UNDERSTANDING GRAPH NEURAL NETWORK GENERALIZATION VIA MESSAGE-PASSING STRUCTURAL COMPLEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph neural networks (GNNs) have become a fundamental paradigm for learning from graph-structured data. Unlike conventional neural networks, their predictions depend not only on learnable parameters but also on the graph structures involved in message passing, making their generalization particularly sensitive to structural variations. However, existing analyses of GNN generalization mainly focus on parameter complexity, while how the complexity of message-passing structures affects generalization remains insufficiently understood. To characterize this effect, we define message-passing structural complexity and derive a structure-aware generalization bound within a transductive Rademacher-complexity framework, theoretically revealing structural complexity as a fundamental factor in GNN generalization. To mitigate the effect of structural complexity on generalization, we develop a graph neural network method with structural entropy regularization, which controls message-passing structural complexity through structural entropy. We theoretically establish an upper-bound relation showing that structural entropy serves as a differentiable surrogate for cross-class information-propagation complexity, thereby enabling principled control of potentially harmful message passing. By incorporating the structural regularization term into several representative GNN architectures, experiments on benchmark datasets show that reducing the structural complexity can consistently improve GNN generalization.

1 INTRODUCTION

Graph neural networks (GNNs) have emerged as a fundamental tool for learning from graph-structured data, achieving strong performance across a wide range of applications (Fan et al., 2025; Yuan et al., 2025; Ju et al., 2025; Yin et al., 2026). However, their dependence on graph structure also makes understanding their generalization fundamentally different from that of conventional neural networks. In classical learning theory, generalization is commonly characterized through model complexity, with quantities such as parameter norms and network depth controlling the capacity of the hypothesis class (Ledent et al., 2021; Jacot et al., 2025; Li & Lyu, 2026). For GNNs, predictions depend not only on learnable parameters but also on the structural relations through which information is propagated. Exploiting more structural relations during prediction may increase the predictor's dependence on graph-specific aggregation patterns, thereby increasing the risk of structural overfitting. This suggests that GNN generalization is particularly sensitive to the complexity of message-passing relations. This motivates the central question of this work: *How does the complexity of message-passing structures affect GNN generalization?*

Existing studies have investigated GNN generalization using a variety of learning-theoretic frameworks. Stability-based analyses (Verma & Zhang, 2019; Yang et al., 2025), Rademacher-complexity analyses (Garg et al., 2020; Esser et al., 2021; Lv, 2021), and PAC-Bayesian approaches (Ju et al., 2023; Brilliantov et al., 2024) have examined how factors such as parameter norms, network depth, graph propagation operators, and their spectral properties affect generalization. However, these analyses typically do not explicitly characterize the structural capacity associated with how many effective message-passing relations are actually exploited by the predictor. As a result, the role of effective structural participation in GNN generalization remains insufficiently understood.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To characterize this missing structural capacity, we define message-passing structural complexity as the number of effective aggregation relations exploited by the predictor. Within a transductive Rademacher-complexity framework, we derive a structure-aware generalization bound that reveals an explicit dependence of the generalization guarantee on message-passing structural complexity. Holding the remaining quantities fixed, increasing structural complexity loosens the corresponding generalization upper bound. This dependence motivates explicit control of the effective aggregation structure.

Building on this analysis, we further investigate how message-passing structural complexity can be regulated selectively. Since different structural relations may contribute differently to prediction, we decompose structural complexity into class-consistent and cross-class components and use the latter as a task-specific proxy for potentially harmful structural complexity under homophily. To translate this task-specific criterion into a differentiable training mechanism, we develop Structural Entropy Regularization (SER) to penalize class-inconsistent message passing. We derive a pointwise upper-bound relation between the proposed structural-entropy quantity and the cross-class component of structural complexity, showing that structural entropy provides a principled differentiable mechanism for selectively regulating potentially harmful message passing. By incorporating this structural regularization into representative GNN architectures, experiments on benchmark datasets show that controlling message-passing structural complexity can lead to improved GNN generalization.

2 PRELIMINARIES

Notation. We consider an undirected graph $G = (V, E)$, where $V = [n] := \{1, \dots, n\}$ denotes the set of node indices with $|V| = n$, and $E \subseteq V \times V$ denotes the set of edges. The adjacency matrix of the graph is denoted by $A \in \{0,1\}^{n\times n}$. For each node $v \in V$, we define its neighborhood as $\mathcal{N}(v) = \{u \in V \mid A_{vu} = 1\}$. Let $D \in \mathbb{R}^{n\times n}$ denote the corresponding degree matrix, where $D_{ii} = \sum_{j=1}^{n} A_{ij}$. The node features are represented by a matrix $X \in \mathbb{R}^{n\times d}$, where $x_v \in \mathbb{R}^d$ is the feature vector of node v. The node labels are denoted by $Y = \{y_v\}_{v\in V}$, where $y_v \in \mathcal{Y}$ represents the label of node v. We focus on semi-supervised node classification on a fixed graph population (G, X, Y). The labeled set $\mathcal{S} \subset V$ with $|\mathcal{S}| = m$ is sampled uniformly from all subsets of V of size m, with labels observed only on $\mathcal{S}$, and $\mathcal{U} = V \setminus \mathcal{S}$ contains the remaining $n - m$ unlabeled nodes. Let $\mathcal{F}$ denote the GNN hypothesis class, specified independently of the sampled split $\mathcal{S}$, where each $f \in \mathcal{F}$ produces node predictions $f_v \in \mathbb{R}$. Given a loss function $\ell(\cdot, y)$, the empirical risk over the labeled nodes and the transductive risk over the unlabeled nodes are defined as $\hat{\epsilon}(f) = \frac{1}{m}\sum_{v\in\mathcal{S}} \ell(f_v, y_v)$, $\epsilon(f) = \frac{1}{n-m}\sum_{v\in\mathcal{U}} \ell(f_v, y_v)$. The generalization gap is $\epsilon(f) - \hat{\epsilon}(f)$. We use $\|\cdot\|_2$, $\|\cdot\|_F$, and $\|\cdot\|_0$ for the spectral, Frobenius, and entrywise ℓ_0 norms.

Graph Neural Networks. GNNs learn node representations by iteratively propagating and aggregating information over graph relations. Let $\Omega^{(k)} = [\omega_{vu}^{(k)}]_{v,u\in V} \in \mathbb{R}^{n\times n}$ denote the structural propagation operator at layer k, where $\omega_{vu}^{(k)}$ is the propagation weight from node u to node v. The layer-wise node representation matrix is updated as:

$$H^{(k)} = \sigma\left(\Omega^{(k)} H^{(k-1)} W^{(k)}\right), \qquad H^{(0)} = X, \tag{1}$$

where $h_v^{(k)} \in \mathbb{R}^{d_k}$ is a column vector, $H^{(k)} \in \mathbb{R}^{n\times d_k}$ stores $\left(h_v^{(k)}\right)^\top$ as its v-th row, and $W^{(k)} \in \mathbb{R}^{d_{k-1}\times d_k}$ is a trainable weight matrix, and $\sigma(\cdot)$ is an element-wise nonlinear activation function. Representative instantiations of this framework, including graph convolutional networks (GCNs), graph attention networks (GATs), and graph transformers (GTs), are introduced in Appendix A.

Generalization Theory for Graph Neural Networks. Understanding the generalization of GNNs has attracted increasing attention in recent years. Unlike classical learning settings that assume independent and identically distributed samples, node-level prediction on graphs involves edge-induced dependencies, making standard generalization analysis challenging. Early theoretical studies extend algorithmic stability frameworks to GNNs. Verma & Zhang (2019) analyze GCN generalization by treating sampled local subgraphs as approximately independent training samples. However, this framework relies on the assumption that sampled subgraphs are sufficiently independent, which may not hold for densely connected graphs. To address this challenge, Lv (2021) employ Rademacher complexity to analyze node-level generalization on a fixed graph, where the graph structure and

node features are treated as fixed. However, these complexity terms are still mainly characterized by model-dependent quantities, such as parameter norms, while the influence of graph structure is captured indirectly through spectral properties of graph propagation operators. Beyond complexity-based analyses, recent works investigate GNN generalization from PAC-Bayesian perspectives (Ma et al., 2021; Choi et al., 2026), optimization dynamics (Tang & Liu, 2023), and graph distribution shifts (Liu et al., 2023; Wang et al., 2024; 2025). Despite these advances, the structural capacity associated with the number of effective aggregation relations exploited by the predictor remains insufficiently characterized. We address this gap by explicitly quantifying these relations through message-passing structural complexity.

3 Structural Complexity and GNN Generalization

In this section, we investigate the role of structural complexity in GNN generalization. We characterize structural complexity through the effective aggregation relations exploited during message passing and theoretically show that it enters the characterization of GNN generalization alongside conventional parameter-dependent complexity.

3.1 Message-Passing Structural Complexity

In message-passing GNNs, node representations are updated by aggregating information along graph relations. Beyond learnable parameters, the effective capacity of a GNN also depends on the structural relations actually involved in message passing. To characterize this additional source of capacity, we define message-passing structural complexity by the number of effective aggregation relations exploited by the predictor. Formally, for the propagation operator $\Omega^{(k)}$, the layer-wise structural complexity is defined as $\eta^{(k)} = \left\|\Omega^{(k)}\right\|_0$, which counts the active ordered propagation relations at layer k. This notion characterizes the structural relations effectively exploited by the predictor during message passing, rather than the complexity of the input graph itself.

3.2 How Structural Complexity Enters GNN Generalization

We next examine how structural complexity enters GNN generalization. Since all node features and graph structure are observed while labels are available only for a subset of nodes, we consider the transductive learning setting and characterize the GNN hypothesis class using transductive Rademacher complexity (Oono & Suzuki, 2020; Esser et al., 2021). The following result relates the risk on the unobserved complement $\mathcal{U}$ to the empirical risk on the uniformly sampled training set $\mathcal{S}$.

Lemma 1 (El-Yaniv & Pechyony (2009))**.** *Let $\mathcal{F}$ be a hypothesis class and assume that the loss function $\ell(\cdot, y)$ is L-Lipschitz. Then for $0 < \delta < 1$, with probability at least $1 - \delta$ for all $f \in \mathcal{F}$ the following holds:*

$$\epsilon(f) \leq \hat{\epsilon}(f) + \mathfrak{R}_{m,n}(\ell \circ \mathcal{F}) + c_1 \frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2 \sqrt{\frac{n}{m(n-m)} \log \frac{1}{\delta}}, \tag{2}$$

where $c_1 < 5.05$ and $c_2 < 0.8$ are universal constants.

The definition of transductive Rademacher complexity is provided in Appendix D.1. Eq. (2) reduces the question of how message-passing structure affects generalization to a more specific one: *does the transductive Rademacher complexity of the GNN hypothesis class explicitly depend on the effective structural relations exploited during message passing?* To answer this question, we track the effective aggregation relations when bounding $\mathfrak{R}_{m,n}(\mathcal{F})$. For tractability, we consider a two-layer message-passing GNN to capture the joint structural contribution across successive layers.

Theorem 1. *Consider the hypothesis class $\mathcal{F}$ induced by GNNs with fixed propagation operators. Assume that $\|x_i\|_2 \leq B$ for all nodes i, the activation function σ is 1-Lipschitz with $\sigma(0) = 0$, the propagation weights satisfy $|\omega_{vu}^{(k)}| \leq 1$ for $k = 1, 2$, the weight matrix in the hidden layer satisfies $\left\|W^{(1)}\right\|_2 \leq R_1$, the output weight vector satisfies $\left\|W^{(2)}\right\|_2 \leq R_2$, and the loss function $\ell(\cdot, y)$ is L-Lipschitz continuous. Then for $0 < \delta < 1$, with probability at least $1 - \delta$ for all $f \in \mathcal{F}$ the*

following holds:

$$\epsilon(f) \leq \hat{\epsilon}(f) + LR_1R_2B\sqrt{\frac{2n}{m(n-m)}}\sqrt{\eta^{(1)}\eta^{(2)}} + c_1\frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2\sqrt{\frac{n}{m(n-m)}\log\frac{1}{\delta}}, \tag{3}$$

where $\eta^{(k)} = \left\|\Omega^{(k)}\right\|_0$ *for* $k = 1, 2$ *denotes the structural complexity of each layer, and* $c_1 < 5.05$ *and* $c_2 < 0.8$ *are universal constants.*

3.3 Structural Overfitting: Implications of the Analysis

Theorem 1 shows that the complexity term in the GNN generalization bound jointly depends on the parameter-norm factor R_1R_2 and the structural factor $\sqrt{\eta^{(1)}\eta^{(2)}}$. The former characterizes the flexibility of the hypothesis class induced by learnable parameters, whereas the latter captures the relational capacity available through message passing across successive layers. Thus, even under fixed parameter-norm constraints, exploiting more effective aggregation relations enlarges the complexity term and yields a looser generalization guarantee. This identifies structural complexity as an additional source of effective model capacity beyond parameter complexity and suggests a greater risk of structural overfitting through increased structural participation.

3.4 Architectural Instantiations

We next instantiate the analysis for representative GNNs with fixed and attention-based aggregation.

Generalization under Fixed Aggregation. We first consider GCNs, where message passing is governed by the normalized adjacency matrix $\hat{A}$. In this case, the structural complexity reduces to $\eta = \|\hat{A}\|_0$, and the corresponding generalization bound is given as follows.

