] \to \tau^4$ and the growth condition $\mathbb{E}|\psi_1|^3 = o(\sqrt{d^{\ell,H}})$ are established in Lemma 18.

Therefore, by Lemma 13, $T^{(\ell)}$ converges in distribution to a centred Gaussian random variable with variance τ^2. Since $\mathcal{L}$ and the test vectors $\{\alpha^{X,a}\}$ are arbitrary, the finite-dimensional distributions of $g^\ell(X)$ converge to those of a Gaussian process with covariance given by the node-level kernel $K^{(\ell)}$. □

Lemma 15 (Head-wise exchangeability)**.** *Under the assumptions of Theorem 7, the random variables $\{\psi_h\}_{h=1}^{d^{\ell,H}}$ are exchangeable over the head index h.*

Proof. In all four of our models, for distinct heads h, the parameters

$$\{W_{\ell,h},\, w_h^{\ell,H}\}$$

(including those appearing in $S_{\mathrm{GNN}}^{(\ell,h)}$) are i.i.d. across h and independent of all previous-layer variables. Conditioning on the previous-layer representation $f^{\ell-1}(\cdot)$ (i.e., fixing $\{f^{\ell-1}(x) : x \in \mathcal{X}\}$), (40) implies that $(\psi_1, \ldots, \psi_{d^{\ell,H}})$ are i.i.d. in h, and hence their conditional joint distribution is invariant under any permutation of the head indices. Therefore, by de Finetti's theorem, $\{\psi_h\}_{h\in\mathbb{N}}$ is infinitely exchangeable.

□

Lemma 16 (Head-wise zero mean and vanishing cross-covariance)**.** $\mathbb{E}[\psi_1] = \mathbb{E}[\psi_1\psi_2] = 0$.

Proof. From (40),

$$\psi_1 = \sum_{(X,a)\in\mathcal{L}} \sqrt{d^{\ell,H}}(\alpha^{X,a})^\top \left(S_{\mathrm{GNN}}^{(\ell,1)} f^\ell(X)\right)_a w_1^{\ell,H}.$$

Condition on the σ-algebra generated by $\{f^\ell(X) : X \in \mathcal{X}_0\}$ and $\{S_{\mathrm{GNN}}^{(\ell,h)} : h \in \mathbb{N}\}$. Under this conditioning, $w_1^{\ell,H}$ is independent of all conditioned variables and is centred Gaussian. Therefore,

$$\mathbb{E}\Big[\psi_1 \,\Big|\, \{f^\ell(X)\}_{X\in\mathcal{X}_0}, \{S_{\mathrm{GNN}}^{(\ell,h)}\}_{h\in\mathbb{N}}\Big] = \sum_{(X,a)\in\mathcal{L}} \sqrt{d^{\ell,H}}(\alpha^{X,a})^\top \left(S_{\mathrm{GNN}}^{(\ell,1)} f^\ell(X)\right)_a \mathbb{E}[w_1^{\ell,H}] = 0,$$

and hence $\mathbb{E}[\psi_1] = 0$.

For $\mathbb{E}[\psi_1\psi_2]$, expand the product using (40):

$$\psi_1\psi_2 = d^{\ell,H} \sum_{(X,a)\in\mathcal{L}} \sum_{(X',b)\in\mathcal{L}} (\alpha^{X,a})^\top \left(S_{\mathrm{GNN}}^{(\ell,1)} f^\ell(X)\right)_a (\alpha^{X',b})^\top \left(S_{\mathrm{GNN}}^{(\ell,2)} f^\ell(X')\right)_b w_1^{\ell,H} w_2^{\ell,H}.$$

Conditioning on $\{f^\ell(X)\}_{X\in\mathcal{X}_0}$ and $\{S_{\mathrm{GNN}}^{(\ell,h)}\}_{h\in\mathbb{N}}$, the random variables $w_1^{\ell,H}$ and $w_2^{\ell,H}$ are independent, centred Gaussian and independent of the remaining factors in each summand. Thus

$$\mathbb{E}[w_1^{\ell,H} w_2^{\ell,H}] = \mathbb{E}[w_1^{\ell,H}]\mathbb{E}[w_2^{\ell,H}] = 0,$$

which implies

$$\mathbb{E}\Big[\psi_1\psi_2 \,\Big|\, \{f^\ell(X)\}_{X\in\mathcal{X}_0}, \{S_{\mathrm{GNN}}^{(\ell,h)}\}_{h\in\mathbb{N}}\Big] = 0, \qquad \text{and hence} \qquad \mathbb{E}[\psi_1\psi_2] = 0.$$

□

Lemma 17 (Head-wise variance convergence)**.** *There exists $\tau^2 \geq 0$ such that $\lim_{d^{\ell,H}\to\infty} \mathbb{E}[\psi_1^2] = \tau^2$.*

Proof. Observe that $\mathbb{E}[\psi_1^2]$ can be written as

$$\mathbb{E}[\psi_1^2] = d^{\ell,H} \sum_{(X,a),(X',b)\in\mathcal{L}} (\alpha^{X,a})^\top \mathbb{E}\Big[g^{\ell 1}(X)\, w_1^{\ell,H} w_1^{\ell,H\top}\, g^{\ell 1}(X')^\top\Big]\alpha^{X',b}$$
$$\frac{\sigma_H^2}{d_\ell} \sum_{(X,a),(X',b)\in\mathcal{L}} (\alpha^{X,a})^\top \mathbb{E}\Big[g^{\ell 1}(X)\, g^{\ell 1}(X')^\top\Big]\alpha^{X',b},$$

where we used $\mathbb{E}[w_1^{\ell,H} w_1^{\ell,H\top}] = \frac{\sigma_H^2}{d_\ell d^{\ell,H}}$.

$$\frac{\sigma_H^2}{d_\ell}\mathbb{E}\Big[g^{\ell 1}(X)\, g^{\ell 1}(X)^\top\Big] = \frac{\sigma_H^2}{d_\ell}\mathbb{E}\Big[\big(S_{\mathrm{GNN}}^{(\ell,1)} f^{\ell-1}(X) W_{\ell,1}\big)\,\big(S_{\mathrm{GNN}}^{(\ell,1)} f^{\ell-1}(X) W_{\ell,1}\big)^\top\Big]$$
$$= \sigma_H^2 \mathbb{E}\Big[S_{\mathrm{GNN}}^{(\ell,1)} \frac{f^{\ell-1}(X) f^{\ell-1}(X)^\top}{d_{\ell-1}} (S_{\mathrm{GNN}}^{(\ell,1)})^\top\Big]$$

We may thus apply Lemma 29, which requires convergence in distribution of the integrands to the relevant limit and uniform integrability of the integrand family. Convergence in distribution follows by combining the model-specific convergence argument in Lemma 19, Lemma 25, Lemma 26, Lemma 36 with Lemma 30 and Hron et al. [2020, Lemma 33] combined continuous mapping theorem. Uniform integrability is obtained by Hölder's inequality together with Hron et al. [2020, Lemma 32] and the corresponding moment propagation result for $S_{\mathrm{GNN}}^{(\ell,h)}$, namely Lemma 31, Lemma 32,Lemma 33, Lemma 34 . Hence the integrand family is uniformly integrable by Lemma 28, concluding the proof. □

Lemma 18. *For any $h, h' \in \mathbb{N}$, $\mathbb{E}[\psi_h^2 \psi_{h'}^2]$ converges to the mean of the weak limit of $\{\psi_h^2 \psi_{h'}^2\}_{d^{\ell,H}\to\infty}$.*

Proof. It is enough to prove convergence of expectations of the form

$$\mathbb{E}\Big[g_a^{\ell h}(X)\, g_b^{\ell h}(X')^\top\, g_c^{\ell h'}(Y)\, g_d^{\ell h'}(Y')^\top\Big],$$

where a, b, c, d range over feature indices and $(X, X', Y, Y') \in \mathcal{L}$.

Using the definition of $g^{\ell h}$, we may rewrite the above as

$$\mathbb{E}\Big[g_a^{\ell h}(X)\, g_b^{\ell h}(X')^\top\, g_c^{\ell h'}(Y)\, g_d^{\ell h'}(Y')^\top\Big]$$
$$= \sigma_w^4\, \mathbb{E}\Big[\big(S_{\mathrm{GNN}}^{(\ell,h)}\big)_a \frac{f^{\ell-1}(X)\, f^{\ell-1}(X')^\top}{d^{\ell-1}} \big(S_{\mathrm{GNN}}^{(\ell,h)}\big)_b$$
$$\big(S_{\mathrm{GNN}}^{(\ell,h')}\big)_c \frac{f^{\ell-1}(Y)\, f^{\ell-1}(Y')^\top}{d^{\ell-1}} \big(S_{\mathrm{GNN}}^{(\ell,h')}\big)_d\Big].$$

By Hron et al. [2020, Lemma 33],

$$\left(\frac{f^{\ell-1}(X)\, f^{\ell-1}(X')^\top}{d^{\ell-1}}, \frac{f^{\ell-1}(Y)\, f^{\ell-1}(Y')^\top}{d^{\ell-1}}\right) \xrightarrow{P} \big(K^{(\ell-1)}(X,X'),\, K^{(\ell-1)}(Y,Y')\big).$$

Since $S_{\mathrm{GNN}}^{(\ell,h)}$ and $S_{\mathrm{GNN}}^{(\ell,h')}$ converge in distribution (and are independent of the weight matrices producing $f^{\ell-1}$). Then the entire integrand converges in distribution by Lemma 30. Finally, Billingsley [1986, Theorem 3.5] together with Hölder's inequality and Hron et al. [2020, Lemma 32] implies uniform integrability of the integrand family and hence convergence of the expectation, concluding the proof. □

C.2 CONVERGENCE OF $S_{\mathrm{GAT}}^{(\ell,h)}$

In Lemma 14, we showed that, under the GP induction hypothesis on the node features, it is sufficient to assume that $S_{\mathrm{GNN}}^{(\ell,h)}$ converges in distribution in order for the layer-(ℓ) pre-activation outputs $g^\ell(X)$ to converge in distribution to a Gaussian process. In the standard GAT parameterisation, $S_{\mathrm{GNN}}^{(\ell,h)}$ is a measurable function of the attention logits $E^{(\ell h)}$, and hence

convergence of $E^{(\ell h)}$ to a centred Gaussian process implies convergence of $S^{(\ell,h)}_{\mathrm{GNN}}$ by the continuous mapping theorem. Consequently, to establish Gaussian process convergence of the GAT outputs, it suffices to prove that the attention logits converge in distribution to a centred Gaussian process.

We now recall the version of the Blum–Kiefer–Rosenblatt central limit theorem that we will use, following Matthews et al. [2018, lemma 10], and apply it to establish convergence of the attention logits. Fix a layer index ℓ. For each head h and graph edge $(a, i) \in \mathcal{E}$, recall the attention logit

$$E^{(\ell h)}_{ai}(X) := \vec{v}^{\top}_{\ell h}\big[W^{\ell,h} f^{\ell}_{a}(X),\ W^{\ell,h} f^{\ell}_{i}(X)\big],$$

with the scaling and independence assumptions on $W_{\ell,h}$ and $\vec{a}_{\ell h}$ given in Theorem 7.

Because the graph structure in GAT only affects the model through subsequent linear aggregation of the attention outputs, and linear transformations preserve Gaussianity, it suffices to establish convergence of the attention logits themselves. More precisely, for the fixed layer ℓ, we show that the collection of logits

$$E^{(\ell)} := \big\{E^{(\ell h)}_{ai}(X) : X \in \mathcal{X}_0,\ (a, i) \in \mathcal{E},\ h \in \mathbb{N}\big\}$$

has finite-dimensional marginals converging, as $d_\ell \to \infty$, to those of a centred process with covariance given by the edge-level kernel $K^{(\ell),E}$ defined in Equation (21) of Theorem 7.

Lemma 19 (Logit-level GP limit of GAT)**.** *Let the assumptions of Theorem 7 hold. Then, for the fixed layer ℓ, the process*

$$E^{(\ell)} := \big\{E^{(\ell h)}_{ai}(X) : X \in \mathcal{X}_0,\ (a, i) \in \mathcal{E},\ h \in \mathbb{N}\big\}$$

converges in distribution (as $d_\ell \to \infty$) to a centred process whose finite-dimensional marginals are Gaussian with covariance given by the edge kernel $K^{(\ell),E}$ of (21) *in Theorem 7.*

Proof. Using Hron et al. [2020, Lemma 27] and the Cramér–Wold device [Billingsley, 1986], we may restrict attention to one-dimensional projections of finite-dimensional marginals of $E^{(\ell)}$. Let $\mathcal{C} \subset \mathcal{X}_0 \times \mathcal{E} \times \mathbb{N}$ be finite, and choose test matrices $\{\beta^{X,ai,h}\}$ of the same shape as $E^{(\ell h)}_{ai}(X)$. We consider linear statistics of the form

$$\begin{aligned}
T^{E^\ell} &:= \sum_{(X,(a,i),h)\in\mathcal{C}} \big\langle \beta^{X,ai,h}, E^{(\ell h)}_{ai}(X)\big\rangle_F \\
&= \sum_{(X,(a,i),h)\in\mathcal{C}} \left\langle \beta^{X,ai,h}, \frac{1}{\sqrt{d_\ell}} \sum_{j=1}^{d_\ell} \Big[\vec{v}^{(L)}_{\ell h,j} W_{\ell,h,\cdot j} f^{\ell-1}_{a}(X)\cdot \mathbf{1}^{\top} + \mathbf{1}\cdot \vec{v}^{(R)}_{\ell h,j} W_{\ell,h,\cdot j} f^{\ell-1}_{i}(X)\Big] \right\rangle_F \\
&= \frac{1}{\sqrt{d_\ell}} \sum_{j=1}^{d_\ell} \underbrace{\sum_{(X,(a,i),h)\in\mathcal{C}} \big\langle \beta^{X,ai,h}, \vec{v}^{(L)}_{\ell h,j} W_{\ell,h,\cdot j} f^{\ell-1}_{a}(X)\cdot \mathbf{1}^{\top} + \mathbf{1}\cdot \vec{v}^{(R)}_{\ell h,j} W_{\ell,h,\cdot j} f^{\ell-1}_{i}(X)\big\rangle_F}_{=:\,\varphi_j}.
\end{aligned}$$

Here $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product (equivalently, the Euclidean inner product on vectorised matrices). We decompose the vector v into two subvectors $v^{(L)}$ and $v^{(R)}$, which have equal length and identical distribution. Thus T^{E^ℓ} is of the form (37) with $X_{n,i}$ identified with φ_j and d_n with d_ℓ. To invoke Lemma 13 it suffices to verify conditions (a)–(c) for $\{\varphi_j\}_{j=1}^{d_\ell}$, which we do in Lemmas 20–24 below:

- Exchangeability in j is established in Lemma 20.
- Zero mean and vanishing cross-covariance follow from Lemma 21.
- Convergence of the variance is established in Lemma 22.
- Convergence of $\mathbb{E}[\varphi_1^2\varphi_2^2]$ is proved in Lemma 23.
- The $o(\sqrt{d_\ell})$ growth of the third absolute moment is implied by Lemma 24.

Hence T^{E^ℓ} converges in distribution to a centred Gaussian random variable with variance determined by the limit of $\mathbb{E}[\varphi_1^2]$, which is exactly the covariance specified by the edge-level kernel $K^{(\ell h),E}$. Since $\mathcal{C}$ and the test coefficients are arbitrary, the claim follows by Cramér–Wold. □

Lemma 20 (Exchangeability over feature index)**.** *Under the assumptions of Theorem 7, the random variables $\{\varphi_j\}_{j=1}^{d_\ell}$ are exchangeable over the index j.*

Lemma 21 (Zero mean and vanishing cross-covariance)**.** *Under the assumptions of Theorem 7, we have*

$$\mathbb{E}[\varphi_1] = 0, \qquad \mathbb{E}[\varphi_1\varphi_2] = 0.$$

Proof. Recall that, for a finite index set $\mathcal{C} \subset \mathcal{X}_0 \times \mathcal{E} \times \mathbb{N}$,

$$\varphi_j = \sum_{(X,(a,i),h)\in\mathcal{C}} \big\langle \beta^{X,ai,h}, \vec{v}_{\ell h,1}^{(L)} W_{\ell,h,\cdot 1} f_a^{\ell-1}(X) \cdot \mathbf{1}^\top + \mathbf{1} \cdot \vec{v}_{\ell h,1}^{(R)} W_{\ell,h,\cdot 1} f_i^{\ell-1}(X) \big\rangle_F, \qquad j = 1, 2,$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product.

Mean. For any fixed $(X, (a, i), h) \in \mathcal{C}$,

$$\mathbb{E}\Big[\big\langle \beta^{X,ai,h}, \vec{v}_{\ell h,1}^{(L)} W_{\ell,h,\cdot 1} f_a^{\ell}(X) \cdot \mathbf{1}^\top + \mathbf{1} \cdot \vec{v}_{\ell h,1}^{(R)} W_{\ell,h,\cdot 1} f_i^{\ell}(X) \big\rangle_F\Big] = \big\langle \beta^{X,ai,h}, 0 \big\rangle_F = 0,$$

since $W_{\ell,h,\cdot 1}$, $\vec{v}_{\ell h,1}^{(L)}$ and $\vec{v}_{\ell h,1}^{(R)}$ are independent, centred, and have finite moments, and the entries of $f^\ell(X)$ are almost surely finite. Summing over $(X, (a, i), h) \in \mathcal{C}$ gives $\mathbb{E}[\varphi_1] = 0$.

Cross-covariance. Define, for $j = 1, 2$,

$$R_j^h(X; a, i) := \vec{v}_{\ell h,j}^{(L)} W_{\ell,h,\cdot j} f_a^{\ell-1}(X) \cdot \mathbf{1}^\top + \mathbf{1} \cdot \vec{v}_{\ell h,j}^{(R)} W_{\ell,h,\cdot j} f_i^{\ell-1}(X).$$

Then

$$\varphi_j = \sum_{(X,(a,i),h)\in\mathcal{C}} \langle \beta^{X,ai,h}, R_j^h(X; a, i) \rangle_F, \qquad j = 1, 2,$$

and hence

$$\mathbb{E}[\varphi_1\varphi_2] = \sum_{(X,(a,i),h)} \sum_{(X',(a',i'),h')} (\beta^{X,ai,h})^\top \mathbb{E}\Big[R_1^h(X; a, i)\, R_2^{h'}(X'; a', i')^\top\Big] \beta^{X',a'i',h'}.$$

We now expand the inner expectation. By definition,

$$\begin{aligned} R_1^h(X; a, i) &= \vec{v}_{\ell h,1}^{(L)} W_{\ell,h,\cdot 1} f_a^{\ell-1}(X) \cdot \mathbf{1}^\top + \mathbf{1} \cdot \vec{v}_{\ell h,1}^{(R)} W_{\ell,h,\cdot 1} f_i^{\ell-1}(X), \\ R_2^{h'}(X'; a', i') &= \vec{v}_{\ell h',2}^{(L)} W_{\ell,h',\cdot 2} f_{a'}^{\ell-1}(X') \cdot \mathbf{1}^\top + \mathbf{1} \cdot \vec{v}_{\ell h',2}^{(R)} W_{\ell,h',\cdot 2} f_{i'}^{\ell-1}(X'). \end{aligned}$$

Thus

$$\begin{aligned} &\mathbb{E}\Big[R_1^h(X; a, i)\, R_2^{h'}(X'; a', i')^\top\Big] \\ &= \mathbb{E}\Big[\vec{v}_{\ell h,1}^{(L)} W_{\ell,h,\cdot 1} f_a^{\ell-1}(X) f_{a'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 2})^\top (\vec{v}_{\ell h',2}^{(R)})^\top \\ &\quad + \vec{v}_{\ell h,1}^{(L)} W_{\ell,h,\cdot 1} f_a^{\ell-1}(X) f_{i'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 2})^\top (\vec{v}_{\ell h',2}^{(R)})^\top \\ &\quad + \vec{v}_{\ell h,1}^{(R)} W_{\ell,h,\cdot 1} f_i^{\ell-1}(X) f_{a'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 2})^\top (\vec{v}_{\ell h',2}^{(L)})^\top \\ &\quad + \vec{v}_{\ell h,1}^{(R)} W_{\ell,h,\cdot 1} f_i^{\ell-1}(X) f_{i'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 2})^\top (\vec{v}_{\ell h',2}^{(R)})^\top \Big]. \end{aligned}$$

In each of the four terms above, the random matrices $W_{\ell,h,\cdot 1}$ and $W_{\ell,h',\cdot 2}$ appear with total degree one. Since the columns $\{W_{\ell,h,\cdot j}\}_j$ are independent and centred, and are also independent of $f^\ell(X)$ and $f^\ell(X')$, we obtain

$$\mathbb{E}\Big[R_1^h(X; a, i)\, R_2^{h'}(X'; a', i')^\top\Big] = 0$$

entrywise, for all $(X, (a, i), h)$ and $(X', (a', i'), h')$. Therefore every term in the double sum for $\mathbb{E}[\varphi_1\varphi_2]$ vanishes, and hence

$$\mathbb{E}[\varphi_1\varphi_2] = 0.$$

This completes the proof. □

Lemma 22 (Convergence of the variance)**.** *Under the assumptions of Theorem 7, there exists $\sigma_*^2 \geq 0$ such that*

$$\lim_{d_\ell \to \infty} \mathbb{E}[\varphi_1^2] = \sigma_*^2.$$

Proof. From the definition of φ_1 in Lemma 21, we can write

$$\mathbb{E}[\varphi_1^2] = \sum_{(X,(a,i),h)} \sum_{(X',(a',i'),h')} (\beta^{X,ai,h})^\top \mathbb{E}\Big[R_1^h(X;a,i)\, R_1^{h'}(X';a',i')^\top\Big] \beta^{X',a'i',h'},$$

where

$$R_1^h(X;a,i) := \vec{v}_1^{(L)} W_{\ell,h,\cdot 1} f_a^{\ell-1}(X) \cdot \mathbf{1}^\top + \mathbf{1} \cdot \vec{v}_1^{(R)} W_{\ell,h,\cdot 1} f_i^{\ell-1}(X).$$

The inner expectation can be expanded as

$$\begin{aligned}
&\mathbb{E}\Big[R_1^h(X;a,i)\, R_1^{h'}(X';a',i')^\top\Big] \\
&\quad = \mathbb{E}\Big[\big(\vec{v}_1^{(L)} W_{\ell,h,\cdot 1} f_a^\ell(X) \cdot \mathbf{1}^\top + \mathbf{1} \cdot \vec{v}_1^{(R)} W_{\ell,h,\cdot 1} f_i^\ell(X)\big)\big(\vec{v}_1^{(L)} W_{\ell,h',\cdot 1} f_{a'}^\ell(X') \cdot \mathbf{1}^\top + \mathbf{1} \cdot \vec{v}_1^{(R)} W_{\ell,h',\cdot 1} f_{i'}^\ell(X')\big)^\top\Big] \\
&\quad = \mathbb{E}\Big[\vec{v}_1^{(L)} W_{\ell,h,\cdot 1} f_a^{\ell-1}(X) f_{a'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 1})^\top (\vec{v}_1^{(L)})^\top \\
&\qquad\quad + \vec{v}_1^{(L)} W_{\ell,h,\cdot 1} f_a^{\ell-1}(X) f_{i'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 1})^\top (\vec{v}_1^{(R)})^\top \\
&\qquad\quad + \vec{v}_1^{(R)} W_{\ell,h,\cdot 1} f_i^{\ell-1}(X) f_{a'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 1})^\top (\vec{v}_1^{(L)})^\top \\
&\qquad\quad + \vec{v}_1^{(R)} W_{\ell,h,\cdot 1} f_i^{\ell-1}(X) f_{i'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 1})^\top (\vec{v}_1^{(R)})^\top\Big].
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}[\varphi_1^2] = \sum_{(X,(a,i),h)} \sum_{(X',(a',i'),h')} (\beta^{X,ai,h})^\top \Big\{ &\mathbb{E}\Big[v_1^{(L)} W_{\ell,h,\cdot 1} f_a^{\ell-1}(X) f_{a'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 1})^\top (v_1^{(L)})^\top\Big] \\
&+ \mathbb{E}\Big[v_1^{(L)} W_{\ell,h,\cdot 1} f_a^{\ell-1}(X) f_{i'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 1})^\top (v_1^{(R)})^\top\Big] \\
&+ \mathbb{E}\Big[v_1^{(R)} W_{\ell,h,\cdot 1} f_i^{\ell-1}(X) f_{a'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 1})^\top (v_1^{(L)})^\top\Big] \\
&+ \mathbb{E}\Big[v_1^{(R)} W_{\ell,h,\cdot 1} f_i^{\ell-1}(X) f_{i'}^{\ell-1}(X')^\top (W_{\ell,h',\cdot 1})^\top (v_1^{(R)})^\top\Big]\Big\} \beta^{X',a'i',h'}.
\end{aligned}$$

Using independence, zero mean and the variance scaling of $W_{\ell,h,\cdot 1}$ and $W_{\ell,h',\cdot 1}$ from Equation (5) of Theorem 1, each of the four expectations above can be rewritten as a constant multiple of

$$\mathbb{E}\left[\frac{f_u^{\ell-1}(X) f_v^{\ell-1}(X')^\top}{d_{\ell-1}}\right]$$

with $(u,v) \in \{a,i\} \times \{a',i'\}$. By the inductive GP hypothesis for layer ℓ, these expectations converge entrywise to the corresponding entries of $K^{(\ell-1)}(X,X')$[Hron et al., 2020, Lemma 33]. Therefore $\mathbb{E}[\varphi_1^2]$ converges to a finite limit, which we denote by σ_*^2. □

Lemma 23 (Convergence of the mixed fourth moment)**.** *Under the assumptions of Theorem 7,*

$$\lim_{d_\ell \to \infty} \mathbb{E}[\varphi_1^2 \varphi_2^2] = \sigma_*^4.$$

Proof. Define

$$R_j^h(X;a,i) := \vec{v}_j^{(L)} W_{\ell,h,\cdot j} f_a^{\ell-1}(X) \cdot \mathbf{1}^\top + \mathbf{1} \cdot \vec{v}_j^{(R)} W_{\ell,h,\cdot j} f_i^{\ell-1}(X), \qquad j = 1,2.$$

Then

$$\varphi_j = \sum_{(X,(a,i),h) \in \mathcal{C}} \langle \beta^{X,ai,h}, R_j^h(X;a,i) \rangle_F, \quad j = 1,2.$$

Hence

$$\mathbb{E}[\varphi_1^2\varphi_2^2] = \sum_{(X,(a,i),h)} \sum_{(X',(a',i'),h')} (\beta^{X,ai,h})^\top \mathbb{E}\Big[R_1^h(X;a,i)R_1^h(X;a,i)^\top\Big]\beta^{X,ai,h} \cdot (\beta^{X',a'i',h'})^\top \mathbb{E}\Big[R_2^{h'}(X';a',i')R_2^{h'}(X';a',i')^\top\Big]\beta^{X',a'i',h'}.$$