Corollary 1. *Consider the hypothesis class* $\mathcal{F}$ *induced by GCNs with a fixed message-passing structure. Under the assumptions of Theorem 1, for* $0 < \delta < 1$*, with probability at least* $1 - \delta$ *for all* $f \in \mathcal{F}$ *the following holds:*

$$\epsilon(f) \leq \hat{\epsilon}(f) + LR_1R_2B\sqrt{\frac{2n}{m(n-m)}}\,\eta + c_1\frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2\sqrt{\frac{n}{m(n-m)}\log\frac{1}{\delta}}, \tag{4}$$

where $\eta = \left\|\hat{A}\right\|_0$ *denotes the structural complexity of the model, and* $c_1 < 5.05$ *and* $c_2 < 0.8$ *are universal constants.*

Generalization under Attention-Based Aggregation. We next consider attention-based architectures, such as GATs and GTs, where the propagation weights depend on the predictor f. At layer k, message passing is governed by an attention matrix $A_f^{(k)} = (\alpha_{vu}^{(k)})$. Since softmax attention typically assigns nonzero weights to all candidate relations, directly counting the support of $A_f^{(k)}$ may include negligible connections. We therefore introduce a threshold $\tau > 0$ and define the thresholded attention matrix as $[\bar{A}_f^{(k)}]_{vu} = \alpha_{vu}^{(k)}$ if $\alpha_{vu}^{(k)} \geq \tau$, and $[\bar{A}_f^{(k)}]_{vu} = 0$ otherwise. The effective structural complexity of a predictor is measured by $\|\bar{A}_f^{(k)}\|_0$, and for a uniform bound we define $\eta_\tau^{(k)} = \sup_{f\in\mathcal{F}} \|\bar{A}_f^{(k)}\|_0$. The generalization bound for a two-layer GAT is then given as follows.

Proposition 1. *Consider the hypothesis class* $\mathcal{F}_{\mathrm{GAT}}$ *induced by GATs with attention-based message passing. Under the input, activation, weight-norm, and loss assumptions of Theorem 1, for* $0 < \delta < 1$*, with probability at least* $1 - \delta$ *for all* $f \in \mathcal{F}_{\mathrm{GAT}}$ *the following holds:*

$$\epsilon(f) \leq \hat{\epsilon}(f) + LR_1R_2B\,n\sqrt{\frac{2}{m(n-m)}}\left(\sqrt{\eta_\tau^{(1)}\eta_\tau^{(2)}} + \mathcal{O}(\tau)\right) + c_1\frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2\sqrt{\frac{n}{m(n-m)}\log\frac{1}{\delta}}, \tag{5}$$

where $\eta_\tau^{(k)} = \sup_{f \in \mathcal{F}_{\mathrm{GAT}}} \|\bar{A}_f^{(k)}\|_0$ *for* $k = 1, 2$ *denotes a uniform upper bound on the effective structural complexity of each layer, and* $c_1 < 5.05$ *and* $c_2 < 0.8$ *are universal constants.*

The extension to multi-head GATs is provided in Appendix D.2, Proposition 2. The analysis also extends to GTs. Despite the additional multi-head attention and normalization components, the resulting bound retains an explicit dependence on effective attention-driven structural participation. The formal result and proof are provided in Appendix D.2, Proposition 3. Together with Corollary 1 and Proposition 1, these results show that the dependence on message-passing structural complexity appears explicitly in the generalization bounds of representative GNNs with both fixed and attention-based aggregation mechanisms, indicating that this dependence is not specific to a particular GNN architecture.

4 Controlling Structural Complexity

In the previous section, we showed that increasing message-passing structural complexity can lead to a looser generalization guarantee. Building on this result, we further investigate how structural complexity can be controlled effectively. Since aggregation relations may play different roles in prediction, effective structural control requires a finer-grained characterization of structural complexity. We therefore focus on identifying and selectively regulating its potentially harmful component.

4.1 From Structural Complexity to Harmful Structural Complexity

For node classification, we instantiate potentially harmful structural complexity according to class consistency. Let $P \in \mathbb{R}^{n \times c}$ denote the class assignment matrix, where P_{vj} represents the probability that node v belongs to class j and $\sum_{j=1}^{c} P_{vj} = 1$. For labeled nodes, P_v is given by the ground-truth one-hot label, while for unlabeled nodes it is given by the model posterior. Accordingly, we decompose the structural complexity at layer k into class-consistent and cross-class components as $\eta^{(k)} = \eta_{\mathrm{inner}}^{(k)} + \eta_{\mathrm{cross}}^{(k)}$, where $\eta_{\mathrm{inner}}^{(k)} = \sum_{v,u \in V} \mathbf{1}\left(\omega_{vu}^{(k)} \neq 0\right) P_v^\top P_u$ denotes the class-consistent structural complexity, and $\eta_{\mathrm{cross}}^{(k)} = \sum_{v,u \in V} \mathbf{1}\left(\omega_{vu}^{(k)} \neq 0\right) \left(1 - P_v^\top P_u\right)$ denotes the cross-class structural complexity. In the homophilous node-classification setting considered here, class-consistent relations are generally more likely to support label-consistent aggregation, whereas the cross-class component can serve as a task-specific proxy for potentially harmful structural complexity.

4.2 Surrogate for Potentially Harmful Structural Complexity

To selectively control potentially harmful structural complexity, we focus on $\eta_{\mathrm{cross}}^{(k)}$ rather than the overall structural complexity. Since the support indicator $\mathbf{1}(\omega_{vu}^{(k)} \neq 0)$ is discontinuous in the propagation weights, $\eta_{\mathrm{cross}}^{(k)}$ is difficult to optimize directly. We therefore construct a differentiable surrogate based on weighted cross-class propagation and establish a pointwise upper-bound relation to $\eta_{\mathrm{cross}}^{(k)}$. For layer k, let $d_v^{(k)} = \sum_u \omega_{vu}^{(k)}$ denote the weighted degree of node v, and define the total volume as $\mathrm{vol}^{(k)}(V) = \sum_v d_v^{(k)}$. For each class j, let C_j denote the soft class represented by the membership weights $\{P_{vj}\}_{v \in V}$. Its weighted volume is defined as $\mathrm{vol}^{(k)}(C_j) = \sum_v P_{vj} d_v^{(k)}$, and the corresponding cross-class propagation weight is $g_{C_j}^{(k)} = \sum_{v \in V} P_{vj} \sum_{u \in \mathcal{N}(v)} \omega_{vu}^{(k)} (1 - P_{uj})$. The following theorem formalizes this pointwise relation.

Theorem 2. *For each layer* k*, consider a fixed propagation operator with nonnegative propagation weights, and let* $\omega_{\min}^{(k)} = \min\left\{\omega_{vu}^{(k)} : \omega_{vu}^{(k)} > 0\right\}$*. Let* $p_j^{(k)} = \frac{\mathrm{vol}^{(k)}(C_j)}{\mathrm{vol}^{(k)}(V)}$ *and* $p_{\max}^{(k)} = \max_{1 \le j \le c} p_j^{(k)}$*, with* $0 < p_j^{(k)} \le p_{\max}^{(k)} < 1$*, and define* $\xi^{(k)} = \frac{\mathrm{vol}^{(k)}(V)}{\omega_{\min}^{(k)} \log\left(1/p_{\max}^{(k)}\right)}$*. Then, the following upper bound holds for the cross-class structural complexity:*

$$\eta_{\mathrm{cross}}^{(k)} \le \xi^{(k)} \left[-\sum_{j=1}^{c} \frac{g_{C_j}^{(k)}}{\mathrm{vol}^{(k)}(V)} \log \frac{\mathrm{vol}^{(k)}(C_j)}{\mathrm{vol}^{(k)}(V)} \right]. \tag{6}$$

Theorem 2 establishes a pointwise upper-bound relation between the cross-class structural complexity and a differentiable class-aware quantity. The latter integrates weighted cross-class propagation with class-level structural-volume information, providing a differentiable criterion for characterizing and penalizing class-inconsistent message passing. Motivated by this theoretical connection, we define the bracketed term in Eq. (6) as the class-aware structural entropy:

$$\mathcal{L}_{\mathrm{SE}}^{(k)} = -\sum_{j=1}^{c} \frac{g_{C_j}^{(k)}}{\mathrm{vol}^{(k)}(V)} \log \frac{\mathrm{vol}^{(k)}(C_j)}{\mathrm{vol}^{(k)}(V)}. \tag{7}$$

4.3 Structural Entropy Regularization

Based on the class-aware structural entropy defined in Eq. (7), we construct SER as a theory-guided regularization mechanism for selectively regulating class-inconsistent message passing. For attention-based architectures, the propagation weights are directly determined by learnable attention coefficients. For fixed-aggregation GCNs, we introduce a learnable edge mask $M^{(k)}$ and define $\Omega^{(k)} = M^{(k)} \odot \hat{A}$, where $\odot$ denotes element-wise multiplication, enabling adaptive regulation of the message-passing structure. Accordingly, for a K-layer GNN, the overall training objective is defined as:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CLS}} + \lambda \sum_{k=1}^{K} \mathcal{L}_{\mathrm{SE}}^{(k)}, \tag{8}$$

where $\mathcal{L}_{\mathrm{CLS}}$ denotes the standard cross-entropy loss evaluated on the labeled node set $\mathcal{S}$, defined as $\mathcal{L}_{\mathrm{CLS}} = -\sum_{v\in\mathcal{S}} \sum_{j=1}^{c} y_{vj} \log \hat{y}_{vj}$, and $\lambda > 0$ controls the strength of the structure-aware regularization. Unlike existing graph regularization methods based on random edge perturbation (Bo et al., 2022) or generic smoothness constraints (Ando & Zhang, 2006; Yang et al., 2021), SER is directly motivated by our structural-complexity analysis. Through $\mathcal{L}_{\mathrm{SE}}^{(k)}$, SER imposes a differentiable penalty on class-inconsistent message passing, enabling selective structural regularization rather than uniformly suppressing the overall message-passing structure. The overall optimization procedure is summarized in Algorithm 1 (Appendix E).

5 Experiments

In this section, we apply structural regularization to representative GNN architectures and evaluate whether reducing message-passing structural complexity improves generalization.

Does Structural Complexity Explain GNN Generalization? To empirically examine the qualitative relationship suggested by our theoretical analysis in Section 3, we examine the relationship between structural complexity and the empirical generalization gap. We estimate the generalization gap by the absolute difference between the empirical errors on the test and training sets. As shown in Figure 1, the generalization gap exhibits an increasing trend as the number of effective propagation relations increases in most settings. These observations are consistent with our theoretical analysis, indicating that larger structural complexity is generally associated with a larger empirical generalization gap and may therefore indicate a greater risk of overfitting.

Is Overall Structural Complexity Control Sufficient? The analysis in Section 3 suggests that excessive structural complexity may be associated with a looser generalization guarantee. A natural strategy is therefore to directly penalize the number of effective message-passing relations. To evaluate its sufficiency, we introduce an L_0-based regularization baseline (L_0-reg) that directly penalizes the structural complexity. As reported in Table 1, we compare L_0-reg with both the vanilla models and SER across different GNN architectures. The results show that L_0-reg generally underperforms SER, suggesting that reducing overall structural complexity alone is insufficient and that selective regulation of class-inconsistent aggregation can be more effective.

Does SER Effectively Control Harmful Structural Complexity? To further validate the mechanism of SER, we analyze its impact on the learned message-passing structures from two perspectives: effective edge usage and cross-class aggregation. Specifically, effective edges are identified using a fixed threshold on the propagation weights, and their proportion is reported as the effective edge ratio. We further quantify class-inconsistent message passing by the fraction of effective

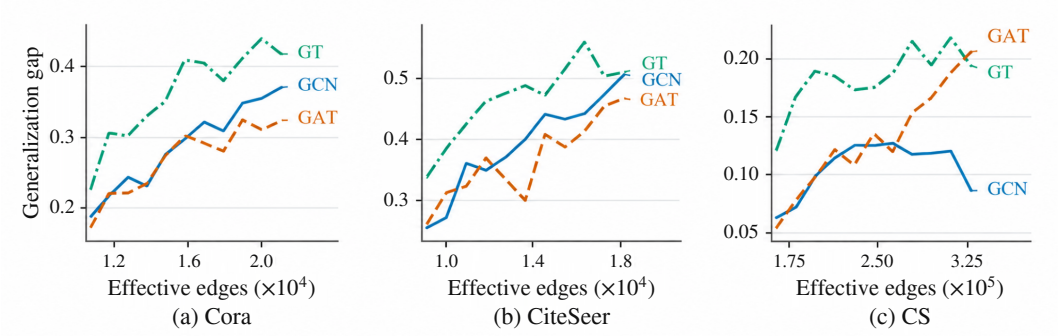


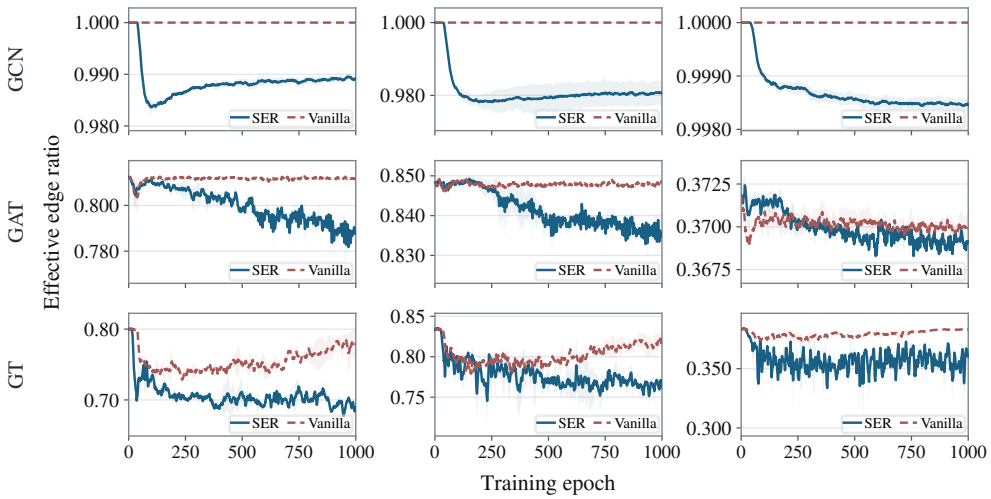
Figure 1: The relationship between the generalization gap and the number of effective propagation relations in GCN, GAT, and GT.

Table 1: Node classification accuracy (%) of different regularization strategies.