We have, for example,

$$\mathbb{E}\Big[R_1^h(X;a,i)R_1^h(X;a,i)^\top\Big] = \mathbb{E}\Big\{\big(\vec{v}_1^{(L)}W_{\ell,h,\cdot 1}f_a^{\ell-1}(X)\cdot\mathbf{1}^\top + \mathbf{1}\cdot\vec{v}_1^{(R)}W_{\ell,h,\cdot 1}f_i^{\ell-1}(X)\big) \times \big(\vec{v}_1^{(L)}W_{\ell,h,\cdot 1}f_a^{\ell-1}(X)\cdot\mathbf{1}^\top + \mathbf{1}\cdot\vec{v}_1^{(R)}W_{\ell,h,\cdot 1}f_i^{\ell-1}(X)\big)^\top\Big\},$$

and similarly for $R_2^{h'}(X';a',i')$ with parameters $(\vec{a}_2,\vec{b}_2)$. Thus a generic term in the expansion of $\mathbb{E}[\varphi_1^2\varphi_2^2]$ takes the form

$$\mathbb{E}\Big\{\big(v_1^{(L)}W_{\ell,h,\cdot 1}f_a^{\ell-1}(X)\,\mathbf{1}^\top + \mathbf{1}\,v_1^{(R)}W_{\ell,h,\cdot 1}f_i^{\ell-1}(X)\big)\big(v_1^{(L)}W_{\ell,h,\cdot 1}f_a^{\ell-1}(X)\,\mathbf{1}^\top + \mathbf{1}\,v_1^{(R)}W_{\ell,h,\cdot 1}f_i^{\ell-1}(X)\big)^\top \times \big(v_2^{(L)}W_{\ell,h',\cdot 2}f_{a'}^{\ell-1}(X')\,\mathbf{1}^\top + \mathbf{1}\,v_2^{(R)}W_{\ell,h',\cdot 2}f_{i'}^{\ell-1}(X')\big)\big(v_2^{(L)}W_{\ell,h',\cdot 2}f_{a'}^{\ell-1}(X')\,\mathbf{1}^\top + \mathbf{1}\,v_2^{(R)}W_{\ell,h',\cdot 2}f_{i'}^{\ell-1}(X')\big)^\top\Big\}.$$

Expanding this product yields a finite sum of terms of the form

$$\mathbb{E}\Big[\vec{u}_1^\top W_{\ell,h,\cdot 1}f_u^{\ell-1}(X)f_z^{\ell-1}(X)^\top(W_{\ell,h,\cdot 1})^\top\vec{u}_2\cdot\vec{z}_1^\top W_{\ell,h',\cdot 2}f_{u'}^{\ell-1}(X')f_{z'}^{\ell-1}(X')^\top(W_{\ell,h',\cdot 2})^\top\vec{z}_2\Big],$$

where $\vec{u}_1,\vec{u}_2,\vec{z}_1,\vec{z}_2$ are fixed coefficient vectors formed from $\vec{v}_j^{(L)},\vec{v}_j^{(R)}$, and $(u,z)\in\{a,i\}^2$, $(u',z')\in\{a',i'\}^2$. Using independence and the variance scaling of $W_{\ell,h,\cdot 1},W_{\ell,h',\cdot 2}$, each such term is proportional to

$$\mathbb{E}\Big[\frac{f_u^{\ell-1}(X)f_z^{\ell-1}(X)^\top}{d_{\ell-1}}\,\frac{f_{u'}^{\ell-1}(X')f_{z'}^{\ell-1}(X')^\top}{d_{\ell-1}}\Big].$$

Thus $\mathbb{E}[\varphi_1^2\varphi_2^2]$ can be written as a weighted sum of expectations of the form

$$\mathbb{E}\Big[\frac{f_u^{\ell-1}(X)f_v^{\ell-1}(X)^\top}{d_{\ell-1}}\,\frac{f_{u'}^{\ell-1}(X')f_{v'}^{\ell-1}(X')^\top}{d_{\ell-1}}\Big],$$

with (u,v,u',v') ranging over a finite index set. By the inductive GP hypothesis for layer ℓ and Lemma 22, each factor

$$\frac{f_u^{\ell-1}(X)f_z^{\ell-1}(X)^\top}{d_{\ell-1}}$$

converges in probability to the corresponding kernel entry $K_{uz}^{(\ell-1)}(X,X)$, and similarly for (X',u',z'). By the continuous mapping theorem the products converge in probability to sums of products of kernel limits such as

$$K_{uz}^{(\ell-1)}(X,X)\,K_{u'z'}^{(\ell-1)}(X',X').$$

Using the eighth-moment bound on f^ℓ (as in Lemma 24) and Hölder's inequality, the collection of these products is uniformly integrable. Hence the expectations converge to the corresponding products of limits, and we obtain

$$\lim_{d_\ell\to\infty}\mathbb{E}[\varphi_1^2\varphi_2^2] = \sigma_*^4,$$

with σ_*^2 as in Lemma 22. □

Lemma 24 (Third absolute moment)**.** *Under the assumptions of Theorem 7,*

$$\mathbb{E}\big|\varphi_1\big|^3 = o\big(\sqrt{d_\ell}\big).$$

Proof. By Hölder's inequality, it suffices to show that

$$\limsup_{d_\ell \to \infty} \mathbb{E}\big|\varphi_1\big|^4 < \infty.$$

Using the same notation $R_1^h(X)$ as above, we can write

$$\mathbb{E}\big|\varphi_1\big|^4 = \sum_{(X,(a,i),h),(X',(a',i'),h')} (\beta^{X,ai,h})^\top \mathbb{E}\Big[R_1^h(X) R_1^h(X)^\top R_1^{h'}(X') R_1^{h'}(X')^\top\Big] \beta^{X',a'i'}.$$

Each expectation is a finite sum of terms bounded by

$$\max_{c\in V} \max_{Z\in\{X,X'\}} \mathbb{E}\big|f_{c1}^\ell(Z)\big|^8,$$

which is uniformly bounded in d_ℓ by the same eighth-moment assumption used in Hron et al. [2020, Lemma 32]. Thus $\mathbb{E}|\varphi_1|^4$ is uniformly bounded, which implies $\mathbb{E}|\varphi_1|^3 = o(\sqrt{d_\ell})$ and completes the proof. □

C.3 CONVERGENCE OF $S^{(\ell,h)}_{\text{Graphormer}}$

Since the infinite-width limit for networks with positional encodings is treated in Hron et al. [2020, Appendix C], and since Lemma 14 reduces the infinite-head argument to convergence in distribution of $S^{(\ell,h)}_{\text{GNN}}$, it remains to establish this convergence for Graphormer with structural bias. In Graphormer, $S^{(\ell,h)}_{\text{GNN}}$ is obtained from the attention logits $E^{(\ell h)}$ by adding the structural bias and then applying a softmax functin. Therefore $S^{(\ell,h)}_{\text{GNN}}$ is a measurable function of the bias-augmented logits, and convergence in distribution of the logits implies convergence of $S^{(\ell,h)}_{\text{GNN}}$ by the continuous mapping theorem. Hence it suffices to prove that the attention logits still converge when the structural bias is included.

Lemma 25 (Logit-level GP limit of Graphormer)**.** *Let the assumptions of Theorem 10 hold. Then, for the fixed layer ℓ, the process*

$$E^{(\ell)} := \big\{E_{ai}^{(\ell h)}(X) : X \in \mathcal{X}_0,\ (a,i) \in (V\times V),\ h \in \mathbb{N}\big\}$$

converges in distribution (as $d_\ell \to \infty$) to a centred process whose finite-dimensional marginals are Gaussian with covariance given by the edge kernel $K^{(\ell),E}$ of (21) *in Theorem 10.*

Proof. By definition of the Graphormer attention logits (Equation (5)), we have

$$E_{ij}^{(\ell h)}(X) = \frac{(\tilde f_i^\ell(X) W^{Q,\ell h})(\tilde f_j^\ell(X) W^{K,\ell h})^\top}{\sqrt{d_\ell}} + b_{\rho(i,j)}.$$

Define $G_{ij}^{(\ell h)}(X) = \frac{\big(f_i^\ell(X)W^{Q,\ell h}\big)\big(f_j^\ell(X)W^{K,\ell h}\big)^\top}{\sqrt{d_\ell}}$. The covariance $\mathbb{E}\big[E_{ai}^{(\ell h)}(X) E_{bj}^{(\ell h')}(X)\big]$ can be rewritten as:

$$\begin{aligned}
\mathbb{E}\big[E_{ai}^{(\ell h)}(X) E_{bj}^{(\ell h')}(X)\big] &= \mathbb{E}\big[(G_{ai}^{\ell h}(X) + b_{\rho(a,i)})(G_{bj}^{\ell h'}(X) + b_{\rho(b,j)})\big] \\
&= \mathbb{E}\big[G_{ai}^{\ell h}(X) G_{bj}^{\ell h'}(X) + b_{\rho(a,i)} G_{bj}^{\ell h'}(X) + b_{\rho(b,j)} G_{ai}^{\ell h}(X) + b_{\rho(a,i)} b_{\rho(b,j)}\big] \\
&= \mathbb{E}\big[G_{ai}^{\ell h}(X) G_{bj}^{\ell h'}(X)\big] + \mathbb{E}\big[b_{\rho(a,i)} b_{\rho(b,j)}\big]
\end{aligned}$$

Since the convergence of $G^{(\ell h)}(X)$ have already been proved in Hron et al. [2020, Lemma 12], and $b_{\phi(v_i,v_j)}$ are independent for different value of $\phi(i,j)$, thus

$$\mathbb{E}\big[E_{ai}^{(\ell h)}(X)\, E_{bj}^{(\ell h')}(X)\big] = \delta_{h=h'}\Big(\sigma_Q^2\sigma_K^2\, \tilde K_{ab}^{(\ell-1)}\, \tilde K_{ij}^{(\ell-1)} + \sigma_b^2\, \mathbf{1}_{\phi(a,i)=\phi(b,j)}\Big).$$

C.4 CONVERGENCE OF $S^h_{\text{Specformer}}$

In Lemma 14, we showed that, under the GP induction hypothesis on the node features, it is sufficient to establish that $S^{(\ell,h)}_{\text{GNN}}$ converges in distribution in order for the layer-ℓ pre-activation outputs $g^\ell(X)$ to converge in distribution to a Gaussian process. In Specformer, the message-passing operator is defined by Equation(2), so $S^{(\ell,h)}_{\text{GNN}} \equiv S^h_{\text{Specformer}}$. Consequently, to establish Gaussian process convergence of the Specformer outputs, it suffices to prove that $S^h_{\text{Specformer}}$ converges in distribution.

Lemma 26. *Let the assumptions of Theorem 12 hold. Then, for the fixed layer ℓ, the process*

$$S_{\text{Specformer}} := \{S^h_{\text{Specformer}} : h \in \mathbb{N}\}$$

converges in distribution.

Proof. Recall from (2) that the Specformer convolution operator in layer ℓ and head h is defined as

$$S^{(\ell,h)}_{\text{Specformer}} = U \operatorname{diag}\big(\bar{\lambda}^{(h)}\big) U^\top, \qquad \bar{\lambda}^{(h)} = \sigma\big(\tilde{\lambda}^{(h)}\big), \qquad \tilde{\lambda}^{(h)} := \bar{H}^{(h)} W_{\lambda,h}, \qquad \bar{H}^{(h)} := H^{T,(h)}.$$

Since the eigenvalue encoding block is a standard Transformer, Hron et al. [2020, Appendix B.1] establishes that H^{t+1} converges in distribution to a centred Gaussian process

$$H^{t+1} \sim \mathcal{GP}\big(0, K_H^{(t+1)}\big)$$

as $d_t, d^{t,H} \to \infty$, with the kernel recursion given in Equation (32). In particular, $\bar{H}^{(h)} = H^{T,(h)}$ converges in distribution to a centred Gaussian process with covariance $K_H^{(T)}$. With the standard $1/d_T$ variance scaling and independence of $W_{\lambda,h}$, each coordinate of $\tilde{\lambda}^{(h)} = \bar{H}^{(h)} W_{\lambda,h}$ is a sum of d_T independent centred terms, so by the CLT $\tilde{\lambda}^{(h)}$ converges in distribution to a centred Gaussian process with covariance

$$C_{\lambda,ab} := \mathbb{E}\Big[(\tilde{\lambda}^{(h)})_a \, (\tilde{\lambda}^{(h)\prime})_b\Big] = \sigma_\lambda^2 \, K^{(T)}_{H,ab}.$$

Consequently, $\bar{\lambda}^{(h)} = \sigma(\tilde{\lambda}^{(h)})$ also converges in distribution. Since U is fixed and the map $\lambda \mapsto U \operatorname{diag}(\lambda) U^\top$ is continuous, the continuous mapping theorem implies that $S^{(\ell,h)}_{\text{Specformer}}$ converges in distribution as $d_T \to \infty$.