Model	Method	Cora	CiteSeer	PubMed	CS	Physics	Computers	Photo
GCN	Vanilla	81.54 ± 0.88	71.40 ± 0.69	79.18 ± 0.29	91.25 ± 0.71	93.25 ± 0.37	69.51 ± 3.40	83.24 ± 2.01
	L_0-reg	81.11 ± 0.35	71.47 ± 0.58	78.75 ± 0.22	91.73 ± 0.57	93.38 ± 0.59	71.38 ± 3.64	84.35 ± 1.68
	LapR	82.13 ± 0.49	71.70 ± 0.75	**79.62** ± 0.32	91.50 ± 0.61	93.89 ± 0.21	73.49 ± 1.39	84.48 ± 1.22
	CP	82.44 ± 2.97	74.53 ± 0.28	78.93 ± 0.42	92.26 ± 0.21	94.12 ± 0.79	72.36 ± 2.04	85.76 ± 2.01
	P-reg	**83.25** ± 0.56	74.56 ± 0.35	78.32 ± 0.37	92.35 ± 0.19	93.71 ± 1.26	72.22 ± 3.35	**87.20** ± 1.58
	NASA	81.82 ± 0.93	71.97 ± 0.27	79.23 ± 0.62	91.65 ± 0.47	93.41 ± 0.37	71.14 ± 2.56	84.44 ± 1.41
	R-reg	82.29 ± 0.56	73.87 ± 0.39	79.43 ± 0.36	92.36 ± 0.20	94.24 ± 0.49	69.10 ± 6.15	84.24 ± 2.82
	SER	82.85 ± 0.54	**74.79** ± 0.54	79.50 ± 0.29	**92.57** ± 0.29	**94.57** ± 0.30	**73.58** ± 2.57	85.11 ± 2.43
GAT	Vanilla	82.90 ± 0.66	71.88 ± 0.86	77.99 ± 0.55	91.22 ± 0.45	93.04 ± 0.51	73.85 ± 2.75	87.04 ± 2.03
	L_0-reg	83.01 ± 0.57	72.06 ± 1.03	78.22 ± 0.59	91.26 ± 0.35	93.13 ± 0.35	73.87 ± 2.26	86.98 ± 2.23
	LapR	82.81 ± 0.93	71.67 ± 0.82	**79.63** ± 0.37	91.54 ± 0.55	93.83 ± 0.23	73.59 ± 1.21	84.39 ± 1.11
	CP	83.30 ± 0.41	72.75 ± 0.81	78.28 ± 0.41	91.77 ± 0.18	93.89 ± 0.39	74.14 ± 2.67	88.62 ± 2.13
	P-reg	82.45 ± 0.80	72.60 ± 0.52	78.51 ± 0.68	91.78 ± 0.28	94.00 ± 0.34	74.41 ± 3.51	88.56 ± 2.27
	NASA	83.07 ± 0.64	71.87 ± 1.27	77.71 ± 0.69	90.80 ± 0.75	93.05 ± 0.79	72.84 ± 2.91	86.54 ± 2.06
	R-reg	83.17 ± 0.71	72.33 ± 0.74	78.60 ± 0.33	91.77 ± 0.22	94.02 ± 0.35	62.80 ± 3.51	62.04 ± 4.09
	SER	**83.44** ± 0.41	**73.59** ± 0.57	78.44 ± 0.63	**91.84** ± 0.22	**94.18** ± 0.34	**75.27** ± 2.84	**88.98** ± 2.22
GT	Vanilla	76.60 ± 1.55	65.09 ± 1.27	75.26 ± 1.32	86.79 ± 1.57	90.70 ± 1.53	77.12 ± 3.50	87.02 ± 2.10
	L_0-reg	73.22 ± 1.66	62.12 ± 1.86	75.39 ± 1.15	87.63 ± 1.87	91.05 ± 1.41	75.61 ± 3.39	87.66 ± 1.84
	LapR	78.69 ± 1.76	68.83 ± 1.89	**76.71** ± 0.80	87.06 ± 6.65	92.66 ± 0.49	78.03 ± 1.90	88.35 ± 1.34
	CP	77.43 ± 1.38	64.78 ± 3.84	75.74 ± 0.77	89.33 ± 1.03	91.44 ± 0.96	80.42 ± 2.40	88.46 ± 1.36
	P-reg	75.80 ± 3.22	62.76 ± 5.24	76.41 ± 1.54	89.72 ± 2.23	93.08 ± 0.77	**82.56** ± 1.88	88.81 ± 1.96
	NASA	78.85 ± 0.87	68.36 ± 1.31	76.48 ± 0.79	88.97 ± 0.75	91.18 ± 1.25	77.23 ± 4.17	86.85 ± 2.32
	R-reg	78.57 ± 1.49	69.05 ± 2.03	76.60 ± 1.39	87.43 ± 6.27	**93.17** ± 0.67	54.69 ± 14.88	70.58 ± 4.80
	SER	**78.96** ± 0.95	**69.58** ± 1.60	76.68 ± 0.86	**90.17** ± 1.53	93.02 ± 0.59	81.19 ± 2.22	**89.13** ± 2.03

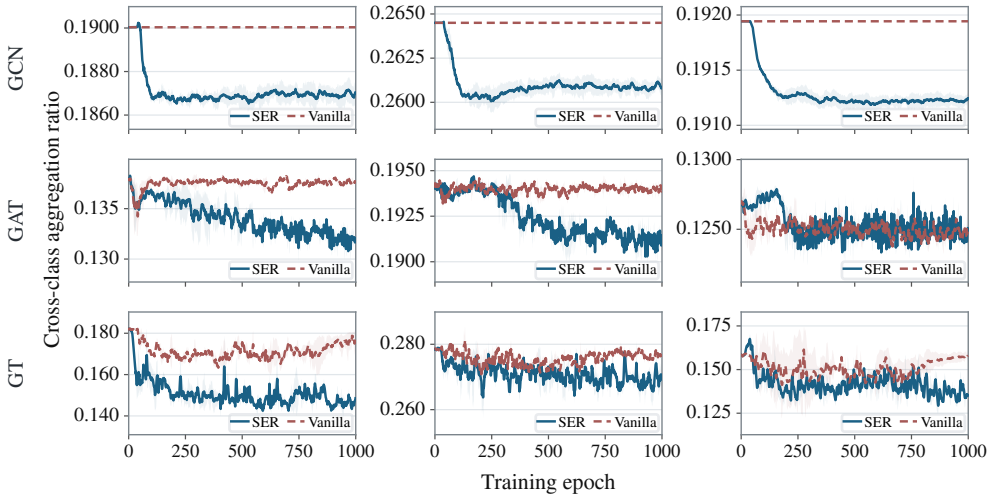
edges connecting nodes from different ground-truth labels. The results are illustrated in Figure 2 and Figure 3, respectively. We observe that incorporating SER consistently reduces the number of effective edges compared with the vanilla models across all architectures, encouraging sparser and more selective message passing. Moreover, SER substantially suppresses cross-class aggregation, which directly corresponds to the class-inconsistent structural component targeted by the surrogate in Eq. (6). This effect becomes more evident in later training stages, where the learned aggregation structure becomes increasingly consistent with the underlying class partition. These results show that SER effectively regulates the potentially harmful structural complexity targeted by our analysis.

Does Controlling Harmful Structural Complexity Improve Generalization? We next compare SER with a broader range of existing regularization methods. Table 1 reports the node classification accuracy of all models under different regularization strategies, where the best results are highlighted in **bold** and the second-best results are underlined. The proposed SER consistently outperforms both the vanilla baselines and competing regularization methods across most datasets and model architectures. Further analyses of existing regularization strategies and the behavior of SER across different



(a) Cora (b) CiteSeer (c) CS

Figure 2: Evolution of the effective edge ratio over training epochs for GCN, GAT, and GT. The red curves denote the vanilla models, while the blue curves represent models with SER.



(a) Cora (b) CiteSeer (c) CS

Figure 3: Evolution of the cross-class aggregation ratio over training epochs for GCN, GAT, and GT. The red curves denote the vanilla models, while the blue curves represent models with SER.

architectures are provided in Appendix G. These results demonstrate the effectiveness of SER in improving generalization through selective control of potentially harmful structural complexity.

Ablation Study. To further assess the contribution of each SER component, we conduct an ablation study, with node classification accuracy (%) reported in Table 2. We consider three variants: w/o SER, which removes the entire structural entropy regularization; SER w/o CAS (class-aware structure), which removes class-aware modeling of message passing; and SER w/o EEC (effective-edge control), which disables the mechanism for regulating message-passing structures. Removing SER consistently degrades performance across different architectures, demonstrating the effectiveness of explicitly controlling structural complexity. Moreover, the performance drop of SER w/o CAS indicates that class-aware structural modeling is important for suppressing cross-class aggregation and improving generalization. The degradation of SER w/o EEC further verifies that the benefit of SER comes from regulating effective message-passing edges rather than merely introducing an additional

Table 2: Ablation study of SER on different GNN architectures.

Model	Method	Cora	CiteSeer	PubMed	CS	Physics	Computers	Photo
GCN	w/o SER	81.54 ± 0.88	71.40 ± 0.69	79.18 ± 0.29	91.25 ± 0.71	93.25 ± 0.37	69.51 ± 3.40	83.24 ± 2.01
	w/o CAS	80.57 ± 0.46	71.09 ± 0.58	78.72 ± 0.47	91.65 ± 0.68	93.42 ± 0.43	71.03 ± 0.36	84.15 ± 1.73
	w/o EEC	82.77 ± 0.71	74.48 ± 0.60	79.37 ± 0.45	92.36 ± 0.20	94.08 ± 0.44	72.38 ± 2.69	84.29 ± 2.72
	SER	**82.85** ± 0.54	**74.79** ± 0.54	**79.50** ± 0.29	**92.57** ± 0.29	**94.57** ± 0.30	**73.58** ± 2.57	**85.11** ± 2.43
GAT	w/o SER	82.90 ± 0.66	71.88 ± 0.86	77.99 ± 0.55	91.22 ± 0.45	93.04 ± 0.51	73.85 ± 2.75	87.04 ± 2.03
	w/o CAS	82.64 ± 0.61	71.60 ± 0.74	78.06 ± 0.41	91.22 ± 0.37	93.08 ± 0.54	73.70 ± 2.48	86.80 ± 2.16
	w/o EEC	83.35 ± 0.58	73.36 ± 0.48	78.31 ± 0.59	91.81 ± 0.29	94.09 ± 0.41	74.49 ± 1.86	88.21 ± 1.61
	SER	**83.44** ± 0.41	**73.59** ± 0.57	**78.44** ± 0.63	**91.84** ± 0.22	**94.18** ± 0.34	**75.27** ± 2.84	**88.98** ± 2.22
GT	w/o SER	76.60 ± 1.55	65.09 ± 1.27	75.26 ± 1.32	86.79 ± 1.57	90.70 ± 1.53	77.12 ± 3.50	87.02 ± 2.10
	w/o CAS	72.49 ± 1.98	62.52 ± 1.62	75.25 ± 0.78	87.82 ± 1.49	90.76 ± 1.48	76.45 ± 3.92	87.36 ± 2.51
	w/o EEC	78.18 ± 0.49	69.26 ± 1.78	76.27 ± 0.89	89.51 ± 0.80	91.00 ± 1.67	80.48 ± 3.46	88.19 ± 2.41
	SER	**78.96** ± 0.95	**69.58** ± 1.60	**76.68** ± 0.86	**90.17** ± 1.53	**93.02** ± 0.59	**81.19** ± 2.22	**89.13** ± 2.03

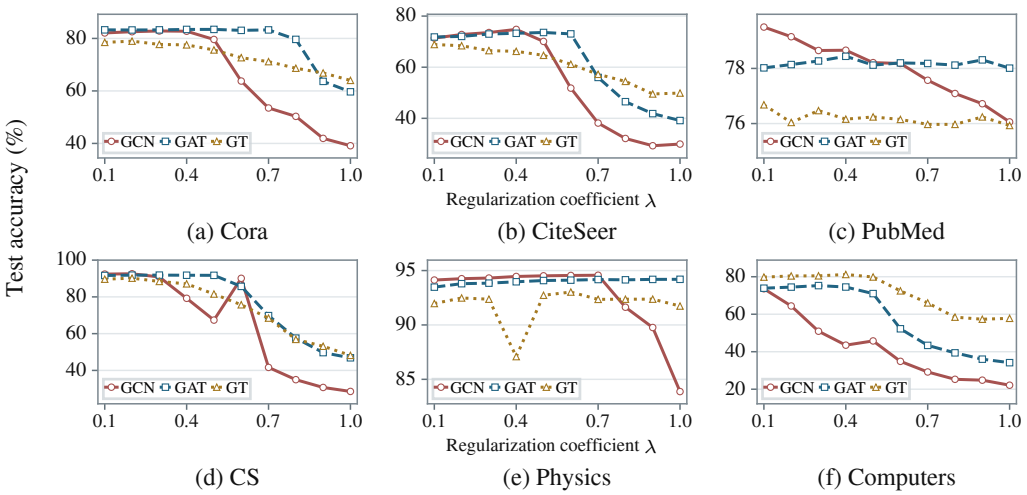


Figure 4: The relationship between test accuracy and the regularization strength λ across six benchmark datasets.

entropy-based penalty. Overall, the ablation results confirm that each component of SER contributes to improving the generalization performance of GNNs.

Sensitivity to Regularization Strength. To further investigate the impact of the regularization strength, we analyze the sensitivity of performance with respect to λ, with results shown in Figure 4. We observe that moderate values of λ consistently improve performance, while accuracy begins to decline once λ exceeds a moderate range, typically around 0.6 to 0.7 in most datasets. This behavior is consistent with the motivation of SER: moderate regularization can selectively suppress potentially harmful aggregation relations, whereas overly strong regularization may also constrain useful structural relations and reduce model capacity, leading to underfitting. Therefore, an appropriate value of λ provides a practical trade-off between overfitting and underfitting.

6 Conclusion

In this paper, we define message-passing structural complexity through effective aggregation relations and theoretically establish its fundamental role in GNN generalization. We further develop SER to penalize a continuous surrogate whose scaled form pointwise upper-bounds the cross-class component of structural complexity. Experiments across representative GNN architectures and benchmark datasets show that reducing structural complexity can improve GNN generalization.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing, primarily to improve the readability, grammar, and clarity of the manuscript. We have reviewed all AI-assisted work and verified the resulting text for accuracy and consistency with our intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rie Ando and Tong Zhang. Learning on graph with laplacian regularization. *Advances in Neural Information Processing System*, 19, 2006.