Finally, writing $K^{(\ell-1)}(X, X') := \mathbb{E}[f^{\ell-1}(X) f^{\ell-1}(X')^\top]$ and letting $(\bar{H}, \bar{H}')$ denote the pair of limiting Gaussian-process draws associated with (X, X'), we obtain

$$\begin{aligned}
K^{(\ell)}(X, X') &:= \mathbb{E}\big[g^\ell(X)\, g^\ell(X')^\top\big] = \sigma_O^2 \sigma_w^2 \, \mathbb{E}\Big[S^1_{\text{Specformer}}(\bar{H})\, K^{(\ell-1)}(X, X')\, S^1_{\text{Specformer}}(\bar{H}')^\top\Big] \\
&= \sigma_O^2 \sigma_w^2 \, \mathbb{E}\Big[U \operatorname{diag}(\bar{\lambda}^1(\bar{H}))\, (U^\top K^{(\ell-1)}(X, X') U) \operatorname{diag}(\bar{\lambda}^1(\bar{H}'))\, U^\top\Big] \\
&= \sigma_O^2 \sigma_w^2 \, U\Big(K_\lambda \odot \big(U^\top K^{(\ell-1)}(X, X') U\big)\Big) U^\top,
\end{aligned}$$

where $K_{\lambda,ab} := \mathbb{E}[\bar{\lambda}_a(\bar{H}) \bar{\lambda}_b(\bar{H}')]$. Therefore the Specformer outputs $g^\ell(X)$ converge in distribution to a centred Gaussian process with covariance kernel $K^{(\ell)}$. □

C.5 AUXILIARY RESULTS

Lemma 27 ([Billingsley, 1986, Theorem 29.4])**.** *Let $X, (X_n)_{n\geq 1}$ be random variables taking values in $(\mathbb{R}^N, \mathcal{B}^N)$, where $\mathcal{B}^N$ denotes the Borel σ-algebra. Then $X_n \xrightarrow{d} X$ if and only if for every finite index set $J \subset \mathbb{N}$ and the corresponding projection $\Gamma^J : \mathbb{R}^N \to \mathbb{R}^J$, we have*

$$\Gamma^J(X_n) \xrightarrow{d} \Gamma^J(X) \quad \text{as } n \to \infty.$$

Lemma 28 ([Billingsley, 1986, p. 31])**.** *A sequence of real-valued random variables $(X_n)_{n\geq 1}$ is uniformly integrable if*

$$\sup_{n\geq 1} \mathbb{E}\big[|X_n|^{1+\varepsilon}\big] < \infty$$

for some $\varepsilon > 0$.

Theorem 29 ([Billingsley, 1986, Theorem 3.5])**.** *If $(X_n)_{n\geq 1}$ is uniformly integrable and $X_n \xrightarrow{d} X$, then X is integrable and*

$$\mathbb{E}[X_n] \to \mathbb{E}[X].$$

Lemma 30 (Slutsky's lemma)**.** *Let $X, (X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ be real-valued random variables defined on the same probability space. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$ for some constant $c \in \mathbb{R}$, then*

$$X_n Y_n \xrightarrow{d} cX, \qquad X_n + Y_n \xrightarrow{d} X + c.$$

Lemma 31 (Moment bound for GAT operators)**.** *Under the assumptions of Theorem 7, suppose that $\phi : \mathbb{R} \to \mathbb{R}$ is entrywise polynomially bounded, i.e., Then for any $t \geq 1$,*

$$\sup_{a \in V} \sup_{i \in \mathcal{N}_a} \sup_{d_\ell \in \mathbb{N}} \mathbb{E}\big|S^{(\ell,h)}_{\mathrm{GAT},ai}(X)\big|^t < \infty. \tag{41}$$

Proof. For each layer width $d_\ell \in \mathbb{N}$, recall that the unnormalized GAT attention score is given by

$$E^{(\ell h, d_\ell)}_{ai}(X) = (\vec{v}^{\ell h})^\top \Big[W^{\ell,h} f^{\ell-1}_a(X) \parallel W^{\ell,h} f^{\ell-1}_i(X)\Big],$$

where $W^{\ell,h}$ and $\vec{v}^{\ell h}$ are independently fan-in scaled Gaussian and independent of $f^{\ell-1}(x)$.

Conditioned on

$$u^{(d_\ell)}_{ai}(X) := \Big[W^{\ell,h} f^{\ell-1}_a(X) \parallel W^{\ell,h} f^{\ell-1}_i(X)\Big],$$

the random variable $E^{(\ell h, d_\ell)}_{ai}(X)$ is centered Gaussian with variance proportional to $\|u^{(d_\ell)}_{ai}(X)\|^2$. Hence for any $t \geq 1$,

$$\mathbb{E}\big(|E^{(\ell h, d_\ell)}_{ai}(X)|^t \mid u^{(d_\ell)}_{ai}(X)\big) \lesssim \|u^{(d_\ell)}_{ai}(X)\|^t.$$

Taking expectations and using the assumptions of Theorem 7, which ensure that all moments of fan-in scaled linear transformations of $f^{\ell-1}(x)$ are uniformly bounded in d_ℓ, we obtain

$$\sup_{a \in V} \sup_{i \in \mathcal{N}_a} \sup_{d_\ell \in \mathbb{N}} \mathbb{E}\big|E^{(\ell h, d_\ell)}_{ai}(X)\big|^t < \infty.$$

Finally, since ϕ is entrywise polynomially bounded, this uniform moment bound implies that

$$\sup_{a \in V} \sup_{i \in \mathcal{N}_a} \sup_{d_\ell \in \mathbb{N}} \mathbb{E}\big|S^{(\ell,h)}_{\mathrm{GAT},ai}(X)\big|^t < \infty,$$

which completes the proof. □

Lemma 32 (Moment propagation for Graphormer operators)**.** *Under the assumptions of Theorem 10, suppose that $\phi : \mathbb{R} \to \mathbb{R}$ is entrywise polynomially bounded. Then for any $t \geq 1$,*

$$\sup_{i,j \in V} \sup_{d_\ell \in \mathbb{N}} \mathbb{E}\big|S^{(\ell,h)}_{\mathrm{Graphormer},ij}(X)\big|^t < \infty. \tag{42}$$

Proof. For each layer width $d_\ell \in \mathbb{N}$, the unnormalized Graphormer attention score is given by

$$E^{(\ell h, d_\ell)}_{ij}(X) = \frac{\big(\tilde{f}^\ell_i(X) W^{Q,\ell h}\big)\big(\tilde{f}^\ell_j(X) W^{K,\ell h}\big)^\top}{\sqrt{d_\ell}} + b_{\rho(i,j)},$$

where $W^{Q,\ell h}$ and $W^{K,\ell h}$ are independently fan-in scaled Gaussian and independent of $\tilde{f}^\ell(X)$, and the bias term satisfies $\sup_{i,j} |b_{\rho(i,j)}| \leq B < \infty$.

Write

$$E^{(\ell h, d_\ell)}_{ij}(X) = T^{(d_\ell)}_{ij}(X) + b_{\rho(i,j)}, \qquad T^{(d_\ell)}_{ij}(X) := d_\ell^{-1/2} \Big\langle \tilde{f}^\ell_i(X) W^{Q,\ell h}, \tilde{f}^\ell_j(X) W^{K,\ell h} \Big\rangle.$$

Using the inequality $|a+b|^t \leq 2^{t-1}(|a|^t + |b|^t)$ and the boundedness of $b_{\rho(i,j)}$, it suffices to control the moments of $T^{(d_\ell)}_{ij}(X)$ uniformly in d_ℓ.

By Cauchy–Schwarz and Hölder's inequality,

$$\mathbb{E}\big|T^{(d_\ell)}_{ij}(X)\big|^t \leq d_\ell^{-t/2} \Big(\mathbb{E}\|\tilde{f}^\ell_i(X) W^{Q,\ell h}\|^{2t}\Big)^{1/2} \Big(\mathbb{E}\|\tilde{f}^\ell_j(X) W^{K,\ell h}\|^{2t}\Big)^{1/2}.$$

Under the assumptions of Theorem 7, fan-in scaling and the moment bounds on the upstream features imply that all moments of the linear transforms $\tilde{f}^\ell_i(X) W^{Q,\ell h}$ and $\tilde{f}^\ell_j(X) W^{K,\ell h}$ are uniformly bounded in d_ℓ. Consequently,

$$\sup_{i,j \in V} \sup_{d_\ell \in \mathbb{N}} \mathbb{E}\big|E^{(\ell h, d_\ell)}_{ij}(X)\big|^t < \infty.$$

Finally, since ϕ is entrywise polynomially bounded, this uniform moment bound implies that

$$\sup_{i,j\in V}\sup_{d_\ell\in\mathbb{N}}\mathbb{E}\big|S^{(\ell,h)}_{\mathrm{Graphormer},ij}(X)\big|^t<\infty,$$

which completes the proof. □

Lemma 33 (Moment propagation for Specformer spectral filters)**.** *Under the assumptions of Theorem 12. Then for any $t\geq 1$ and any random vector $u\in\mathbb{R}^{|V|}$,*

$$\sup_{d_t,d^{t,H}}\mathbb{E}\big\|S_{\mathrm{Specformer}}u\big\|_2^t<\infty,\qquad \textit{whenever}\qquad \sup_{d_t,d^{t,H}}\mathbb{E}\|u\|_2^t<\infty. \tag{43}$$

In particular, for any fixed coordinate $a\in V$,

$$\sup_{d_t,d^{t,H}}\mathbb{E}\big|(S^{(h)}_{\mathrm{Specformer}}u)_a\big|^t<\infty.$$

Proof. By Lemma 32 of Hron et al. [2020], the upstream hidden representation $\bar{H}$ produced by the Infinite Attention mechanism admits uniform moment bounds of all orders; in particular, for any $t\geq 1$,

$$\sup_{d_t,d^{t,H}}\mathbb{E}\|\bar{H}\|_\infty^t<\infty.$$

Applying a fan-in scaled linear transformation preserves uniform moment bounds, and since σ is entrywise polynomially bounded, it follows that the spectral coefficients admit uniform moment bounds, namely

$$\sup_{d_t,d^{t,H}}\mathbb{E}\|\bar{\lambda}^{(h)}\|_\infty^t<\infty.$$

Now write

$$S^{(h)}_{\mathrm{Specformer}}=U\ \mathrm{diag}\big(\bar{\lambda}^{(h)}\big)\,U^\top.$$

Because U is orthogonal, we have

$$\|S^{(h)}_{\mathrm{Specformer}}\|_{2\to 2}=\|\,\mathrm{diag}(\bar{\lambda}^{(h)})\|_{2\to 2}=\|\bar{\lambda}^{(h)}\|_\infty.$$

Therefore, for any random vector u,

$$\|S^{(h)}_{\mathrm{Specformer}}u\|_2\leq\|\bar{\lambda}^{(h)}\|_\infty\,\|u\|_2.$$

Raising both sides to the power t and taking expectations yields

$$\mathbb{E}\big\|S^{(h)}_{\mathrm{Specformer}}u\big\|_2^t\leq\Big(\mathbb{E}\|\bar{\lambda}^{(h)}\|_\infty^{2t}\Big)^{1/2}\Big(\mathbb{E}\|u\|_2^{2t}\Big)^{1/2},$$

by Cauchy–Schwarz. The claimed bound follows from the established uniform moment bounds. □

Lemma 34 (Moment propagation for GTN operators)**.** *Under the assumptions of Theorem 35, fix a head h and a meta-path length K. Then for any $t\geq 1$ and any random vector u,*

$$\sup_n\sup_{a\in V}\mathbb{E}\big|(S^{(K,h)}(X)u)_a\big|^t\ \leq\ \sup_n\sup_{i\in V}\mathbb{E}|u_i|^t.$$

Proof. By construction, each relation adjacency matrix is row-normalized and entrywise nonnegative. Convex combinations with simplex weights preserve these properties, and products of row-stochastic, entrywise nonnegative matrices remain row-stochastic and entrywise nonnegative. Hence each row of $S^{(K,h)}(x)$ is a probability vector.

For any $a\in V$,

$$(S^{(K,h)}(x)u)_a=\sum_{i\in V}S^{(K,h)}_{ai}(x)\,u_i.$$

Since the function $z \mapsto |z|^t$ is convex for $t \geq 1$, Jensen's inequality yields

$$\left|(S^{(K,h)}(X)u)_a\right|^t \leq \sum_{i \in V} S_{ai}^{(K,h)}(X)\,|u_i|^t.$$

Taking expectations and using $\sum_i S_{ai}^{(K,h)}(X) = 1$ gives

$$\mathbb{E}\left|(S^{(K,h)}(X)u)_a\right|^t \leq \sup_{i \in V} \mathbb{E}|u_i|^t.$$

Taking the supremum over $a \in V$ and n completes the proof. □

D GRAPH TRANSFORMERS ON HETEROGENEOUS GRAPHS

In this Appendix, the GP setting is extended to heterogeneous graphs with multiple edge types.

Heterogeneous Graphs Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T})$ be an undirected graph with N nodes, where $\mathcal{V}$ denotes the set of nodes, $\mathcal{E}$ the set of edges, and $\mathcal{T}$ the set of edge types (relations). Heterogeneous graphs contain multiple edge types, i.e., $|\mathcal{T}| > 1$. For each edge type $t \in \mathcal{T}$, let $A_t \in \mathbb{R}^{N \times N}$ denote the corresponding adjacency matrix.The node feature matrix $X \in \mathbb{R}^{N \times d_{\text{in}}}$ remains unchanged.

Graph Transformers Network (GTN) GTN operates on heterogeneous graphs by recursively constructing meta-path adjacency matrices through soft selections over edge types. For each GTN head $h \in \{1, \ldots, d^{\ell,H}\}$, define $\tilde{A}_i := D_i^{-1} A_i$ for $i \in \mathcal{T}$, where D_i denotes the diagonal degree matrix associated with A_i. The recursion is initialized as $A^{(1,h)} = \sum_{i=1}^{\mathcal{T}} \beta_i^{(1,h)} \tilde{A}_i$. For a fixed meta-path length K,

$$S_{\text{GTN}}^{(K,h)} = A^{(K,h)} = A^{(K-1,h)} \left(\sum_{i=1}^{\mathcal{T}} \beta_i^{(K,h)} A_i \right), \tag{44}$$

where $\beta^{(k,h)}$ for $k = 1, \ldots, K$ are learnable parameters that weight the relation-specific adjacency matrices for the h-th head. The design choices underlying the reformulated GTN formulation are detailed in the Appendix D.