Deyu Bo, Binbin Hu, Xiao Wang, Zhiqiang Zhang, Chuan Shi, and Jun Zhou. Regularizing graph neural networks via consistency-diversity graph augmentations. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, volume 36, pp. 3913–3921, 2022.

Kirill Brilliantov, Amauri H Souza, and Vikas Garg. Compositional pac-bayes: Generalization of gnns with persistence and beyond. *Advances in Neural Information Processing Systems*, 37: 107890–107931, 2024.

Yoonhyuk Choi, Jiho Choi, Taewook Ko, JongWook Kim, and Chong-Kwon Kim. Sheaf graph neural networks via pac-bayes spectral optimization. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, volume 40, pp. 20570–20578, 2026.

Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.

Pascal Esser, Leena Chennuru Vankadara, and Debarghya Ghoshdastidar. Learning theory can (sometimes) explain generalisation in graph neural networks. *Advances in Neural Information Processing Systems*, 34:27043–27056, 2021.

Wei Fan, Jingru Fei, Dingyu Guo, Kun Yi, Xiaozhuang Song, Haolong Xiang, Hangting Ye, and Min Li. Towards multi-resolution spatiotemporal graph learning for medical time series classification. In *Proceedings of the ACM on Web Conference 2025*, pp. 5054–5064, 2025.

Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *Proceedings of the Thirty-Seventh International Conference on Machine Learning*, pp. 3419–3430, 2020.

Arthur Jacot, Seok Hoan Choi, and Yuxiao Wen. How dnns break the curse of dimensionality: Compositionality and symmetry learning. In *Proceedings of the Thirteenth International Conference on Learning Representations*, pp. 74504–74538, 2025.

Feng Ji, Yanan Zhao, See Hian Lee, Kai Zhao, Wee Peng Tay, and Jielong Yang. Graph distributional signals for regularization in graph neural networks. *IEEE Transactions on Signal and Information Processing over Networks*, 11:670–682, 2025.

Haotian Ju, Dongyue Li, Aneesh Sharma, and Hongyang R Zhang. Generalization in graph neural networks: Improved pac-bayesian bounds on graph diffusion. In *Proceedings of the Twenty-Sixth International Conference on Artificial Intelligence and Statistics*, pp. 6314–6341, 2023.

Wei Ju, Zhengyang Mao, Siyu Yi, Yifang Qin, Yiyang Gu, Zhiping Xiao, Jianhao Shen, Ziyue Qiao, and Ming Zhang. Cluster-guided contrastive class-imbalanced graph classification. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, pp. 11924–11932, 2025.

Antoine Ledent, Waleed Mustafa, Yunwen Lei, and Marius Kloft. Norm-based generalisation bounds for deep multi-class convolutional neural networks. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, volume 35, pp. 8279–8287, 2021.

Jia Li, Jianwei Yu, Jiajin Li, Honglei Zhang, Kangfei Zhao, Yu Rong, Hong Cheng, and Junzhou Huang. Dirichlet graph variational autoencoder. *Advances in Neural Information Processing Systems*, 33:5274–5283, 2020.

Mang Li and Wei Lyu. Adaptive regularization for large-scale sparse feature embedding models. In *Proceedings of the Fourteenth International Conference on Learning Representations*, pp. 44098–44118, 2026.

Shikun Liu, Tianchun Li, Yongbin Feng, Nhan Tran, Han Zhao, Qiang Qiu, and Pan Li. Structural re-weighting improves graph domain adaptation. In *Proceedings of the Fortieth International Conference on Machine Learning*, pp. 21778–21793, 2023.

Shaogao Lv. Generalization bounds for graph convolutional neural networks via rademacher complexity. *arXiv preprint arXiv:2102.10234*, 2021.

Jiaqi Ma, Junwei Deng, and Qiaozhu Mei. Subgroup generalization and fairness of graph neural networks. *Advances in Neural Information Processing Systems*, 34:1048–1061, 2021.

Sohir Maskey, Raffaele Paolino, Aras Bacho, and Gitta Kutyniok. A fractional graph laplacian approach to oversmoothing. *Advances in Neural Information Processing Systems*, 36:13022–13063, 2023.

Hoang Nt and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*, 2019.

Kenta Oono and Taiji Suzuki. Optimization and generalization analysis of transduction through gradient boosting and application to multi-scale graph neural networks. *Advances in Neural Information Processing Systems*, 33:18917–18930, 2020.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *Proceedings of the Fifth International Conference on Learning Representations*, 2017.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–93, 2008.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

Huayi Tang and Yong Liu. Towards understanding generalization of graph neural networks. In *Proceedings of the Fortieth International Conference on Machine Learning*, pp. 33674–33719, 2023.

Saurabh Verma and Zhi-Li Zhang. Stability and generalization of graph convolutional neural networks. In *Proceedings of the Twenty-Fifth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1539–1548, 2019.

Qixun Wang, Yifei Wang, Yisen Wang, and Xianghua Ying. Dissecting the failure of invariant learning on graphs. *Advances in Neural Information Processing Systems*, 37:80383–80438, 2024.

Zhiyang Wang, Juan Cerviño, and Alejandro Ribeiro. Generalization of graph neural networks is robust to model mismatch. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, pp. 21402–21410, 2025.

Lei Xu, Lei Chen, Rong Wang, Feiping Nie, and Xuelong Li. Joint feature and differentiable k-nn graph learning using dirichlet energy. *Advances in Neural Information Processing Systems*, 36: 24497–24518, 2023.

Guangrui Yang, Ming Li, Han Feng, and Xiaosheng Zhuang. Deeper insights into deep graph convolutional networks: Stability and generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1707–1719, 2025.

Han Yang, Kaili Ma, and James Cheng. Rethinking graph regularization for graph neural networks. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, volume 35, pp. 4573–4581, 2021.

Jun Yin, Senzhang Wang, Ziluowen Luo, Peng Huo, Hao Yan, Hao Miao, Chaozhuo Li, Shirui Pan, and Chengqi Zhang. Paradigm shift of gnn explainer from label space to prototypical representation space. In *Proceedings of the Fourteenth International Conference on Learning Representations*, pp. 13710–13734, 2026.

Haonan Yuan, Qingyun Sun, Junhua Shi, Xingcheng Fu, Bryan Hooi, Jianxin Li, and Philip S Yu. How much can transfer? bridge: Bounded multi-domain graph foundation model with generalization guarantees. In *Proceedings of the Forty-Second International Conference on Machine Learning*, 2025.

A Graph Neural Network Architectures

A representative instantiation of GNNs is GCN, in which neighborhood aggregation is realized through a fixed, normalized graph operator derived from the adjacency structure. We define $\widetilde{A} = A + I_n$ and $\widetilde{D} = D + I_n$ as the adjacency and degree matrices of the graph augmented with self-loops. In matrix form, the update of a GCN at layer k can be written as:

$$H^{(k)} = \sigma\left(\hat{A}H^{(k-1)}W^{(k)}\right), \tag{9}$$

where $\hat{A} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$ is the normalized adjacency matrix with self-loops, $W^{(k)}$ is the trainable weight matrix at layer k, and $\sigma(\cdot)$ denotes a nonlinear activation function. GATs extend this framework by introducing learnable attention weights:

$$h_v^{(k)} = \sigma\left(\sum_{u\in\mathcal{N}(v)} \alpha_{vu}^{(k)}\left(W^{(k)}\right)^{\top} h_u^{(k-1)}\right), \tag{10}$$

where $\alpha_{vu}^{(k)}$ are normalized attention coefficients. More recently, GTs generalize attention-based message passing by extending aggregation from local neighborhoods to a potentially global set of nodes. At each layer, node representations are updated through multi-head self-attention, which can be written as:

$$\hat{h}_v^{(k)} = O^{(k)} \overset{\Delta}{\underset{r=1}{\Big\|}} \left(\sum_{u\in V} \alpha_{vu}^{(k,r)} V^{(k,r)} h_u^{(k-1)}\right), \tag{11}$$

where $\alpha_{vu}^{(k,r)}$ denotes attention weights, $\Big\|$ denotes concatenation across heads, $V^{(k,r)} \in \mathbb{R}^{d_r\times\widetilde{d}}$ is a trainable projection matrix, and $O^{(k)} \in \mathbb{R}^{\widetilde{d}\times\widetilde{d}}$ is a learnable output projection. Following the attention-based aggregation, residual connections, normalization and a position-wise feed-forward network are applied:

$$\widetilde{h}_v^{(k)} = \mathrm{Norm}\left(h_v^{(k-1)} + \hat{h}_v^{(k)}\right), \tag{12}$$

$$\hat{\widetilde{h}}_v^{(k)} = W_2^{(k)}\sigma\left(W_1^{(k)}\widetilde{h}_v^{(k)}\right), \tag{13}$$

$$h_v^{(k)} = \mathrm{Norm}\left(\widetilde{h}_v^{(k)} + \hat{\widetilde{h}}_v^{(k)}\right), \tag{14}$$

where $W_1^{(k)} \in \mathbb{R}^{2\widetilde{d}\times\widetilde{d}}$ and $W_2^{(k)} \in \mathbb{R}^{\widetilde{d}\times 2\widetilde{d}}$ are trainable parameters, and $\mathrm{Norm}(\cdot)$ denotes either Layer Normalization or Batch Normalization. For the theoretical analysis, we obtain a scalar-valued node prediction from the final GT representation through a linear readout:

$$f(x_v) = w_{\mathrm{out}}^{\top} h_v^{(K)}, \tag{15}$$

where $w_{\mathrm{out}} \in \mathbb{R}^{\widetilde{d}}$ denotes the output weight vector.

Table 3: Notations used in this paper.

Notation	Description
$G = (V, E)$	input graph with node set V and edge set E
$X \in \mathbb{R}^{n \times d}$	node feature matrix
$A \in \{0,1\}^{n \times n}$	adjacency matrix
$D \in \mathbb{R}^{n \times n}$	degree matrix
$\widetilde{A}, \widetilde{D}$	adjacency and degree matrices with self-loops
$\widetilde{L}$	augmented normalized Laplacian matrix
$Y = \{y_v\}_{v \in V}$	node-label collection
$\mathcal{S}$,$\mathcal{U}$	labeled and unlabeled node sets
m	number of labeled nodes, $\lvert\mathcal{S}\rvert$
$\mathcal{F}$	hypothesis class induced by the GNN
$\ell \circ \mathcal{F}$	loss-composed hypothesis class
$H^{(k)}$	node representations at layer k
$E(H^{(k)})$	Dirichlet energy at layer k
$\hat{A}$	normalized aggregation matrix
$A^{(k)}$	attention aggregation matrix at layer k
$\bar{A}_f^{(k)}$	thresholded attention matrix at layer k
η	structural complexity under fixed aggregation
$\eta_\tau^{(k)}$	effective structural complexity at layer k
τ	support threshold for continuously weighted propagation
Δ	number of attention heads
$\hat{\epsilon}(f)$	empirical risk over labeled nodes
$\epsilon(f)$	transductive test risk
$\mathfrak{R}_{m,n}(\mathcal{F})$	transductive Rademacher complexity of $\mathcal{F}$
λ	regularization coefficient

B Notation Summary

For clarity and ease of reference, we summarize the main notations used throughout the paper in Table 3.

C Supplementary Analysis of Structural Overfitting in GNNs

The analysis in Section 3 characterizes structural complexity through the effective aggregation relations exploited during message passing. In this appendix, we provide a complementary perspective on how the participation of graph structure in message passing may contribute to structural overfitting. We show that normalized GCN message passing does not increase the Dirichlet energy of pre-activation representations, revealing a structure-induced smoothing effect. When the target signal is aligned with the graph structure, such smoothing may facilitate fitting graph-consistent patterns. However, the same structural aggregation that induces smoothing may also propagate graph-specific or task-irrelevant correlations as more relations participate in message passing. Consequently, extensive structural participation may increase the dependence of learned representations on such correlations, providing a possible mechanism for structural overfitting.

C.1 Smoothness as a Measure of Fitting Capacity

The ability of a GNN to fit the data is influenced by the structural priors imposed by graph edges. Edges encourage node representations to vary smoothly across the graph, and smoother representations correspond to target functions that are easier to capture (Nt & Maehara, 2019). To quantify this effect, we first need a measure for smoothness on graphs. To this end, we employ the Dirichlet energy (Li et al., 2020; Maskey et al., 2023; Xu et al., 2023), defined as follows.

Definition 1 (Dirichlet energy)**.** *Let $G = (V, E)$ be an undirected graph with augmented normalized Laplacian $\widetilde{L} = I_n - \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$, and let $H^{(k)} = [h_1^{(k)}, \ldots, h_n^{(k)}]^\top \in \mathbb{R}^{n \times d}$ denote the node*

representation matrix learned by a graph neural network at the k-th layer. The Dirichlet energy of $H^{(k)}$ with respect to graph G is defined as:

$$E(H^{(k)}) = \mathrm{tr}\left(H^{(k)\top}\widetilde{L}H^{(k)}\right) = \frac{1}{2}\sum_{v\in[n]}\sum_{u\in\mathcal{N}(v)}\widetilde{a}_{vu}\left\|\frac{h_v^{(k)}}{\sqrt{1+d_v}} - \frac{h_u^{(k)}}{\sqrt{1+d_u}}\right\|_2^2, \tag{16}$$

where $\widetilde{a}_{vu}$ denotes the (v,u)-th entry of the augmented adjacency matrix, and d_v is the degree of node v in the original graph.

The definition above shows that Dirichlet energy aggregates representation discrepancies across edges under the normalization induced by $\widetilde{L}$. As a quadratic form associated with the graph Laplacian, Dirichlet energy admits an equivalent characterization in the spectral domain, as stated in the following Lemma.

Lemma 2. *Let $\widetilde{L}$ be the augmented normalized Laplacian of graph G, with eigenpairs $\{(\lambda_i, u_i)\}_{i=1}^n$, where $0 = \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$. For any graph signal $f \in \mathbb{R}^n$ with expansion:*

$$f = \sum_{i=1}^n \gamma_i u_i, \tag{17}$$

the Dirichlet energy of f is given by:

$$E(f) = f^\top \widetilde{L} f = \sum_{i=1}^n \lambda_i \gamma_i^2. \tag{18}$$

Proof. Since $\widetilde{L}$ is symmetric and positive semi-definite, its eigenvectors $\{u_i\}_{i=1}^n$ form an orthonormal basis of $\mathbb{R}^n$. Substituting the spectral expansion of f into the quadratic form yields:

$$f^\top \widetilde{L} f = \left(\sum_{i=1}^n \gamma_i u_i\right)^\top \widetilde{L}\left(\sum_{j=1}^n \gamma_j u_j\right) = \sum_{i=1}^n \lambda_i \gamma_i^2,$$

where we use the orthonormality of eigenvectors and the eigenvalue equation $\widetilde{L}u_i = \lambda_i u_i$. □

Remark 1. *Lemma 2 shows that Dirichlet energy quantifies the spectral complexity of a graph signal by weighting its Laplacian eigencomponents according to their frequencies. Since larger eigenvalues correspond to rapidly varying modes across edges, a large Dirichlet energy indicates strong local variation on the graph, whereas a small value implies smoothness with respect to the graph structure.*

This property makes Dirichlet energy a natural tool for analyzing representation smoothness in GNNs. Since many message-passing GNNs exhibit a low-pass bias, they preferentially capture low-frequency components of graph signals. When the target signal is aligned with the graph structure and is dominated by low-frequency components, smoother representations with lower Dirichlet energy can facilitate fitting the target patterns. In the next subsection, we analyze how incorporating graph structure into GNNs affects the Dirichlet energy.

C.2 Impact of Graph Structure on the Fitting Capacity of GNNs

To analyze how graph structure affects fitting capacity, we compare a standard GNN against its edge-free extreme, where the model degrades to a multi-layer perceptron (MLP) that updates node representations using only their own features. We take GCNs as a representative example and focus on the pre-activation representations to isolate the effect of graph propagation from that of the nonlinear activation. The difference in Dirichlet energy between these two scenarios is characterized by the following theorem.

Theorem 3. *Let $Z_{\mathrm{GCN}}^{(1)} = \hat{A}XW^{(1)}$ and $Z_{\mathrm{MLP}}^{(1)} = XW^{(1)}$ denote the pre-activation representations of a one-layer GCN with and without graph propagation, respectively. Then the Dirichlet energy satisfies:*

$$E(Z_{\mathrm{GCN}}^{(1)}) \le \left\|\hat{A}\right\|_2^2 E(Z_{\mathrm{MLP}}^{(1)}). \tag{19}$$

Proof. According to the definition of Dirichlet energy in Eq.(16), the Dirichlet energy of the pre-activation representation of a one-layer GCN is:

$$E(Z_{\mathrm{GCN}}^{(1)}) = \mathrm{tr}\left(Z_{\mathrm{GCN}}^{(1)\top}\widetilde{L}Z_{\mathrm{GCN}}^{(1)}\right). \tag{20}$$

Similarly, the Dirichlet energy of the corresponding pre-activation representation without graph propagation is:

$$E(Z_{\mathrm{MLP}}^{(1)}) = \mathrm{tr}\left(Z_{\mathrm{MLP}}^{(1)\top}\widetilde{L}Z_{\mathrm{MLP}}^{(1)}\right). \tag{21}$$

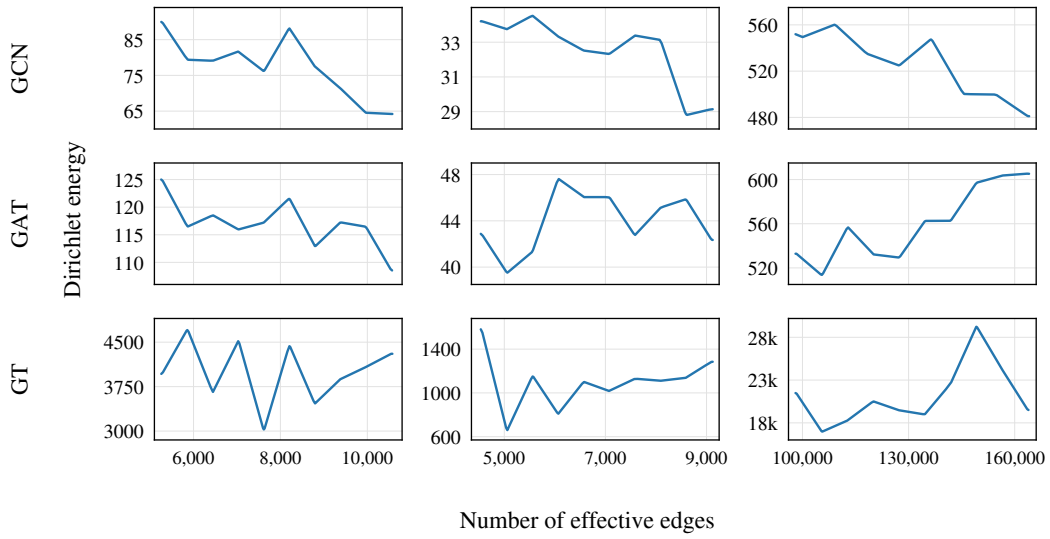
Substituting the pre-activation GCN representation into Eq.(20), we have:

$$\begin{aligned}
&\mathrm{tr}\left(Z_{\mathrm{GCN}}^{(1)\top}\widetilde{L}Z_{\mathrm{GCN}}^{(1)}\right)\\
=&\,\mathrm{tr}\left(\left(\hat{A}XW^{(1)}\right)^{\top}\widetilde{L}\left(\hat{A}XW^{(1)}\right)\right)\\
=&\,\mathrm{tr}\left(\left(W^{(1)}\right)^{\top}X^{\top}\hat{A}^{\top}\widetilde{L}\hat{A}XW^{(1)}\right)\\
\leq&\,\mathrm{tr}\left(\left(W^{(1)}\right)^{\top}X^{\top}\widetilde{L}XW^{(1)}\right)\mu_{\max}\left(\hat{A}^{\top}\hat{A}\right)\\
=&\left\|\hat{A}\right\|_2^2\,\mathrm{tr}\left(\left(XW^{(1)}\right)^{\top}\widetilde{L}\left(XW^{(1)}\right)\right)\\
=&\left\|\hat{A}\right\|_2^2\,\mathrm{tr}\left(Z_{\mathrm{MLP}}^{(1)\top}\widetilde{L}Z_{\mathrm{MLP}}^{(1)}\right)\\
=&\left\|\hat{A}\right\|_2^2 E(Z_{\mathrm{MLP}}^{(1)}),
\end{aligned} \tag{22}$$

where the equality in the second step follows from the cyclic property of the trace operator together with the standard transpose rules for matrix products. The inequality in the third step follows from the fact that for any symmetric positive semidefinite matrix M and any matrix B, it holds that $\mathrm{tr}(B^{\top}MB) \leq \mu_{\max}(M)\,\mathrm{tr}(B^{\top}B)$, where $\mu_{\max}(\cdot)$ denotes the largest eigenvalue. Here, since $\widetilde{L} = I_n - \hat{A}$, $\widetilde{L}$ and $\hat{A}$ commute, allowing us to rewrite $\hat{A}^{\top}\widetilde{L}\hat{A} = \hat{A}^{\top}\hat{A}\widetilde{L}$. The trace inequality can then be applied with $M = \hat{A}^{\top}\hat{A}$ and $B = \widetilde{L}^{1/2}XW^{(1)}$. The fourth step is due to the definition of the spectral norm $\mu_{\max}(M^{\top}M) = \|M\|_2^2$ for any matrix M. □

Remark 2. *Theorem 3 provides insight into the role of the graph structure in controlling smoothness. For GCNs, the spectral norm of the normalized adjacency matrix satisfies $\|\hat{A}\|_2^2 \equiv 1$, and hence $E(Z_{\mathrm{GCN}}^{(1)}) \leq E(Z_{\mathrm{MLP}}^{(1)})$. Thus, the GCN message-passing operation does not increase the Dirichlet energy of the pre-activation representations, revealing a structure-induced smoothing effect. Similar energy-contraction results can be established for attention-based architectures under additional conditions on the attention operator. In particular, if an attention matrix $A^{(1)}$ is symmetric and commutes with the graph Laplacian $\widetilde{L}$, then $E(A^{(1)}Z) \leq \|A^{(1)}\|_2^2 E(Z)$. Hence, when $\|A^{(1)}\|_2 \leq 1$, the corresponding attention-based message-passing operation does not increase the Dirichlet energy. Moreover, the spectral norm of the propagation operator appears as a multiplicative factor controlling the upper bound on the Dirichlet energy after message passing. Under the conditions above, a smaller spectral norm yields a stronger guaranteed energy contraction.*

Remark 3. *To further validate our theoretical findings, we measure the Dirichlet energy of node representations under varying numbers of edges participating in the decision process. Specifically, we gradually increase the number of effective edges involved in aggregation and record the resulting Dirichlet energy for three representative GNN architectures: GCN, GAT, and GT. The results are presented in Figure 5. The first row corresponds to GCN across three datasets, the second row to GAT, and the third row to GT. Across most settings, we observe a decreasing trend in Dirichlet energy as the number of decision-participating edges increases. These empirical observations corroborate our theoretical analysis, suggesting that increased edge-based aggregation is generally associated with lower representation energy and a stronger smoothing effect.*



(a) Cora (b) CiteSeer (c) CS

Figure 5: The relationship between Dirichlet energy and the number of effective edges in GCN, GAT and GT.

Taken together, these results provide a smoothness-based perspective on how structural information may contribute to overfitting in message-passing GNNs. Edge-based aggregation generally promotes smoother representations by reducing local variation across the graph. When the target signal is aligned with the graph structure, such smoothing can facilitate fitting graph-consistent patterns. However, as more structural relations participate in message passing, aggregation may also propagate graph-specific or task-irrelevant correlations, thereby increasing the dependence of learned representations on the observed graph. When some of these relations are noisy or class-inconsistent, such dependence may increase the risk of structural overfitting. This smoothness-based view provides complementary intuition for the structural-overfitting phenomenon discussed in the main text.

D Theoretical Preliminaries and Proofs

D.1 Transductive Rademacher Complexity

We provide the formal definition of transductive Rademacher complexity used in our analysis.

Definition 2 (Transductive Rademacher Complexity)**.** *Let $\mathcal{F}$ be a class of node-level predictors on a fixed graph $G = (V, E)$ with $|V| = n$, where m nodes are labeled and the remaining $n - m$ nodes are unlabeled. The transductive Rademacher complexity of $\mathcal{F}$ is defined as:*

$$\mathfrak{R}_{m,n}(\mathcal{F}) = \left(\frac{1}{m} + \frac{1}{n-m}\right) \mathbb{E}_\zeta \left[\sup_{f \in \mathcal{F}} \sum_{j=1}^{n} \zeta_j f(x_j)\right], \tag{23}$$

where $\zeta_1, \ldots, \zeta_n$ are independent random variables taking values $+1$ and -1 with probability p_0 each, and 0 with probability $1 - 2p_0$, where $p_0 = \frac{m(n-m)}{n^2}$.

The expectation over ζ in Definition 2 is auxiliary to the definition of transductive Rademacher complexity and is distinct from the probability over the random training-set draw $\mathcal{S}$ in Lemma 1.

D.2 Proofs of Main Theoretical Results

Theorem 1. *Consider the hypothesis class $\mathcal{F}$ induced by GNNs with fixed propagation operators. Assume that $\|x_i\|_2 \leq B$ for all nodes i, the activation function σ is 1-Lipschitz with $\sigma(0) = 0$, the propagation weights satisfy $|\omega_{vu}^{(k)}| \leq 1$ for $k = 1, 2$, the weight matrix in the hidden layer satisfies*

$\left\|W^{(1)}\right\|_2 \leq R_1$, *the output weight vector satisfies* $\left\|W^{(2)}\right\|_2 \leq R_2$, *and the loss function* $\ell(\cdot, y)$ *is* L*-Lipschitz continuous. Then for* $0 < \delta < 1$, *with probability at least* $1-\delta$ *for all* $f \in \mathcal{F}$ *the following holds:*

$$\begin{aligned}\epsilon(f) \leq & \hat{\epsilon}(f) + LR_1R_2B\sqrt{\frac{2n}{m(n-m)}}\sqrt{\eta^{(1)}\eta^{(2)}} \\ & + c_1\frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2\sqrt{\frac{n}{m(n-m)}\log\frac{1}{\delta}},\end{aligned}$$

where $\eta^{(k)} = \left\|\Omega^{(k)}\right\|_0$ *for* $k = 1, 2$ *denotes the structural complexity of each layer, and* $c_1 < 5.05$ *and* $c_2 < 0.8$ *are universal constants.*

Proof. Let $Q_{m,n} = \frac{1}{m} + \frac{1}{n-m}$. Substituting the two-layer message-passing predictor defined in Eq.(1) into the definition of the transductive Rademacher complexity in Eq.(23), we obtain:

$$\begin{aligned}\mathfrak{R}_{m,n}(\mathcal{F}) &= Q_{m,n}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}\zeta_j f(x_j)\right] \\ &= Q_{m,n}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}\zeta_j\sigma\left(\sum_{v\in\mathcal{N}(j)}[\Omega^{(2)}]_{jv}\left(W^{(2)}\right)^\top h_v^{(1)}\right)\right] \\ &\leq Q_{m,n}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}\zeta_j\sum_{v\in\mathcal{N}(j)}[\Omega^{(2)}]_{jv}\left(W^{(2)}\right)^\top h_v^{(1)}\right] \\ &\leq Q_{m,n}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}}\left\|\sum_{j=1}^{n}\zeta_j\sum_{v\in\mathcal{N}(j)}[\Omega^{(2)}]_{jv}h_v^{(1)}\right\|_2\|W^{(2)}\|_2\right] \\ &\leq Q_{m,n}R_2\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}}\left\|\sum_{j=1}^{n}\zeta_j\sum_{v\in\mathcal{N}(j)}[\Omega^{(2)}]_{jv}h_v^{(1)}\right\|_2\right], \qquad (24)\end{aligned}$$

where the first inequality follows from the contraction property of the 1-Lipschitz activation function σ with $\sigma(0) = 0$, the second inequality follows by expressing the output as an inner product between the aggregated hidden representation and the output weight vector and applying the Cauchy-Schwarz inequality, and the last inequality follows from the norm constraint $\|W^{(2)}\|_2 \leq R_2$. For the first-layer representation, Eq. (1) gives:

$$h_v^{(1)} = \sigma\left(\sum_{i\in\mathcal{N}(v)}[\Omega^{(1)}]_{vi}\left(W^{(1)}\right)^\top x_i\right).$$