Design Choices for GTN

Reparameterizing GTN Coefficients for Tractable Kernel Construction In the original Graph Transformer Network (GTN) formulation, the coefficients $\{\beta^{(k)}\}_{k=0}^{K}$ are obtained by applying a softmax function to a trainable parameter vector $\{W_\phi^{(k)}\}_{k=0}^{K}$, i.e.,

$$\beta^{(k)} = \phi\Big(W_\phi^{(k)}\Big), \qquad k = 0, \ldots, K,$$

where $\phi(\cdot)$ denotes the softmax operator. Under random initialization of W_ϕ, the softmax transformation induces complex dependencies among the coefficients $\beta^{(k)}$, making their joint distribution difficult to characterize. In particular, the resulting $\beta^{(k)}$ are no longer independent, and moments such as $\mathbb{E}[\beta_i \beta_j]$ do not admit a tractable closed-form expression.

To facilitate a principled Gaussian kernel construction, for each head of graph convolution layer heads $d^{\ell,H}$, therefore depart from the original parameterization and directly initialize the coefficients $\{\beta^{(k)}\}$. This choice enables explicit control over their statistical properties and allows the corresponding kernel expectations to be computed in a tractable manner.

Decoupled Degree Normalization for Well-Defined Graph Kernels Meanwhile,in original paper the GTN adjacency matrix at layer K is given by the recursive construction

$$A^{(K)} = \left(D^{(K-1)}\right)^{-1} A^{(K-1)} \left(\sum_{i=1}^{\mathcal{T}} \beta_i^{(K)} A_i \right), \tag{45}$$

where the associated degree matrix is defined as

$$D_{uu}^{(K)} = \sum_{v=1}^{n} A_{uv}^{(K)}, \qquad D^{(K)} = \sum_{u=1}^{n} D_{uu}^{(K)}\, e_u e_u^\top. \tag{46}$$

By definition, both $A^{(K)}$ and $D^{(K)}$ are random matrices depending on the same collection of random variables $\{\beta_i^{(k)}\}_{k\le K,\, i\le \mathcal{T}}$.

Substituting (45) into (46), the diagonal entry $D_{uu}^{(K)}$ can be written explicitly as

$$D_{uu}^{(K)} = \sum_{v=1}^{n} \left[\left(D^{(K-1)}\right)^{-1} A^{(K-1)} \left(\sum_{i=1}^{\mathcal{T}} \beta_i^{(K)} A_i \right) \right]_{uv}. \tag{47}$$

Consequently, the degree-normalized adjacency satisfies

$$\left(D^{(K)}\right)^{-1} A^{(K)} = \sum_{u=1}^{n} \frac{1}{D_{uu}^{(K)}}\, e_u e_u^\top A^{(K)}. \tag{48}$$

Assume that the coefficients are independently initialized as $\beta_i^{(k)} \sim \mathcal{N}(0, \sigma_k^2)$, for all $i = 1, \dots, \mathcal{T},\ k = 1, \dots, K$. Under this initialization, both $A_{uv}^{(K)}$ and $D_{uu}^{(K)}$ are nonlinear functions of the same Gaussian random variables. In particular, $D_{uu}^{(K)}$ is a centered random variable with a continuous distribution supported on $\mathbb{R}$, and hence satisfies

$$\mathbb{P}\Big(|D_{uu}^{(K)}| < \varepsilon\Big) > 0 \quad \text{for all } \varepsilon > 0. \tag{49}$$

As a result, the random matrix $\left(D^{(K)}\right)^{-1} A^{(K)}$ involves ratios of strongly coupled random variables, where the numerator and denominator depend on the same Gaussian coefficients. This nonlinear coupling prevents any decoupling under expectation, and the matrix-valued expectation $\mathbb{E}\Big[\left(D^{(K)}\right)^{-1} A^{(K)}\Big]$ does not admit a finite or closed-form expression. In particular, the presence of $D_{uu}^{(K)}$ in the denominator precludes the existence of well-defined second-order moments required for a Gaussian process or kernel interpretation.

To avoid this issue, we modify the GTN construction by normalizing each base adjacency matrix independently.

Specifically, define

$$\tilde{A}_i := D_i^{-1} A_i, \qquad i \in \mathcal{T},$$

where D_i denotes the diagonal degree matrix associated with A_i, i.e., $(D_i)_{uu} = \sum_v (A_i)_{uv}$. For each head of graph convolution layer heads $d^{\ell,H}$, recursion is initialized as $A^{(1)} = \sum_{i=1}^{\mathcal{T}} \beta_i^{(1)} \tilde{A}_i$, and, for a fixed meta-path length $K \ge 2$, proceeds according to

$$S_{\mathrm{GTN}}^{(K)} = A^{(K)} = A^{(K-1)} \left(\sum_{i=1}^{\mathcal{T}} \beta_i^{(K)} \tilde{A}_i \right).$$

By performing degree normalization at the level of individual base adjacency matrices, rather than on the recursively constructed adjacency $A^{(K)}$, the resulting graph structure avoids nonlinear coupling between normalization terms and random combination coefficients.

GTN-GP

Theorem 35 (Infinite-width and infinite-head limit of a GTN layer)**.** *If for the ℓ-st GTN layer with meta-path length K, the parameters satisfy $\beta_i^{(k)} \sim \mathcal{N}(0, \sigma_k^2), W_{ij}^{\ell} \sim \mathcal{N}\Big(0, \frac{\sigma_w^2}{d_{\ell-1}}\Big), W_{ij}^{\ell,H} \sim \mathcal{N}\Big(0, \frac{\sigma_H^2}{d^{\ell,H} d_{\ell-1}}\Big)$, independently for all i, j and $k \in [K]$, and define $\tilde{A}_i := D_i^{-1} A_i$ for $i \in \mathcal{T}$, then, as $d_\ell \to \infty$, $g^{\ell+1}(X)$ converges in distribution to $g^{\ell}(X) \sim \mathcal{GP}(0, \Sigma^{(\ell)})$ with*

$$\begin{aligned} \Sigma_{ab}^{(\ell)} &= \mathbb{E}\Big[g_a^{\ell,1}(X)\, g_b^{\ell,1}(X)\Big] \\ &= \left(\sigma_H^2 \sigma_w^2 \prod_{k=1}^{K} \sigma_k^2 \sum_{i_1=1}^{\mathcal{T}} \cdots \sum_{i_K=1}^{\mathcal{T}} \left(\tilde{A}_{i_1} \cdots \tilde{A}_{i_K}\right) K^{(\ell-1)} \left(\tilde{A}_{i_K}^\top \cdots \tilde{A}_{i_1}^\top\right) \right)_{ab}, \end{aligned} \tag{50}$$

where $a, b \in V$.

Equation (50) shows that, in the infinite-width and infinite-head limit, the dependence between node pre-activations at layer ℓ is obtained by aggregating over all possible length-K meta-path compositions. The GTN layer enumerates all sequences of relation types $(i_1, \dots, i_K) \in \mathcal{T}^K$ and combines the corresponding normalized adjacency products $\tilde{A}_{i_1} \cdots \tilde{A}_{i_K}$ to construct the resulting graph structure, so the update can be viewed as forming a composite graph that accounts for every possible meta-path of length K and accumulating their contributions.

Proof. In Lemma 14, we showed that, under the GP induction hypothesis on the node features, it is sufficient to establish that $S_{\mathrm{GNN}}^{(\ell,h)}$ converges in distribution in order for the layer-ℓ pre-activation outputs $g^\ell(X)$ to converge in distribution to a Gaussian process. In GTN, the message-passing operator in head h is $S_{\mathrm{GNN}}^{(\ell,h)} \equiv S_{\mathrm{GTN}}^{(K,h)}$. Consequently, to establish Gaussian process convergence of the GTN outputs, it suffices to prove that $S_{\mathrm{GTN}}^{(K,h)}$ converges in distribution.

Convergence of $S_{\mathrm{GTN}}^{(K,h)}$

Lemma 36. *Let the assumptions of Theorem 35 hold. Then, for the fixed layer ℓ, the process*

$$S_{\mathrm{GTN}}^{K} := \big\{ S_{\mathrm{GTN}}^{(K,h)} : h \in \mathbb{N} \big\}$$

converges in distribution.

Proof. Recall from Equation (44) that the GTN propagation operator with meta-path length K in layer ℓ and head h is generated by the random coefficients $\{\beta^{(k,h)}\}_{k=1}^{K}$ and can be written as

$$S_{\mathrm{GTN}}^{(K,h)} = \Big(\sum_{i_K=1}^{\mathcal{T}} \beta_{i_K}^{(K,h)} \tilde{A}_{i_K} \Big) \cdots \Big(\sum_{i_1=1}^{\mathcal{T}} \beta_{i_1}^{(1,h)} \tilde{A}_{i_1} \Big), \qquad \tilde{A}_i := D_i^{-1} A_i.$$

Assume that for each $k \in \{1, \dots, K\}$ the vectors $\beta^{(k,h)} = (\beta_1^{(k,h)}, \dots, \beta_{\mathcal{T}}^{(k,h)})$ are i.i.d. across h, independent across k, and satisfy $\mathbb{E}[\beta_i^{(k,h)}] = 0$ and $\mathbb{E}[\beta_i^{(k,h)} \beta_j^{(k,h)}] = \sigma_k^2 \delta_{ij}$. Since the matrices $\{\tilde{A}_i\}_{i=1}^{\mathcal{T}}$ are fixed, the map $\{\beta^{(k,h)}\}_{k=1}^{K} \mapsto S_{\mathrm{GTN}}^{(K,h)}$ is a polynomial function, hence measurable and continuous. Therefore $S_{\mathrm{GTN}}^{(K,h)}$ has a well-defined distribution which does not depend on the width d_ℓ, and in particular it is tight and thus converges in distribution along any sequence $d_\ell \to \infty$.

Finally, writing $K^{(\ell-1)}(X, X') := \mathbb{E}[f^{\ell-1}(X) f^{\ell-1}(X')^\top]$ and letting $S_{\mathrm{GTN}}^{(K,h)}$ be an independent draw of the GTN operator, we obtain the kernel recursion

$$\begin{aligned} K^{(\ell)}(X, X') &:= \mathbb{E}\big[g^\ell(X)\, g^\ell(X')^\top \big] = \sigma_H^2 \sigma_w^2\, \mathbb{E}\Big[S_{\mathrm{GTN}}^{(K,1)}\, K^{(\ell-1)}(X, X')\, S_{\mathrm{GTN}}^{(K,1)\top} \Big] \\ &= \sigma_H^2 \sigma_w^2 \left(\prod_{k=1}^{K} \sigma_k^2 \sum_{i_1=1}^{\mathcal{T}} \cdots \sum_{i_K=1}^{\mathcal{T}} \tilde{A}_{i_1} \cdots \tilde{A}_{i_K}\, K^{(\ell-1)}(X, X')\, \tilde{A}_{i_K}^\top \cdots \tilde{A}_{i_1}^\top \right), \end{aligned}$$

where we used $\mathbb{E}[\beta_i^{(k,h)} \beta_j^{(k,h)}] = \sigma_k^2 \delta_{ij}$ and independence across k. Therefore the GTN outputs $g^\ell(X)$ converge in distribution to a centred Gaussian process with covariance kernel $K^{(\ell)}$. □

□

D.1 GTN-GP EXPERIMENTS ON HETEROGENEOUS GRAPHS

To evaluate the practical utility of the derived **GTN-GP**, we compare its performance against several state-of-the-art baselines, including GCN, GAT, HAN, and the standard finite-width GTN. We conduct experiments on three standard heterogeneous graph benchmarks [Wang et al., 2019]: **DBLP**, **ACM**, and **IMDB**.

The results, summarized in Table 5, demonstrate that the infinite-width limit approach is highly competitive.

Table 5: Node classification accuracy on heterogeneous graphs.