Since σ is 1-Lipschitz and satisfies $\sigma(0) = 0$, and $\|W^{(1)}\|_2 \leq R_1$, we have:

$$\|h_v^{(1)}\|_2 \leq \left\|\sum_{i\in\mathcal{N}(v)}[\Omega^{(1)}]_{vi}\left(W^{(1)}\right)^\top x_i\right\|_2 \leq R_1\left\|\sum_{i\in\mathcal{N}(v)}[\Omega^{(1)}]_{vi}x_i\right\|_2.$$

By exchanging the order of summation over j and v and applying the triangle inequality, we obtain:

$$\begin{aligned}&\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}}\left\|\sum_{j=1}^{n}\zeta_j\sum_{v\in\mathcal{N}(j)}[\Omega^{(2)}]_{jv}h_v^{(1)}\right\|_2\right] \\ =&\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}}\left\|\sum_{v\in V}\left(\sum_{j=1}^{n}\zeta_j[\Omega^{(2)}]_{jv}\right)h_v^{(1)}\right\|_2\right]\end{aligned}$$

$$\leq \mathbb{E}_\zeta \left[\sup_{f\in\mathcal{F}} \sum_{v\in V} \left| \sum_{j=1}^{n} \zeta_j [\Omega^{(2)}]_{jv} \right| \|h_v^{(1)}\|_2 \right]$$
$$\leq R_1 \sum_{v\in V} \mathbb{E}_\zeta \left[\left| \sum_{j=1}^{n} \zeta_j [\Omega^{(2)}]_{jv} \right| \right] \left\| \sum_{i\in\mathcal{N}(v)} [\Omega^{(1)}]_{vi} x_i \right\|_2 . \tag{25}$$

Since the transductive Rademacher variables are independent and satisfy $\mathbb{E}[\zeta_j] = 0$ and $\mathbb{E}[\zeta_j^2] = 2p_0$, Jensen's inequality gives:

$$\mathbb{E}_\zeta \left[\left| \sum_{j=1}^{n} \zeta_j [\Omega^{(2)}]_{jv} \right| \right] \leq \left(\mathbb{E}_\zeta \left[\left(\sum_{j=1}^{n} \zeta_j [\Omega^{(2)}]_{jv} \right)^2 \right] \right)^{1/2}$$
$$= \left(2p_0 \sum_{j=1}^{n} [\Omega^{(2)}]_{jv}^2 \right)^{1/2}$$
$$= \sqrt{2p_0} \left(\sum_{j=1}^{n} [\Omega^{(2)}]_{jv}^2 \right)^{1/2}, \tag{26}$$

where the first equality follows from the independence of the transductive Rademacher variables, since for $j \neq j'$, $\mathbb{E}[\zeta_j \zeta_{j'}] = \mathbb{E}[\zeta_j]\mathbb{E}[\zeta_{j'}] = 0$. Substituting Eq. (26) into Eq. (25) and applying the Cauchy-Schwarz inequality over v, we obtain:

$$\mathbb{E}_\zeta \left[\sup_{f\in\mathcal{F}} \left\| \sum_{j=1}^{n} \zeta_j \sum_{v\in\mathcal{N}(j)} [\Omega^{(2)}]_{jv} h_v^{(1)} \right\|_2 \right]$$
$$\leq R_1 \sqrt{2p_0} \sum_{v\in V} \left(\sum_{j=1}^{n} [\Omega^{(2)}]_{jv}^2 \right)^{1/2} \left\| \sum_{i\in\mathcal{N}(v)} [\Omega^{(1)}]_{vi} x_i \right\|_2$$
$$\leq R_1 \sqrt{2p_0} \left(\sum_{v\in V} \sum_{j=1}^{n} [\Omega^{(2)}]_{jv}^2 \right)^{1/2} \left(\sum_{v\in V} \left\| \sum_{i\in\mathcal{N}(v)} [\Omega^{(1)}]_{vi} x_i \right\|_2^2 \right)^{1/2}$$
$$\leq R_1 \sqrt{2p_0} \|\Omega^{(2)}\|_F \|\Omega^{(1)} X\|_F$$
$$\leq R_1 \sqrt{2p_0} \|\Omega^{(2)}\|_F \|\Omega^{(1)}\|_2 \|X\|_F .$$

By the definition of the Frobenius norm and the assumption that $\|x_i\|_2 \leq B$ for all nodes i, we have:

$$\|X\|_F^2 = \sum_{i=1}^{n} \|x_i\|_2^2 \leq \sum_{i=1}^{n} B^2 = nB^2 .$$

Moreover, since $|\omega_{vu}^{(k)}| \leq 1$, we have:

$$\|\Omega^{(k)}\|_2 \leq \|\Omega^{(k)}\|_F \leq \sqrt{\|\Omega^{(k)}\|_0} = \sqrt{\eta^{(k)}} .$$

Combining the above results gives:

$$\mathbb{E}_\zeta \left[\sup_{f\in\mathcal{F}} \left\| \sum_{j=1}^{n} \zeta_j \sum_{v\in\mathcal{N}(j)} [\Omega^{(2)}]_{jv} h_v^{(1)} \right\|_2 \right] \leq R_1 B \sqrt{2p_0 n} \sqrt{\eta^{(1)}\eta^{(2)}} . \tag{27}$$

Substituting Eq. (27) into Eq. (24), we obtain:

$$\mathfrak{R}_{m,n}(\mathcal{F}) \leq Q_{m,n} R_1 R_2 B \sqrt{2p_0 n} \sqrt{\eta^{(1)}\eta^{(2)}}.$$

Using $Q_{m,n} = \frac{1}{m} + \frac{1}{n-m} = \frac{n}{m(n-m)}$ and $p_0 = \frac{m(n-m)}{n^2}$, we have:

$$Q_{m,n}\sqrt{2p_0 n} = \frac{n}{m(n-m)} \sqrt{\frac{2m(n-m)}{n}} = \sqrt{\frac{2n}{m(n-m)}}.$$

Therefore, we obtain:

$$\mathfrak{R}_{m,n}(\mathcal{F}) \leq R_1 R_2 B \sqrt{\frac{2n}{m(n-m)}} \sqrt{\eta^{(1)}\eta^{(2)}}. \tag{28}$$

Let $\ell \circ \mathcal{F} = \left\{ \big(\ell(f(x_1), y_1), \ldots, \ell(f(x_n), y_n)\big) : f \in \mathcal{F} \right\}$ denote the loss-composed hypothesis class. Since $\ell(\cdot, y)$ is L-Lipschitz, the contraction inequality gives:

$$\mathfrak{R}_{m,n}(\ell \circ \mathcal{F}) \leq L\, \mathfrak{R}_{m,n}(\mathcal{F}). \tag{29}$$

Finally, Combining Eq.(28), Eq.(29) and Eq.(2), for $0 < \delta < 1$, with probability at least $1 - \delta$ for all $f \in \mathcal{F}$ the following holds:

$$\begin{aligned}\epsilon(f) \leq & \hat{\epsilon}(f) + L R_1 R_2 B \sqrt{\frac{2n}{m(n-m)}} \sqrt{\eta^{(1)}\eta^{(2)}} \\ & + c_1 \frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2 \sqrt{\frac{n}{m(n-m)} \log \frac{1}{\delta}},\end{aligned}$$

where $c_1 < 5.05$ and $c_2 < 0.8$ are universal constants. This completes the proof. □

Corollary 1. *Consider the hypothesis class $\mathcal{F}$ induced by GCNs with a fixed message-passing structure. Under the assumptions of Theorem 1, for $0 < \delta < 1$, with probability at least $1 - \delta$ for all $f \in \mathcal{F}$ the following holds:*

$$\epsilon(f) \leq \hat{\epsilon}(f) + L R_1 R_2 B \sqrt{\frac{2n}{m(n-m)}}\, \eta + c_1 \frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2 \sqrt{\frac{n}{m(n-m)} \log \frac{1}{\delta}},$$

where $\eta = \left\| \hat{A} \right\|_0$ denotes the structural complexity of the model, and $c_1 < 5.05$ and $c_2 < 0.8$ are universal constants.

Proof. For a two-layer GCN, the propagation operators are fixed as $\Omega^{(1)} = \Omega^{(2)} = \hat{A}$. Therefore, the structural complexity at each layer satisfies:

$$\eta^{(1)} = \eta^{(2)} = \|\hat{A}\|_0 = \eta.$$

Applying Theorem 1 and substituting $\eta^{(1)} = \eta^{(2)} = \eta$, we obtain:

$$\epsilon(f) \leq \hat{\epsilon}(f) + L R_1 R_2 B \sqrt{\frac{2n}{m(n-m)}}\, \eta + c_1 \frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2 \sqrt{\frac{n}{m(n-m)} \log \frac{1}{\delta}}.$$

This completes the proof. □

Proposition 1. *Consider the hypothesis class $\mathcal{F}_{\mathrm{GAT}}$ induced by GATs with attention-based message passing. Under the input, activation, weight-norm, and loss assumptions of Theorem 1, for $0 < \delta < 1$, with probability at least $1 - \delta$ for all $f \in \mathcal{F}_{\mathrm{GAT}}$ the following holds:*

$$\epsilon(f) \leq \hat{\epsilon}(f) + LR_1R_2B\, n\sqrt{\frac{2}{m(n-m)}}\left(\sqrt{\eta_\tau^{(1)}\eta_\tau^{(2)}} + \mathcal{O}(\tau)\right) + c_1\frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2\sqrt{\frac{n}{m(n-m)}\log\frac{1}{\delta}},$$

where $\eta_\tau^{(k)} = \sup_{f\in\mathcal{F}_{\mathrm{GAT}}} \|\bar{A}_f^{(k)}\|_0$ for $k = 1, 2$ denotes a uniform upper bound on the effective structural complexity of each layer, and $c_1 < 5.05$ and $c_2 < 0.8$ are universal constants.

Proof. Unlike fixed propagation operators, the attention matrices in GATs depend on the learnable model parameters and therefore remain inside the supremum defining the transductive Rademacher complexity. For notational simplicity, we suppress the dependence of $A^{(k)}$ and $\bar{A}^{(k)}$ on f below. Let $Q_{m,n} = \frac{1}{m} + \frac{1}{n-m}$. Starting from the definition of transductive Rademacher complexity and applying the contraction inequality to the output activation, we have:

$$\mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GAT}}) \leq Q_{m,n}R_2\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}_{\mathrm{GAT}}}\left\|\sum_{v\in V}\left(\sum_{j=1}^{n}\zeta_j\alpha_{jv}^{(2)}\right)h_v^{(1)}\right\|_2\right]. \tag{30}$$

Applying the Cauchy-Schwarz inequality over v gives:

$$\left\|\sum_{v\in V}\left(\sum_{j=1}^{n}\zeta_j\alpha_{jv}^{(2)}\right)h_v^{(1)}\right\|_2 \leq \left(\sum_{v\in V}\left|\sum_{j=1}^{n}\zeta_j\alpha_{jv}^{(2)}\right|^2\right)^{1/2}\left(\sum_{v\in V}\|h_v^{(1)}\|_2^2\right)^{1/2}.$$

For the first factor, another application of the Cauchy-Schwarz inequality yields:

$$\left|\sum_{j=1}^{n}\zeta_j\alpha_{jv}^{(2)}\right|^2 \leq \left(\sum_{j=1}^{n}\zeta_j^2\right)\left(\sum_{j=1}^{n}[\alpha_{jv}^{(2)}]^2\right).$$

Therefore, we have:

$$\left(\sum_{v\in V}\left|\sum_{j=1}^{n}\zeta_j\alpha_{jv}^{(2)}\right|^2\right)^{1/2} \leq \left(\sum_{j=1}^{n}\zeta_j^2\right)^{1/2}\|A^{(2)}\|_F. \tag{31}$$

For the first-layer representations, since σ is 1-Lipschitz with $\sigma(0) = 0$ and $\|W^{(1)}\|_2 \leq R_1$, we have:

$$\begin{aligned}\left(\sum_{v\in V}\|h_v^{(1)}\|_2^2\right)^{1/2} &\leq R_1\left(\sum_{v\in V}\left\|\sum_{i\in\mathcal{N}(v)}\alpha_{vi}^{(1)}x_i\right\|_2^2\right)^{1/2}\\ &= R_1\|A^{(1)}X\|_F\\ &\leq R_1\|A^{(1)}\|_2\|X\|_F\\ &\leq R_1B\sqrt{n}\,\|A^{(1)}\|_2. \end{aligned}\tag{32}$$

Combining Eqs. (30), (31), and (32), we obtain:

$$\mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GAT}}) \leq Q_{m,n}R_1R_2B\sqrt{n}\,\mathbb{E}_\zeta\left[\left(\sum_{j=1}^{n}\zeta_j^2\right)^{1/2}\right]\sup_{f\in\mathcal{F}_{\mathrm{GAT}}}\left(\|A^{(2)}\|_F\|A^{(1)}\|_2\right).$$

Since the transductive Rademacher variables satisfy $\mathbb{E}[\zeta_j^2] = 2p_0$, Jensen's inequality gives:

$$\mathbb{E}_\zeta\left[\left(\sum_{j=1}^{n}\zeta_j^2\right)^{1/2}\right] \le \left(\mathbb{E}_\zeta\left[\sum_{j=1}^{n}\zeta_j^2\right]\right)^{1/2} = \sqrt{2p_0 n}.$$

Hence, we obtain:

$$\mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GAT}}) \le Q_{m,n} R_1 R_2 B\sqrt{2p_0}\, n \sup_{f\in\mathcal{F}_{\mathrm{GAT}}}\left(\|A^{(2)}\|_F \|A^{(1)}\|_2\right). \tag{33}$$