Dataset	GCN	GAT	HAN	GTN	GTN-GP
DBLP	87.30	93.71	92.83	94.18	**94.33**
ACM	91.60	92.33	90.96	**92.68**	92.19
IMDB	56.89	58.14	56.77	**60.92**	60.50

E SUPPLEMENTARY MATERIAL OF SECTION 4

E.1 GCN-GP KERNEL UNDER SBM

Corollary 37 (GCN kernel under SBM)**.** *Consider the GCN kernel in Niu et al. [2023], given by $K^{(\ell)} = A^\top K^{(\ell-1)} A$ then, for all $\ell \geq 1$, $K^{(\ell)}$ under SBM admits the following closed-form expression:*

$$x_\ell = \frac{1}{2}\left(\frac{n}{2}\right)^{2\ell}\left[(x_0+y_0)((p+q))^{2\ell} + (x_0-y_0)(p-q)^{2\ell}\right], \tag{51}$$

$$y_\ell = \frac{1}{2}\left(\frac{n}{2}\right)^{2\ell}\left[(x_0+y_0)(p+q)^{2\ell} - (x_0-y_0)(p-q)^{2\ell}\right]. \tag{52}$$

Define $J = \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{\frac{n}{2}\times\frac{n}{2}}$. The matrices A and $K^{(0)}$ share the same eigenbasis $\{v_1, v_2\}$ where $v_1 = \frac{1}{\sqrt{n}}\mathbf{1}_n$ and $v_2 = \frac{1}{\sqrt{n}}[\mathbf{1}_{n/2}^\top, -\mathbf{1}_{n/2}^\top]^\top$. The spectral decompositions are:

$$A = \lambda_{A,1} v_1 v_1^\top + \lambda_{A,2} v_2 v_2^\top, \quad \lambda_{A,1} = (p+q)\frac{n}{2}, \lambda_{A,2} = (p-q)\frac{n}{2}$$
$$K^{(0)} = \lambda_{K,1} v_1 v_1^\top + \lambda_{K,2} v_2 v_2^\top, \quad \lambda_{K,1} = (x_0+y_0)\frac{n}{2}, \lambda_{K,2} = (x_0-y_0)\frac{n}{2}$$

Proof of Corollary 37 Under the GCN update rule $K^{(\ell)} = AK^{(\ell-1)}A$, we have $K^{(\ell)} = A^\ell K^{(0)} A^\ell$. Utilizing the orthogonality $v_i^\top v_j = \delta_{ij}$:

$$\begin{aligned} K^{(\ell)} &= \lambda_{A,1}^{2\ell}\lambda_{K,1} v_1 v_1^\top + \lambda_{A,2}^{2\ell}\lambda_{K,2} v_2 v_2^\top \\ &= \frac{\lambda_{A,1}^{2\ell}\lambda_{K,1}}{n}\begin{pmatrix} J & J \\ J & J \end{pmatrix} + \frac{\lambda_{A,2}^{2\ell}\lambda_{K,2}}{n}\begin{pmatrix} J & -J \\ -J & J \end{pmatrix} \end{aligned}$$

Substituting $\lambda_{A,i}$ and $\lambda_{K,i}$ and simplifying the leading coefficient $\frac{1}{n}\cdot\frac{n}{2} = \frac{1}{2}$ yields the scalar entries:

$$x_\ell = \frac{1}{2}\left[(x_0+y_0)\left((p+q)\frac{n}{2}\right)^{2\ell} + (x_0-y_0)\left((p-q)\frac{n}{2}\right)^{2\ell}\right]$$
$$y_\ell = \frac{1}{2}\left[(x_0+y_0)\left((p+q)\frac{n}{2}\right)^{2\ell} - (x_0-y_0)\left((p-q)\frac{n}{2}\right)^{2\ell}\right]$$

Proof of Corollary 1 From the expressions for x_ℓ and y_ℓ in Table 3, we observe that the diagonal entries of the kernel matrix are all equal to x_ℓ. Thus, the normalization term is $\frac{1}{n}\text{tr}(K^{(\ell)}) = \frac{1}{n}(nx_\ell) = x_\ell$. The normalized kernel is then:

$$\frac{K^{(\ell)}}{\frac{1}{n}\text{tr}(K^{(\ell)})} = \frac{1}{x_\ell}\begin{pmatrix} x_\ell \mathbf{1}\mathbf{1}^\top & y_\ell \mathbf{1}\mathbf{1}^\top \\ y_\ell \mathbf{1}\mathbf{1}^\top & x_\ell \mathbf{1}\mathbf{1}^\top \end{pmatrix} = \begin{pmatrix} \mathbf{1}\mathbf{1}^\top & \frac{y_\ell}{x_\ell}\mathbf{1}\mathbf{1}^\top \\ \frac{y_\ell}{x_\ell}\mathbf{1}\mathbf{1}^\top & \mathbf{1}\mathbf{1}^\top \end{pmatrix}.$$

Using the closed-form expressions for GCN-GP, the ratio of inter- to intra-community similarity is:

$$\frac{y_\ell}{x_\ell} = \frac{(x_0+y_0)(p+q)^{2\ell} - (x_0-y_0)(p-q)^{2\ell}}{(x_0+y_0)(p+q)^{2\ell} + (x_0-y_0)(p-q)^{2\ell}} = \frac{1 - \left(\frac{x_0-y_0}{x_0+y_0}\right)\left(\frac{p-q}{p+q}\right)^{2\ell}}{1 + \left(\frac{x_0-y_0}{x_0+y_0}\right)\left(\frac{p-q}{p+q}\right)^{2\ell}}.$$

Since $|p-q| < |p+q|$, the term $\left(\frac{p-q}{p+q}\right)^{2\ell} \to 0$ as $\ell \to \infty$. Therefore, $\lim_{\ell\to\infty}\frac{y_\ell}{x_\ell} = 1$, and the normalized kernel converges to the rank-one all-ones matrix.

E.2 GAT-GP KERNEL UNDER SBM

Corollary 38 (GAT kernel under SBM)**.** *Consider the GAT kernel in Corollary 9, then, for all $\ell \geq 1$, $K^{(\ell)}$ under SBM admits the following closed-form expression:*

$$x_\ell = \frac{1}{2}\left[\frac{G^{2^\ell}}{\left(\frac{n}{2}\right)^2 (p+q)^2}\right]\left[1 + \left(\frac{x_0 - y_0}{x_0 + y_0}\right) F^\ell\right],$$
$$y_\ell = \frac{1}{2}\left[\frac{G^{2^\ell}}{\left(\frac{n}{2}\right)^2 (p+q)^2}\right]\left[1 - \left(\frac{x_0 - y_0}{x_0 + y_0}\right) F^\ell\right],$$

where the global growth factor G and the structural preservation factor F are defined as:

$$G = (x_0 + y_0)(p+q)^2 \left(\frac{n}{2}\right)^2 \quad \textit{and} \quad F = 2\frac{p^2 - pq + q^2}{(p+q)^2}.$$

Proof of Corollary 38 The evolution of the Gaussian Process (GP) kernel for a Graph Attention Network (GAT) at layer $l+1$ is given by:

$$K^{(l+1)} = K^{(l)} \odot (AK^{(l)}A) + A(K^{(l)} \odot K^{(l)})A, \tag{53}$$

where A is the adjacency matrix and $K^{(l)}$ the kernel at layer l.

We consider a Stochastic Block Model with two communities of size $n/2$. The expected adjacency matrix A and the kernel $K^{(l)}$ exhibit the following block structures:

$$A = \begin{bmatrix} p\mathbf{1}\mathbf{1}^\top & q\mathbf{1}\mathbf{1}^\top \\ q\mathbf{1}\mathbf{1}^\top & p\mathbf{1}\mathbf{1}^\top \end{bmatrix}, \quad K^{(l)} = \begin{bmatrix} x_l\mathbf{1}\mathbf{1}^\top & y_l\mathbf{1}\mathbf{1}^\top \\ y_l\mathbf{1}\mathbf{1}^\top & x_l\mathbf{1}\mathbf{1}^\top \end{bmatrix}. \tag{54}$$

where $\mathbf{1}\mathbf{1}^\top$ is the all-ones matrix of size $\frac{n}{2} \times \frac{n}{2}$.

E.2.1 Recurrence Relation

By substituting the block structures of A and $K^{(l)}$ into the evolution equation and noting that $(\mathbf{1}\mathbf{1}^\top)^2 = \frac{n}{2}\mathbf{1}\mathbf{1}^\top$ (yielding a factor of $(\frac{n}{2})^2$ for the triple products), we derive the following recurrence relations for the scalar components x_{l+1} and y_{l+1}:

$$x_{l+1} = \left(\frac{n}{2}\right)^2 \left[2(p^2+q^2)x_l^2 + 2pq(y_l x_l + y_l^2)\right] \tag{55}$$
$$y_{l+1} = \left(\frac{n}{2}\right)^2 \left[2(p^2+q^2)y_l^2 + 2pq(y_l x_l + x_l^2)\right] \tag{56}$$

We introduce a change of variables:

$$S_{l+1} = x_{l+1} + y_{l+1} \tag{57}$$
$$\Delta_{l+1} = x_{l+1} - y_{l+1} \tag{58}$$

Substituting the recurrence relations for x_{l+1} and y_{l+1}, the evolution of the discrepancy Δ_{l+1} is given by:

$$\begin{aligned}
\Delta_{l+1} &= \left(\frac{n}{2}\right)^2 \left[[2(p^2+q^2)x_l^2 + 2pq(y_l x_l + y_l^2)] - [2(p^2+q^2)y_l^2 + 2pq(y_l x_l + x_l^2)]\right] \\
&= \left(\frac{n}{2}\right)^2 \left[2(p^2+q^2)(x_l^2 - y_l^2) - 2pq(x_l^2 - y_l^2)\right] \\
&= 2\left(\frac{n}{2}\right)^2 (p^2 - pq + q^2)(x_l - y_l)(x_l + y_l) \\
&= 2\left(\frac{n}{2}\right)^2 (p^2 - pq + q^2)\Delta_l S_l
\end{aligned} \tag{59}$$

Similarly, we derive the evolution of the total kernel sum S_{l+1}:

$$\begin{aligned} S_{l+1} &= \left(\frac{n}{2}\right)^2 \left[[2(p^2+q^2)x_l^2 + 2pq(y_l x_l + y_l^2)] + [2(p^2+q^2)y_l^2 + 2pq(y_l x_l + x_l^2)]\right] \\ &= \left(\frac{n}{2}\right)^2 \left[2(p^2+q^2)(x_l^2+y_l^2) + 2pq(x_l+y_l)^2\right] \\ &= \left(\frac{n}{2}\right)^2 \left[(p^2+q^2)((x_l-y_l)^2 + (x_l+y_l)^2) + 2pq(x_l+y_l)^2\right] \\ &= \left(\frac{n}{2}\right)^2 (p^2+q^2)\Delta_l^2 + \left(\frac{n}{2}\right)^2 (p+q)^2 S_l^2 \end{aligned} \tag{60}$$

In order to simplify the problem to a single variable recurrence relation, we express Δ_l as a product of previous terms. By iteratively applying the recurrence $\Delta_{j+1} = 2(\frac{n}{2})^2(p^2 - pq + q^2)\Delta_j S_j$ starting from $j = 0$, we obtain:

$$\Delta_l = \Delta_0 \prod_{j=0}^{l-1} \left[2\left(\frac{n}{2}\right)^2 (p^2 - pq + q^2) S_j\right] = \Delta_0 [2\left(\frac{n}{2}\right)^2 (p^2 - pq + q^2)]^l \prod_{j=0}^{l-1} S_j \tag{61}$$

Substituting the square of this expression into the equation for S_{l+1}, we have:

$$\begin{aligned} S_{l+1} &= \left(\frac{n}{2}\right)^2 (p^2+q^2) \left(\Delta_0 [2\left(\frac{n}{2}\right)^2 (p^2 - pq + q^2)]^l \prod_{j=0}^{l-1} S_j\right)^2 + \left(\frac{n}{2}\right)^2 (p+q)^2 S_l^2 \\ &= \left(\frac{n}{2}\right)^2 (p^2+q^2) [2\left(\frac{n}{2}\right)^2 (p^2 - pq + q^2)]^{2l} \Delta_0^2 \prod_{j=0}^{l-1} S_j^2 + \left(\frac{n}{2}\right)^2 (p+q)^2 S_l^2 \end{aligned} \tag{62}$$

Using the initial condition $\Delta_0 = x_0 - y_0$, we arrive at the final one-variable recurrence relation:

$$S_{l+1} = \left(\frac{n}{2}\right)^2 (p^2+q^2) [2\left(\frac{n}{2}\right)^2 (p^2 - pq + q^2)]^{2l} (x_0 - y_0)^2 \prod_{j=0}^{l-1} S_j^2 + \left(\frac{n}{2}\right)^2 (p+q)^2 S_l^2 \tag{63}$$

This equation demonstrates that the total kernel sum at layer $l+1$ depends on the entire history of the sums from previous layers $\{S_0, \ldots, S_l\}$, weighted by the SBM parameters and the initial discrepancy.

E.2.2 Analytical Solution of the Recurrence Relation

Consider the sequence S_l:

$$S_l = \alpha \prod_{j=0}^{l-2} S_j^2 + \beta S_{l-1}^2, \tag{64}$$

where $\alpha = \left(\frac{n}{2}\right)^2 (p^2+q^2) \left[2\left(\frac{n}{2}\right)^2 (p^2 - pq + q^2)\right]^{2(l-1)} (x_0 - y_0)^2$ and $\beta = \left(\frac{n}{2}\right)^2 (p+q)^2$.