Using $Q_{m,n} = \frac{n}{m(n-m)}$ and $p_0 = \frac{m(n-m)}{n^2}$ gives:

$$Q_{m,n}\sqrt{2p_0}n = n\sqrt{\frac{2}{m(n-m)}}.$$

To express the bound in terms of effective structural complexity, we decompose each attention matrix as $A^{(k)} = \bar{A}^{(k)} + E^{(k)}$ for $k = 1, 2$, where $\bar{A}^{(k)}$ is the thresholded attention matrix defined in Section 3.4 and $E^{(k)}$ is the residual matrix containing elements smaller than τ. For the first layer, by the triangle inequality of the spectral norm, we have:

$$\|A^{(1)}\|_2 \le \|\bar{A}^{(1)}\|_2 + \|E^{(1)}\|_2.$$

Since the attention coefficients are bounded by 1 and $\|\bar{A}^{(1)}\|_0 \le \eta_\tau^{(1)}$ uniformly over $f \in \mathcal{F}_{\mathrm{GAT}}$, we have:

$$\|\bar{A}^{(1)}\|_2 \le \|\bar{A}^{(1)}\|_F \le \sqrt{\eta_\tau^{(1)}}.$$

For the residual component $E^{(1)}$, since all entries satisfy $0 \le E_{ij}^{(1)} < \tau$, both its maximum row sum and maximum column sum are bounded by $d_{\max}\tau$. Applying the standard norm inequality $\|E\|_2 \le \sqrt{\|E\|_1\|E\|_\infty}$, its spectral norm is bounded by $\|E^{(1)}\|_2 < d_{\max}\tau = \mathcal{O}(\tau)$, where $d_{\max}$ denotes the maximum node degree. Combining these yields:

$$\left\|A^{(1)}\right\|_2 \le \sqrt{\eta_\tau^{(1)}} + \mathcal{O}(\tau).$$

For the second layer, by the triangle inequality of the Frobenius norm, we have:

$$\|A^{(2)}\|_F \le \|\bar{A}^{(2)}\|_F + \|E^{(2)}\|_F.$$

Similarly, since the attention coefficients are bounded by 1 and $\|\bar{A}^{(2)}\|_0 \le \eta_\tau^{(2)}$, we have:

$$\|\bar{A}^{(2)}\|_F \le \sqrt{\eta_\tau^{(2)}}.$$

Moreover, since $E^{(2)}$ contains at most $nd_{\max}$ nonzero entries, each smaller than τ, its Frobenius norm satisfies:

$$\|E^{(2)}\|_F^2 = \sum_{v,u}|E_{vu}^{(2)}|^2 \le nd_{\max}\tau^2,$$

and hence $\|E^{(2)}\|_F \le \sqrt{nd_{\max}}\,\tau = \mathcal{O}(\tau)$. Therefore, we have:

$$\|A^{(2)}\|_F \le \sqrt{\eta_\tau^{(2)}} + \mathcal{O}(\tau).$$

Substituting the above bounds for $\|A^{(1)}\|_2$ and $\|A^{(2)}\|_F$ into Eq.(33), we obtain:

$$\begin{aligned}\mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GAT}}) \le& R_1 R_2 B\, n\sqrt{\frac{2}{m(n-m)}}\left(\sqrt{\eta_\tau^{(2)}} + \mathcal{O}(\tau)\right)\left(\sqrt{\eta_\tau^{(1)}} + \mathcal{O}(\tau)\right)\\ =& R_1 R_2 B\, n\sqrt{\frac{2}{m(n-m)}}\left(\sqrt{\eta_\tau^{(1)}\eta_\tau^{(2)}} + \mathcal{O}(\tau)\right).\end{aligned} \tag{34}$$

Let $\ell \circ \mathcal{F}_{\mathrm{GAT}}$ denote the corresponding loss-composed hypothesis class. Since $\ell(\cdot, y)$ is L-Lipschitz, the contraction inequality gives:

$$\mathfrak{R}_{m,n}(\ell \circ \mathcal{F}_{\mathrm{GAT}}) \leq L\, \mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GAT}}). \tag{35}$$

Finally, combining Eqs. (34) and (35) with Eq. (2), we conclude that for $0 < \delta < 1$, with probability at least $1 - \delta$ for all $f \in \mathcal{F}_{\mathrm{GAT}}$ the following holds:

$$\begin{aligned}\epsilon(f) \leq & \hat{\epsilon}(f) + LR_1R_2B\, n\sqrt{\frac{2}{m(n-m)}}\left(\sqrt{\eta_\tau^{(1)}\eta_\tau^{(2)}} + \mathcal{O}(\tau)\right) \\ & + c_1\frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2\sqrt{\frac{n}{m(n-m)}\log\frac{1}{\delta}}.\end{aligned}$$

This completes the proof. □

Having established Proposition 1 for single-head attention-based GATs, we extend that architecture-specific analysis to the multi-head setting, which constitutes the canonical and widely adopted form of GAT architectures. Consider a two-layer GAT with Δ attention heads per layer. For each layer $k \in \{1, 2\}$ and each head $r_k \in \{1, \ldots, \Delta\}$, let $\bar{A}_f^{(k,r_k)}$ denote the thresholded attention matrix of the r_k-th head at layer k. Message passing within each head is governed independently by its corresponding attention matrix, and the outputs of different heads are aggregated through concatenation or averaging.

From the perspective of structural complexity, each attention head induces its own set of effective aggregation relations. Accordingly, we define the aggregated effective structural complexity at layer k as:

$$\eta_{\mathrm{MH},\tau}^{(k)} = \sup_{f \in \mathcal{F}_{\mathrm{GAT}}} \sum_{r_k=1}^{\Delta} \left\| \bar{A}_f^{(k,r_k)} \right\|_0, \tag{36}$$

which provides a uniform upper bound on the total number of effective aggregation relations across all heads. The following corollary extends the single-head analysis and characterizes how this aggregated structural complexity enters the generalization bound of multi-head GATs.

Proposition 2. *Consider the hypothesis class $\mathcal{F}_{\mathrm{GAT}}$ induced by a two-layer multi-head GAT with Δ attention heads per layer. Assume that the input node features satisfy $\|x_i\|_2 \leq B$, the activation function σ is 1-Lipschitz with $\sigma(0) = 0$, the weight matrices satisfy $\left\|W^{(1,r_1)}\right\|_2 \leq R_1$ and $\left\|W^{(2,r_2)}\right\|_2 \leq R_2$, and the loss function $\ell(\cdot, y)$ is L-Lipschitz continuous. Then for any $0 < \delta < 1$, with probability at least $1 - \delta$ for all $f \in \mathcal{F}_{\mathrm{GAT}}$ the following holds:*

$$\begin{aligned}\epsilon(f) \leq & \hat{\epsilon}(f) + LR_1R_2B\, n\sqrt{\frac{2}{m(n-m)\Delta}}\left(\sqrt{\eta_{\mathrm{MH},\tau}^{(1)}\eta_{\mathrm{MH},\tau}^{(2)}} + \mathcal{O}(\tau)\right) \\ & + c_1\frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2\sqrt{\frac{n}{m(n-m)}\log\frac{1}{\delta}},\end{aligned} \tag{37}$$

where $\eta_{\mathrm{MH},\tau}^{(k)} = \sup_{f \in \mathcal{F}_{\mathrm{GAT}}} \sum_{r_k=1}^{\Delta} \left\| \bar{A}_f^{(k,r_k)} \right\|_0$ for $k = 1, 2$ denotes a uniform upper bound on the aggregated effective structural complexity, and $c_1 < 5.05$ and $c_2 < 0.8$ are universal constants.

Proof. We sketch the proof by extending the single-head GAT analysis to the multi-head setting. For simplicity, we omit the dependence of the attention matrices on f. Let $Q_{m,n} = \frac{1}{m} + \frac{1}{n-m}$. For a two-layer GAT with Δ attention heads per layer, the hidden representation produced by the second layer can be written as an aggregation over all heads. The prediction function then takes the following form:

$$f(x_j) = \sigma\left(\frac{1}{\Delta}\sum_{r_2=1}^{\Delta}\sum_{v \in \mathcal{N}(j)} \alpha_{jv}^{(2,r_2)}\left(W^{(2,r_2)}\right)^\top h_v^{(1)}\right),$$

where the first-layer representations are given by:

$$h_v^{(1)} = \Big\Vert_{r_1=1}^{\Delta} \sigma\left(\sum_{i\in\mathcal{N}(v)} \alpha_{vi}^{(1,r_1)} \left(W^{(1,r_1)}\right)^\top x_i\right), \qquad r_1 = 1,\dots,\Delta.$$

Substituting the above multi-head expression into the definition of the transductive Rademacher complexity in Eq.(23), and applying the Cauchy–Schwarz inequality together with the 1-Lipschitz property of the activation function, we obtain:

$$\begin{aligned}
\mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GAT}}) &= Q_{m,n}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}_{\mathrm{GAT}}} \sum_{j=1}^n \zeta_j f(x_j)\right] \\
&= Q_{m,n}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}_{\mathrm{GAT}}} \sum_{j=1}^n \zeta_j \sigma\left(\frac{1}{\Delta}\sum_{r_2=1}^{\Delta}\sum_{v\in\mathcal{N}(j)} \alpha_{jv}^{(2,r_2)} \left(W^{(2,r_2)}\right)^\top h_v^{(1)}\right)\right] \\
&\le \frac{Q_{m,n}}{\Delta}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}_{\mathrm{GAT}}} \sum_{j=1}^n \zeta_j \sum_{r_2=1}^{\Delta}\sum_{v\in\mathcal{N}(j)} \alpha_{jv}^{(2,r_2)} \left(W^{(2,r_2)}\right)^\top h_v^{(1)}\right] \\
&\le \frac{Q_{m,n}}{\Delta}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}_{\mathrm{GAT}}} \sum_{r_2=1}^{\Delta}\left\|\sum_{j=1}^n \zeta_j \sum_{v\in\mathcal{N}(j)} \alpha_{jv}^{(2,r_2)} h_v^{(1)}\right\|_2 \left\|W^{(2,r_2)}\right\|_2\right] \\
&\le \frac{Q_{m,n}R_2}{\Delta}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}_{\mathrm{GAT}}} \sum_{r_2=1}^{\Delta}\left\|\sum_{j=1}^n \zeta_j \sum_{v\in\mathcal{N}(j)} \alpha_{jv}^{(2,r_2)} h_v^{(1)}\right\|_2\right] \\
&\le \frac{Q_{m,n}R_2}{\Delta}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}_{\mathrm{GAT}}} \sqrt{\Delta}\left(\sum_{r_2=1}^{\Delta}\left\|\sum_{j=1}^n \zeta_j \sum_{v\in\mathcal{N}(j)} \alpha_{jv}^{(2,r_2)} h_v^{(1)}\right\|_2^2\right)^{1/2}\right] \\
&= \frac{Q_{m,n}R_2}{\sqrt{\Delta}}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}_{\mathrm{GAT}}} \left(\sum_{r_2=1}^{\Delta}\left\|\sum_{j=1}^n \zeta_j \sum_{v\in\mathcal{N}(j)} \alpha_{jv}^{(2,r_2)} h_v^{(1)}\right\|_2^2\right)^{1/2}\right],
\end{aligned}$$

where the third inequality follows from the norm constraint $\|W^{(2,r_2)}\|_2 \le R_2$. For each second-layer attention head, exchanging the order of summation over j and v and applying the Cauchy-Schwarz inequality over v gives:

$$\begin{aligned}
&\left\|\sum_{j=1}^n \zeta_j \sum_{v\in\mathcal{N}(j)} \alpha_{jv}^{(2,r_2)} h_v^{(1)}\right\|_2 \\
&= \left\|\sum_{v\in V}\left(\sum_{j=1}^n \zeta_j \alpha_{jv}^{(2,r_2)}\right) h_v^{(1)}\right\|_2 \\
&\le \left(\sum_{v\in V}\left|\sum_{j=1}^n \zeta_j \alpha_{jv}^{(2,r_2)}\right|^2\right)^{1/2}\left(\sum_{v\in V}\|h_v^{(1)}\|_2^2\right)^{1/2}.
\end{aligned}$$

Therefore, we have:

$$\left(\sum_{r_2=1}^{\Delta}\left\|\sum_{j=1}^n \zeta_j \sum_{v\in\mathcal{N}(j)} \alpha_{jv}^{(2,r_2)} h_v^{(1)}\right\|_2^2\right)^{1/2} \le \left(\sum_{r_2=1}^{\Delta}\sum_{v\in V}\left|\sum_{j=1}^n \zeta_j \alpha_{jv}^{(2,r_2)}\right|^2\right)^{1/2}\left(\sum_{v\in V}\|h_v^{(1)}\|_2^2\right)^{1/2}.$$

For each r_2 and v, the Cauchy–Schwarz inequality gives:

$$\left|\sum_{j=1}^{n}\zeta_j\alpha_{jv}^{(2,r_2)}\right|^2 \le \left(\sum_{j=1}^{n}\zeta_j^2\right)\left(\sum_{j=1}^{n}\left(\alpha_{jv}^{(2,r_2)}\right)^2\right).$$

Hence, we obtain:

$$\begin{aligned}
&\left(\sum_{r_2=1}^{\Delta}\sum_{v\in V}\left|\sum_{j=1}^{n}\zeta_j\alpha_{jv}^{(2,r_2)}\right|^2\right)^{1/2}\\
&\le\left(\sum_{j=1}^{n}\zeta_j^2\right)^{1/2}\left(\sum_{r_2=1}^{\Delta}\sum_{v\in V}\sum_{j=1}^{n}\left(\alpha_{jv}^{(2,r_2)}\right)^2\right)^{1/2}\\
&\le\left(\sum_{j=1}^{n}\zeta_j^2\right)^{1/2}\left(\sum_{r_2=1}^{\Delta}\|A^{(2,r_2)}\|_F^2\right)^{1/2}.
\end{aligned}$$