To eliminate the product term, we examine S_{l-1}. By shifting the index $l \to l-1$:

$$S_{l-1} = \alpha \prod_{j=0}^{l-3} S_j^2 + \beta S_{l-2}^2 \tag{65}$$

Rearranging to isolate the product:

$$\alpha \prod_{j=0}^{l-3} S_j^2 = S_{l-1} - \beta S_{l-2}^2 \tag{66}$$

Returning to S_l and peeling off the last term:

$$S_l = \left(\alpha \prod_{j=0}^{l-3} S_j^2 \right) S_{l-2}^2 + \beta S_{l-1}^2 \tag{67}$$

Substituting Equation 66:

$$S_l = (S_{l-1} - \beta S_{l-2}^2) S_{l-2}^2 + \beta S_{l-1}^2 \tag{68}$$

Distributing the S_{l-2}^2 term:

$$S_l = S_{l-1} S_{l-2}^2 - \beta S_{l-2}^4 + \beta S_{l-1}^2 \tag{69}$$

Using the ansatz $S_l = k\gamma^{2^l}$, we derive the characteristic polynomial for k:

$$\beta k^4 - k^3 - \beta k^2 + k = 0 \tag{70}$$

The non-trivial solutions are 1 and $1/\beta$. Given S_0 and S_1, the specific choice of k ensures consistency with the initial data. For the remainder of the paper we will consider $k = 1/\beta = [(\frac{n}{2})^2 (p+q)^2]^{-1}$. Using $S_0 = x_0 + y_0$:

$$S_l = k \left(\frac{x_0 + y_0}{k} \right)^{2^l} = \frac{1}{\left(\frac{n}{2}\right)^2 (p+q)^2} \left[\left(\frac{n}{2}\right)^2 (p+q)^2 (x_0 + y_0) \right]^{2^l} \tag{71}$$

The difference is:

$$\Delta_l = [2 \left(\frac{n}{2}\right)^2 (p^2 - pq + q^2)]^l (x_0 - y_0) \prod_{j=0}^{l-1} S_j \tag{72}$$

The product simplifies to $\prod_{j=0}^{l-1} S_j = k^{-2^l + l + 1} (x_0 + y_0)^{2^l - 1}$. Thus:

$$\Delta_l = [2 \left(\frac{n}{2}\right)^2 (p^2 - pq + q^2)]^l (x_0 - y_0) k^{-2^l + l + 1} (x_0 + y_0)^{2^l - 1} \tag{73}$$

E.3 FINAL EXPRESSIONS FOR x_l AND y_l

Solving $x_l = \frac{S_l + \Delta_l}{2}$ and $y_l = \frac{S_l - \Delta_l}{2}$ by defining:

$$G = (x_0 + y_0)(p+q)^2 \left(\frac{n}{2}\right)^2 \quad \text{and} \quad F = 2\frac{p^2 - pq + q^2}{(p+q)^2} \tag{74}$$

we obtain:

$$x_l = \frac{1}{2} \left[\frac{G^{2^l}}{\left(\frac{n}{2}\right)^2 (p+q)^2} \right] \left[1 + \left(\frac{x_0 - y_0}{x_0 + y_0} \right) F^l \right] \tag{75}$$

$$y_l = \frac{1}{2} \left[\frac{G^{2^l}}{\left(\frac{n}{2}\right)^2 (p+q)^2} \right] \left[1 - \left(\frac{x_0 - y_0}{x_0 + y_0} \right) F^l \right] \tag{76}$$

E.3.1 Proof of Corollary 2

Proof. Using the expressions for x_ℓ and y_ℓ from Corollary 38, the ratio of inter- to intra-community similarity is:

$$\frac{y_\ell}{x_\ell} = \frac{1 - \left(\frac{x_0 - y_0}{x_0 + y_0} \right) F^\ell}{1 + \left(\frac{x_0 - y_0}{x_0 + y_0} \right) F^\ell}.$$

The normalized kernel, defined by dividing by the average trace $\frac{1}{n}\text{tr}(K^{(\ell)}) = x_\ell$, is given by:

$$\frac{K^{(\ell)}}{\frac{1}{n}\text{tr}(K^{(\ell)})} = \begin{pmatrix} \mathbf{1}\mathbf{1}^\top & \left(\frac{1-\left(\frac{x_0-y_0}{x_0+y_0}\right)F^\ell}{1+\left(\frac{x_0-y_0}{x_0+y_0}\right)F^\ell}\right)\mathbf{1}\mathbf{1}^\top \\ \left(\frac{1-\left(\frac{x_0-y_0}{x_0+y_0}\right)F^\ell}{1+\left(\frac{x_0-y_0}{x_0+y_0}\right)F^\ell}\right)\mathbf{1}\mathbf{1}^\top & \mathbf{1}\mathbf{1}^\top \end{pmatrix}.$$

As $\ell \to \infty$, if $F < 1$, the second term of y_ℓ vanishes, leading to rank collapse (oversmoothing). However, if $F \geq 1$, the second term does not vanish. Setting $F = 2\frac{p^2-pq+q^2}{(p+q)^2} \geq 1$ yields:

$$\begin{aligned} 2(p^2 - pq + q^2) &\geq p^2 + 2pq + q^2 \\ p^2 - 4pq + q^2 &\geq 0 \end{aligned}$$

This condition is satisfied when the ratio p/q is sufficiently large (or small), specifically $p/q \geq 2 + \sqrt{3}$ or $p/q \leq 2 - \sqrt{3}$. Under this condition, the limit matrix retains rank 2, preserving the community structure. □

E.4 GRAPHORMER-GP KERNEL UNDER SBM

Corollary 39 (Graphormer Kernel under SBM)**.** *Consider the Graphormer kernel from Corollary 11 without the bias term, and assume that the positional encodings capture all structural information of the graph. In the SBM setting, this implies that the spatial relation matrix coincides with the adjacency matrix, i.e., $R = A$. Under SBM, the block entries $\tilde{x}_\ell$ and $\tilde{y}_\ell$ of $\tilde{K}^{(\ell)}$ are determined by the recurrence:*

$$\begin{aligned} \tilde{x}_\ell &= \alpha^\ell x_0 + (1-\alpha^\ell)p, \\ \tilde{y}_\ell &= \alpha^\ell y_0 + (1-\alpha^\ell)q, \end{aligned}$$

where the normalized kernel is $K^{(\ell)} = \tilde{K}^{(\ell)} \cdot Z_{\ell-1}$, with $Z_{\ell-1} = Tr(A^\top(\tilde{K}^{(\ell-1)} \odot \tilde{K}^{(\ell-1)}))$.

Since our analysis focuses on whether the kernel preserves the block structure induced by the SBM, the trace term does not play a role. Consequently, it suffices to study the evolution of $\tilde{K}^{(\ell)}$.

Proof. By the recursive definition of the Graphormer kernel, the unnormalized kernel $\tilde{K}^{(\ell)}$ is a weighted sum of the previous layer's normalized kernel $K^{(\ell-1)}$ and the spatial relation matrix $R = A$. Under the SBM assumption, both $K^{(\ell-1)}$ and A are block-constant. Solving these first-order recurrences with initial conditions x_0, y_0 yields the closed-form expressions:

$$\tilde{K}^{(\ell)} = \begin{pmatrix} \alpha^\ell x_0 + (1-\alpha^\ell)p & \alpha^\ell y_0 + (1-\alpha^\ell)q \\ \alpha^\ell y_0 + (1-\alpha^\ell)q & \alpha^\ell x_0 + (1-\alpha^\ell)p \end{pmatrix}$$

The normalized kernel $K^{(\ell)}_{ab}$ is obtained by multiplying the previous unnormalized entries by the sum over the entries of the spatial relation matrix (here A):

$$K^{(\ell)}_{ab} = \tilde{K}^{(\ell-1)}_{ab} \sum_{i,j \in V} A_{ij}\tilde{K}ij^{(\ell-1)}\tilde{K}ij^{(\ell-1)} = \tilde{K}^{(\ell-1)}_{ab} \cdot \text{Tr}\left(A^\top(\tilde{K}^{(\ell-1)} \odot \tilde{K}^{(\ell-1)})\right)$$

Hence the normalized kernel is defined as $K^{(\ell)} = \tilde{K}^{(\ell)} \cdot Z_{\ell-1}$ with $Z_{\ell-1} = \text{Tr}(A^\top(\tilde{K}^{(\ell-1)} \odot \tilde{K}^{(\ell-1)}))$. □

E.5 SPECFORMER-GP KERNEL UNDER SBM

Corollary 40 (SpecFormer Kernel under SBM)**.** *Consider the Specformer kernel from Theorem 12, defined as For $\ell \geq 1$, the block entries of $K^{(\ell)}$ admit the following closed-form expressions:*

$$\begin{aligned} x_\ell &= \frac{1}{2}\left[(x_0+y_0)\bar{\lambda}_1^{2\ell} + (x_0-y_0)\bar{\lambda}_2^{2\ell}\right], \\ y_\ell &= \frac{1}{2}\left[(x_0+y_0)\bar{\lambda}_1^{2\ell} - (x_0-y_0)\bar{\lambda}_2^{2\ell}\right], \end{aligned}$$

where $\bar{\lambda}_1$ and $\bar{\lambda}_2$ $\bar{\lambda}_2$ denote the learned eigenvalues of the Specformer's spectral filter; these parameters determine the cross-covariance K_λ.

Proof. As detailed in Theorem 12, the Specformer kernel at layer ℓ is obtained via $K^{(\ell)} = U(K_\lambda \odot (U^\top K^{(\ell-1)} U))U^\top$. We define $M^{(\ell)} := U^\top K^{(\ell)} U$. Given the recurrence relation, we have:

$$M^{(\ell)} = K_\lambda \odot M^{(\ell-1)} = (K_\lambda \odot K_\lambda \odot \cdots \odot K_\lambda) \odot M^{(0)} = (K_\lambda)^{\odot \ell} \odot M^{(0)},$$

where $(K_\lambda)_{ab} = \bar{\lambda}_a \bar{\lambda}_b$. Thus, the ℓ-th Hadamard power is $((K_\lambda)^{\odot \ell})_{ab} = \bar{\lambda}_a^\ell \bar{\lambda}_b^\ell$. For an SBM with two equal communities, $U = [v_1\ v_2]$ with $v_1 = \frac{1}{\sqrt{n}}\mathbf{1}_n$ and $v_2 = \frac{1}{\sqrt{n}}[\mathbf{1}_{n/2}^\top, -\mathbf{1}_{n/2}^\top]^\top$.

The initial spectral kernel $M^{(0)} = U^\top K^{(0)} U$ is a diagonal matrix containing the eigenvalues of the block-constant kernel $K^{(0)}$:

$$M^{(0)} = \begin{pmatrix} \frac{n}{2}(x_0 + y_0) & 0 \\ 0 & \frac{n}{2}(x_0 - y_0) \end{pmatrix}.$$

The non-zero diagonal entries of $M^{(\ell)}$ evolve as:

$$M_{11}^{(\ell)} = (\bar{\lambda}_1 \bar{\lambda}_1)^\ell \frac{n}{2}(x_0 + y_0) = \bar{\lambda}_1^{2\ell} \frac{n}{2}(x_0 + y_0),$$
$$M_{22}^{(\ell)} = (\bar{\lambda}_2 \bar{\lambda}_2)^\ell \frac{n}{2}(x_0 - y_0) = \bar{\lambda}_2^{2\ell} \frac{n}{2}(x_0 - y_0).$$

Reconstructing the kernel in the spatial domain via $K^{(\ell)} = M_{11}^{(\ell)} v_1 v_1^\top + M_{22}^{(\ell)} v_2 v_2^\top$, we use the outer product definitions:

$$v_1 v_1^\top = \frac{1}{n}\begin{pmatrix} \mathbf{J} & \mathbf{J} \\ \mathbf{J} & \mathbf{J} \end{pmatrix}, \quad v_2 v_2^\top = \frac{1}{n}\begin{pmatrix} \mathbf{J} & -\mathbf{J} \\ -\mathbf{J} & \mathbf{J} \end{pmatrix},$$

where $\mathbf{J} = \mathbf{1}\mathbf{1}^\top$ of size $\frac{n}{2} \times \frac{n}{2}$. Substituting these yields:

$$K^{(\ell)} = \frac{\bar{\lambda}_1^{2\ell}(x_0 + y_0)}{2}\begin{pmatrix} \mathbf{J} & \mathbf{J} \\ \mathbf{J} & \mathbf{J} \end{pmatrix} + \frac{\bar{\lambda}_2^{2\ell}(x_0 - y_0)}{2}\begin{pmatrix} \mathbf{J} & -\mathbf{J} \\ -\mathbf{J} & \mathbf{J} \end{pmatrix}.$$

Identifying the diagonal block scalar x_ℓ and the off-diagonal block scalar y_ℓ gives the desired result. □

Proof of Remark 5 Using the expressions for x_ℓ and y_ℓ from Corollary 40, the ratio of inter- to intra-community similarity is:

$$\frac{y_\ell}{x_\ell} = \frac{(x_0 + y_0)\bar{\lambda}_1^{2\ell} - (x_0 - y_0)\bar{\lambda}_2^{2\ell}}{(x_0 + y_0)\bar{\lambda}_1^{2\ell} + (x_0 - y_0)\bar{\lambda}_2^{2\ell}} = \frac{\left(\frac{x_0+y_0}{x_0-y_0}\right)\left(\frac{\bar{\lambda}_1}{\bar{\lambda}_2}\right)^{2\ell} - 1}{\left(\frac{x_0+y_0}{x_0-y_0}\right)\left(\frac{\bar{\lambda}_1}{\bar{\lambda}_2}\right)^{2\ell} + 1}.$$

If $|\bar{\lambda}_2| \geq |\bar{\lambda}_1|$, the term $(\bar{\lambda}_1/\bar{\lambda}_2)^{2\ell}$ either vanishes or stays constant as $\ell \to \infty$. Consequently, the ratio y_ℓ/x_ℓ does not converge to 1, and the kernel maintains its rank-2 structure. This allows Specformer to avoid oversmoothing by amplifying or preserving the second spectral direction v_2, which corresponds to the community partition.

F SUPPLEMENTARY MATERIAL FOR EXPERIMENTS

The code used to reproduce all experiments is available at `https://figshare.com/s/4ad5245d4d6c405f46f0`. In this section we provide additional information regarding datasets and experiment detauls. In Section F.3 we present additional experiments.

F.1 DATASETS

We conduct experiments on both real-world benchmarks and synthetic datasets to evaluate model performance under varying conditions.

Benchmark Datasets We utilize two widely recognized benchmarks: **PubMed** [Sen et al., 2008], a homopholic citation network, and **Chameleon** [Rozemberczki et al., 2019], a heterophilic Wikipedia page-page network.

Synthetic Data To analyze the interplay between graph topology and feature information, we employ two synthetic generative models, considering two-class versions for both:

- **SBM (Random Features):** The graph is generated according to the Stochastic Block Model (SBM) rule, where the probability of an edge existing between nodes in the same community is p and between different communities is q. Node features are drawn from an independent Gaussian distribution, making them uninformative of the community structure.
- **CSBM (Aligned Features):** In the Contextual Stochastic Block Model (CSBM), node features are generated as a mixture of Gaussians such that the feature distributions are aligned with the graph's community labels. This represents a more realistic scenario common in literature where the graph structure and node attributes provide complementary information regarding the underlying classes.

All datasets are accessed via the PyTorch Geometric (PyG) data loaders. We follow the standard PyG splits for each dataset. In particular, for datasets that come with multiple predefined splits (notably the heterophilous benchmarks), we run the full training–validation–test procedure on every split and report the mean of the test accuracy across splits.

Table 6: Dataset statistics for node-level classification tasks.

Dataset	Level	Splits	Nodes	Edges	Classes	Features	Homophily
Pubmed	Node	1	19,717	44,338	3	500	0.80
Chameleon	Node	10	2,277	31,421	10	2,325	0.23

Table 7: Datasets for node classification on heterogeneous graphs.

Dataset	Nodes	Edges	Edge type	Features	Training	Validation	Test
DBLP	18405	67946	4	334	800	400	2857
ACM	8994	25922	4	1902	600	300	2125
IMDB	12772	37288	4	1256	300	300	2339

F.2 EXPERIMENTAL ENVIRONMENT, TRAINING DETAILS, AND HYPERPARAMETERS

We compare several graph GP kernels derived from the infinite-width limits of representative architectures, including GCN-GP, GAT-GP, Graphormer-GP, and Specformer-GP. For each kernel, we perform node classification using a GP-style multi-output regression setup: class labels are encoded as one-hot targets, and predictions are obtained by selecting the class with the largest output score.

Replacing the softmax with the identity map requires an explicit normalisation step; consistent with Hron et al. [2020], we incorporate a per-node LayerNorm. Concretely, at the end of each layer we apply the following LayerNorm kernel: $K_{ab}\left[K_{aa}K_{bb}\right]^{-1/2}$. All GP models in our experiments include a LayerNorm module at the end of every layer.

For each split, model selection is conducted on the validation set (e.g., via a grid search over the ridge regularization coefficient) and the resulting configuration is evaluated on the test set. For datasets with multiple splits, we aggregate results by reporting the mean across splits.

F.3 ADDITIONAL EXPERIMENTS

Distribution Experiments Figure 4 complements Figure 1 in the introduction by further illustrating the distribution of eigenvalue encoding of Specformer.

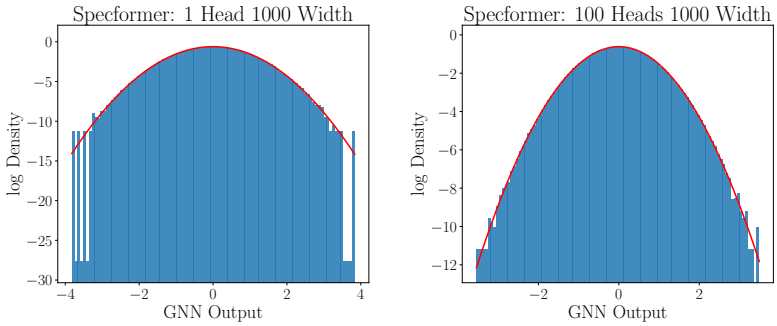

Figure 4: Histogram of the eigenvalue encoding of Specformer for different number of heads. The output distribution converges to a Gaussian (red line) fitted with mean and variance of the empirical distribution when both width and number of heads are large.

Oversmoothing Experiments Figure 5 further illustrates the effect of positional encodings on oversmoothing. In both datasets, Laplacian-based positional encodings consistently achieve high test accuracy across depth. In the CSBM setting, where node features are already informative, these encodings have perfect accuracy.

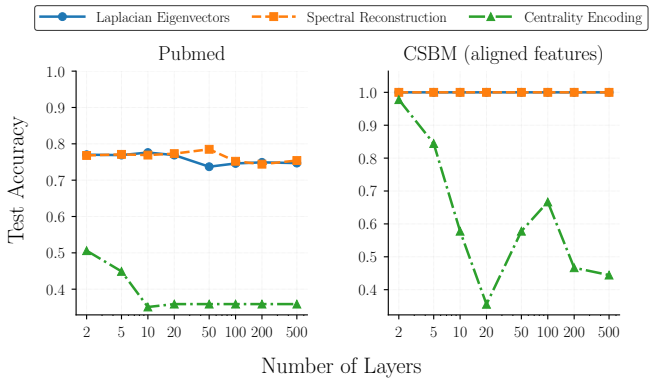

Figure 5: **Oversmoothing behaviour and the impact of positional encodings.** Test accuracy of **Graphormer-GP** on Pubmed (left) and CSBM (right) as a function of the number of layers.

Empirical and Theoretical Kernels under CSBM To empirically validate our theoretical predictions regarding oversmoothing (Table 3), we evaluate the empirical kernels of trained finite-width GAT under varying signal regimes.

Specifically, we test the network's capacity to preserve class boundaries on the CSBM across two distinct regimes:

1. **Strong Community Signal regime** ($F \geq 1$)**:** The graph has a pronounced community structure, with the within- and cross-community connection probabilities sufficiently different, i.e., the ratio p/q is bounded away from one. This condition holds both in strongly homophilic and strongly heterophilic settings, and failure occurs only when p and q become close, corresponding to a regime with weak or no community structure rather than a specific type of homophily or heterophily
2. **Weak Community Signal regime** ($F < 1$)**:** The within- and cross-community connection probabilities are similar ($p \approx q$), resulting in a weak community signal and making the graph structure less informative for distinguishing the communities.

We trained finite-width GAT models on both configurations and computed their empirical node-level kernel representations at deep layers. Figure 6 illustrates the results.

The empirical findings match the theoretical predictions. In the strong-signal regime, the empirical kernel perfectly preserves the underlying block-diagonal community structure, matching the formulas for x_l and y_l in Table 3 and demonstrating that representations remain highly discriminative even at depth. Conversely, in the weak-signal regime, the community boundaries fade into a uniform representation space. **On Linear vs. Softmax Attention** While our main closed-form

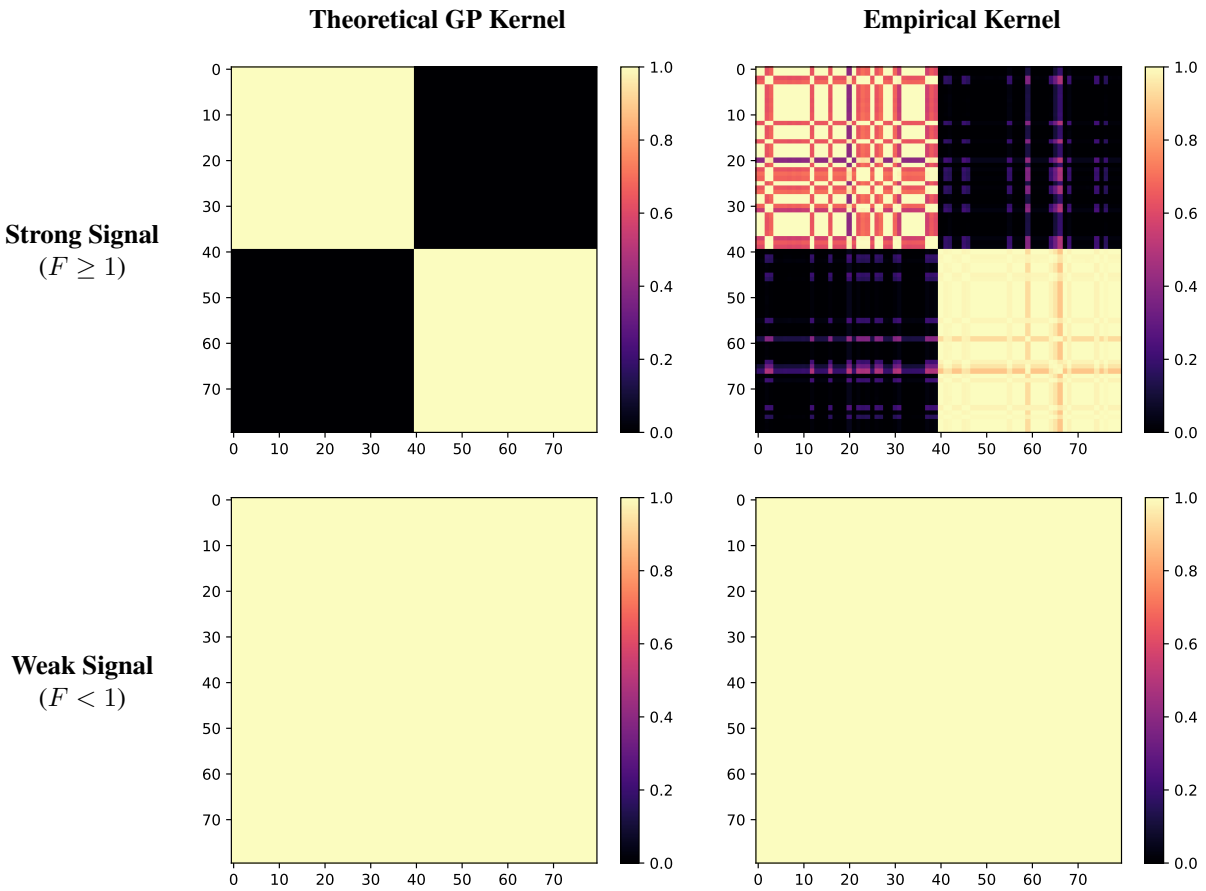

Figure 6: **Comparison of Empirical and Theoretical Kernels under CSBM.** This figure compares the empirical kernel from trained GAT with 15 layers with the theoretical kernel from Table 3 across two distinct regimes of the Contextual Stochastic Block Model (CSBM). (Top) In the *Strong Signal* regime ($F = 1.903 \geq 1$), where classes are adequately separated, both the empirical and theoretical kernels recover the underlying block structure of the graph. (Bottom) In the *Weak Signal* regime ($F = 0.512 < 1$), the class information is lost.

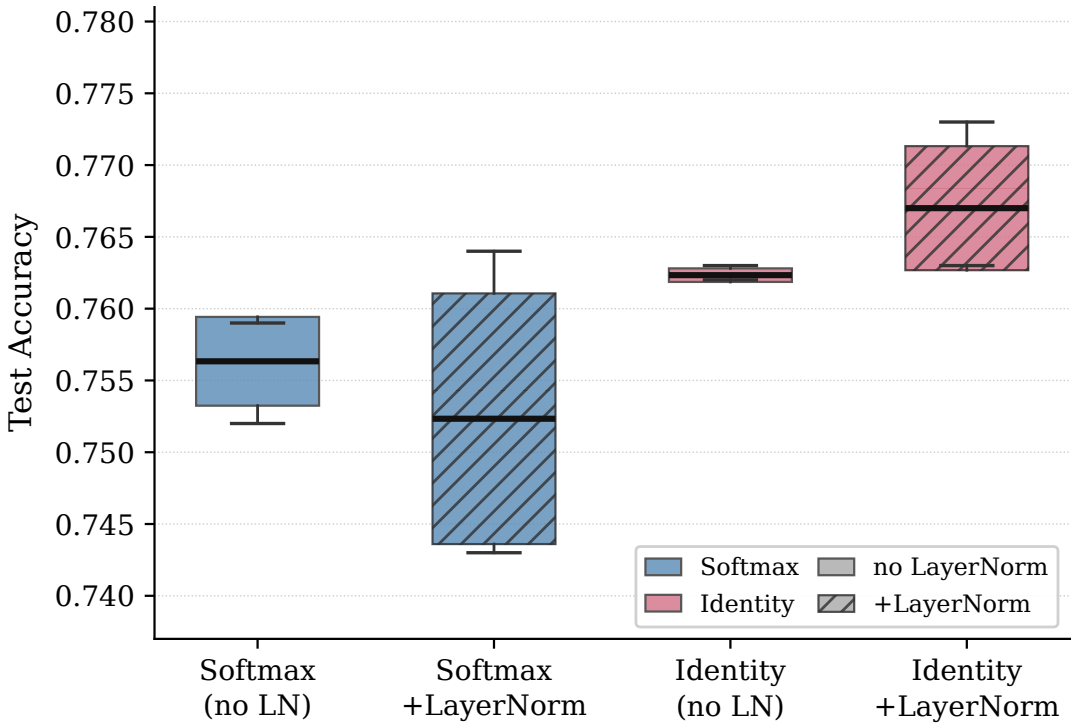

Figure 7: **Comparison of attention mechanisms on the PubMed dataset.** This figure presents the test accuracy of Graph Attention Network (GAT) for different choices of attention nonlinearity and layer-normalization (LN) in a two-layer model. The results show that linear attention is a strong alternative to softmax-based attention and achieves higher test accuracy on this dataset.

analysis relies on linear attention (i.e., the identity activation function), standard practical architectures often employ softmax attention. Mathematically, deriving closed-form GP kernels for softmax attention remains an open challenge due to the lack of tractable recursions.

To justify our theoretical focus, we empirically compare the performance of linear and softmax attention within a two-layer GAT on the PubMed dataset. As illustrated in Figure 7, linear attention is an acceptable alternative, consistently achieving comparable or superior test accuracy compared to its softmax counterpart—particularly when combined with Layer Normalization (LN). Analogous empirical findings and theoretical limitations regarding softmax attention have also been documented in the analysis of standard transformers [Hron et al., 2020]. Therefore, focusing on linear attention allows for theoretical tractability that is still empirically relevant.