For the first-layer representations, since the representations are concatenated across attention heads, σ is 1-Lipschitz with $\sigma(0)=0$, and $\|W^{(1,r_1)}\|_2 \le R_1$, we have:

$$\begin{aligned}
\sum_{v\in V}\|h_v^{(1)}\|_2^2 &= \sum_{v\in V}\left\|\Big\|_{r_1=1}^{\Delta}\sigma\left(\sum_{i\in\mathcal{N}(v)}\alpha_{vi}^{(1,r_1)}\left(W^{(1,r_1)}\right)^\top x_i\right)\right\|_2^2\\
&\le R_1^2\sum_{r_1=1}^{\Delta}\sum_{v\in V}\left\|\sum_{i\in\mathcal{N}(v)}\alpha_{vi}^{(1,r_1)}x_i\right\|_2^2\\
&= R_1^2\sum_{r_1=1}^{\Delta}\|A^{(1,r_1)}X\|_F^2\\
&\le R_1^2\|X\|_F^2\sum_{r_1=1}^{\Delta}\|A^{(1,r_1)}\|_2^2.
\end{aligned}$$

Therefore, using $\|X\|_F \le \sqrt{n}B$, we obtain:

$$\left(\sum_{v\in V}\|h_v^{(1)}\|_2^2\right)^{1/2} \le R_1B\sqrt{n}\left(\sum_{r_1=1}^{\Delta}\|A^{(1,r_1)}\|_2^2\right)^{1/2}.$$

Combining the above results, we obtain:

$$\begin{aligned}
\mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GAT}}) \le& Q_{m,n}R_1R_2B\sqrt{\frac{n}{\Delta}}\mathbb{E}_\zeta\left[\left(\sum_{j=1}^{n}\zeta_j^2\right)^{1/2}\right]\\
&\times\sup_{f\in\mathcal{F}_{\mathrm{GAT}}}\left(\sum_{r_2=1}^{\Delta}\|A^{(2,r_2)}\|_F^2\right)^{1/2}\left(\sum_{r_1=1}^{\Delta}\|A^{(1,r_1)}\|_2^2\right)^{1/2}. \qquad (38)
\end{aligned}$$

Since the transductive Rademacher variables satisfy $\mathbb{E}[\zeta_j^2]=2p_0$, Jensen's inequality gives:

$$\mathbb{E}_\zeta\left[\left(\sum_{j=1}^{n}\zeta_j^2\right)^{1/2}\right] \le \left(\mathbb{E}_\zeta\left[\sum_{j=1}^{n}\zeta_j^2\right]\right)^{1/2} = \sqrt{2p_0n}.$$

Substituting this result into Eq. (38), we obtain:

$$\mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GAT}}) \leq Q_{m,n} R_1 R_2 B\, n \sqrt{\frac{2p_0}{\Delta}} \sup_{f \in \mathcal{F}_{\mathrm{GAT}}} \left(\sum_{r_2=1}^{\Delta} \|A^{(2,r_2)}\|_F^2 \right)^{1/2} \left(\sum_{r_1=1}^{\Delta} \|A^{(1,r_1)}\|_2^2 \right)^{1/2}.$$

Using $Q_{m,n} = \frac{1}{m} + \frac{1}{n-m} = \frac{n}{m(n-m)}$ and $p_0 = \frac{m(n-m)}{n^2}$, we have:

$$Q_{m,n} n \sqrt{\frac{2p_0}{\Delta}} = \frac{n^2}{m(n-m)} \sqrt{\frac{2m(n-m)}{n^2 \Delta}} = n \sqrt{\frac{2}{m(n-m)\Delta}}.$$

Therefore, we obtain:

$$\begin{aligned} \mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GAT}}) \leq & R_1 R_2 B\, n \sqrt{\frac{2}{m(n-m)\Delta}} \\ & \times \sup_{f \in \mathcal{F}_{\mathrm{GAT}}} \left(\sum_{r_2=1}^{\Delta} \|A^{(2,r_2)}\|_F^2 \right)^{1/2} \left(\sum_{r_1=1}^{\Delta} \|A^{(1,r_1)}\|_2^2 \right)^{1/2}. \end{aligned} \tag{39}$$

To express the bound in terms of effective structural complexity, we decompose each attention matrix as $A^{(k,r_k)} = \bar{A}^{(k,r_k)} + E^{(k,r_k)}$ for $k = 1, 2$ and $r_k = 1, \ldots, \Delta$, where $\bar{A}^{(k,r_k)}$ is the thresholded matrix defined in Section 3.4 and $E^{(k,r_k)}$ is the residual matrix containing entries smaller than τ. For the first layer, we have:

$$\|A^{(1,r_1)}\|_2 \leq \|\bar{A}^{(1,r_1)}\|_2 + \|E^{(1,r_1)}\|_2.$$

Since the attention coefficients are bounded by 1, $\|\bar{A}^{(1,r_1)}\|_2 \leq \|\bar{A}^{(1,r_1)}\|_F \leq \sqrt{\|\bar{A}^{(1,r_1)}\|_0}$. For the residual component $E^{(k,r_k)}$, since all entries satisfy $0 \leq E_{ij}^{(k,r_k)} < \tau$, both its maximum row sum and maximum column sum are bounded by $d_{\max}\tau$, where $d_{\max}$ denotes the maximum node degree. Applying the standard norm inequality $\|E\|_2 \leq \sqrt{\|E\|_1 \|E\|_\infty}$, its spectral norm is bounded by $\|E^{(1,r_1)}\|_2 < d_{\max}\tau = \mathcal{O}(\tau)$. Therefore, we have:

$$\left(\sum_{r_1=1}^{\Delta} \|A^{(1,r_1)}\|_2^2 \right)^{1/2} \leq \left[\sum_{r_1=1}^{\Delta} \left(\sqrt{\|\bar{A}^{(1,r_1)}\|_0} + \mathcal{O}(\tau) \right)^2 \right]^{1/2} \leq \sqrt{\eta_{\mathrm{MH},\tau}^{(1)}} + \mathcal{O}(\tau).$$

For the second layer, by the triangle inequality of the Frobenius norm, we have:

$$\|A^{(2,r_2)}\|_F \leq \|\bar{A}^{(2,r_2)}\|_F + \|E^{(2,r_2)}\|_F.$$

Since the attention coefficients are bounded by 1, $\|\bar{A}^{(2,r_2)}\|_F \leq \sqrt{\|\bar{A}^{(2,r_2)}\|_0}$. Moreover, since $E^{(2,r_2)}$ contains at most $n d_{\max}$ nonzero entries, each smaller than τ, its Frobenius norm satisfies:

$$\|E^{(2,r_2)}\|_F^2 = \sum_{v,u} |E_{vu}^{(2,r_2)}|^2 \leq n d_{\max} \tau^2,$$

and hence $\|E^{(2,r_2)}\|_F \leq \sqrt{n d_{\max}}\, \tau = \mathcal{O}(\tau)$. Hence, we have:

$$\left(\sum_{r_2=1}^{\Delta} \|A^{(2,r_2)}\|_F^2 \right)^{1/2} \leq \left[\sum_{r_2=1}^{\Delta} \left(\sqrt{\|\bar{A}^{(2,r_2)}\|_0} + \mathcal{O}(\tau) \right)^2 \right]^{1/2} \leq \sqrt{\eta_{\mathrm{MH},\tau}^{(2)}} + \mathcal{O}(\tau).$$

Substituting these bounds into Eq. (39) gives:

$$\begin{aligned} \mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GAT}}) \leq & R_1 R_2 B\, n \sqrt{\frac{2}{m(n-m)\Delta}} \left(\sqrt{\eta_{\mathrm{MH},\tau}^{(2)}} + \mathcal{O}(\tau) \right) \left(\sqrt{\eta_{\mathrm{MH},\tau}^{(1)}} + \mathcal{O}(\tau) \right) \\ = & R_1 R_2 B\, n \sqrt{\frac{2}{m(n-m)\Delta}} \left(\sqrt{\eta_{\mathrm{MH},\tau}^{(1)} \eta_{\mathrm{MH},\tau}^{(2)}} + \mathcal{O}(\tau) \right). \end{aligned} \tag{40}$$

Let $\ell \circ \mathcal{F}_{\mathrm{GAT}}$ denote the corresponding loss-composed hypothesis class. Since $\ell(\cdot, y)$ is L-Lipschitz, the contraction inequality gives:

$$\mathfrak{R}_{m,n}(\ell \circ \mathcal{F}_{\mathrm{GAT}}) \leq L\, \mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GAT}}). \tag{41}$$

Finally, combining Eq. (40), Eq. (41), and Eq. (2), we conclude that for $0 < \delta < 1$, with probability at least $1 - \delta$ for all $f \in \mathcal{F}_{\mathrm{GAT}}$, the following holds:

$$\begin{aligned}\epsilon(f) \leq & \hat{\epsilon}(f) + LR_1R_2B\, n\sqrt{\frac{2}{m(n-m)\Delta}}\left(\sqrt{\eta^{(1)}_{\mathrm{MH},\tau}\eta^{(2)}_{\mathrm{MH},\tau}} + \mathcal{O}(\tau)\right) \\ & + c_1\frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2\sqrt{\frac{n}{m(n-m)}\log\frac{1}{\delta}}.\end{aligned}$$

This completes the proof. □

Assumption 1. *The normalization operator* $\mathrm{Norm}(\cdot)$ *is* Π_{Norm}*-Lipschitz and satisfies* $\mathrm{Norm}(0) = 0$*. Specifically, for all* $a, b \in \mathbb{R}^{\widetilde{d}}$*,* $\|\mathrm{Norm}(a) - \mathrm{Norm}(b)\|_2 \leq \Pi_{\mathrm{Norm}}\|a - b\|_2$.

For LayerNorm, Assumption 1 holds under a positive stabilization constant $\epsilon_{\mathrm{LN}} > 0$, a bounded affine scaling parameter γ satisfying $\|\gamma\|_\infty \leq R_\gamma$, and zero affine shift. Under Assumption 1, we obtain the following generalization bound for a single-layer GT.

Proposition 3. *Consider the hypothesis class* $\mathcal{F}_{\mathrm{GT}}$ *induced by a single-layer GT. Under the input, activation, and loss assumptions of Theorem 1, assume further that the projection matrices satisfy* $\left\|V^{(1,r_1)}\right\|_2 \leq R_0$*,* $\left\|W_1^{(1)}\right\|_2 \leq R_1$ *and* $\left\|W_2^{(1)}\right\|_2 \leq R_2$*, the output projection matrix satisfies* $\left\|O^{(1)}\right\|_2 \leq \widetilde{R}$*, the output weight vector satisfies* $\|w_{\mathrm{out}}\|_2 \leq R_{\mathrm{out}}$*, and Assumption 1 holds. Then for any* $0 < \delta < 1$*, with probability at least* $1 - \delta$ *for all* $f \in \mathcal{F}_{\mathrm{GT}}$ *the following holds:*

$$\begin{aligned}\epsilon(f) \leq & \hat{\epsilon}(f) + \Pi^2_{\mathrm{Norm}}LBR_{\mathrm{out}}(1 + R_1R_2)n\sqrt{\frac{2}{m(n-m)}}\left[1 + R_0\widetilde{R}\left(\sqrt{\eta^{(1)}_{\mathrm{MHT},\tau}} + \mathcal{O}(\tau)\right)\right] \\ & + c_1\frac{n\sqrt{\min\{m, n-m\}}}{m(n-m)} + c_2\sqrt{\frac{n}{m(n-m)}\log\frac{1}{\delta}},\end{aligned} \tag{42}$$

where $\eta^{(1)}_{\mathrm{MHT},\tau} = \sup_{f\in\mathcal{F}_{\mathrm{GT}}}\sum_{r_1=1}^{\Delta}\left\|\bar{A}_f^{(1,r_1)}\right\|_0$ *denotes a uniform upper bound on the aggregated attention-driven structural complexity, and* $c_1 < 5.05$ *and* $c_2 < 0.8$ *are universal constants.*

Proof. As in Proposition 1, the attention operators depend on the predictor, so this result is derived from the GT-specific complexity expression rather than as a direct specialization of Theorem 1. For notational simplicity, we suppress the dependence of the intermediate representations and attention matrices on f whenever there is no ambiguity. Let $Q_{m,n} = \frac{1}{m} + \frac{1}{n-m}$. Let $H^{(1)} \in \mathbb{R}^{n\times\widetilde{d}}$ denote the final representation matrix of the single-layer GT, whose j-th row is $(h_j^{(1)})^\top$. Using the scalar readout in Eq. (15), the transductive Rademacher complexity satisfies:

$$\begin{aligned}\mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GT}}) &= Q_{m,n}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}_{\mathrm{GT}}}\sum_{j=1}^{n}\zeta_j w_{\mathrm{out}}^\top h_j^{(1)}\right] \\ &\leq Q_{m,n}R_{\mathrm{out}}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}_{\mathrm{GT}}}\left\|\sum_{j=1}^{n}\zeta_j h_j^{(1)}\right\|_2\right] \\ &= Q_{m,n}R_{\mathrm{out}}\mathbb{E}_\zeta\left[\sup_{f\in\mathcal{F}_{\mathrm{GT}}}\left\|(H^{(1)})^\top\zeta\right\|_2\right] \\ &\leq Q_{m,n}R_{\mathrm{out}}\mathbb{E}_\zeta\left[\|\zeta\|_2\sup_{f\in\mathcal{F}_{\mathrm{GT}}}\|H^{(1)}\|_F\right]\end{aligned}$$

$$= Q_{m,n} R_{\text{out}} \mathbb{E}_\zeta \|\zeta\|_2 \sup_{f \in \mathcal{F}_{\text{GT}}} \|H^{(1)}\|_F$$
$$\leq Q_{m,n} R_{\text{out}} \sqrt{2p_0 n} \sup_{f \in \mathcal{F}_{\text{GT}}} \|H^{(1)}\|_F. \tag{43}$$

where the first inequality follows from the Cauchy-Schwarz inequality and $\|w_{\text{out}}\|_2 \leq R_{\text{out}}$. The second inequality uses $\|(H^{(1)})^\top \zeta\|_2 \leq \|H^{(1)}\|_F \|\zeta\|_2$. Finally, since $\mathbb{E}[\zeta_j^2] = 2p_0$, Jensen's inequality gives $\mathbb{E}_\zeta \|\zeta\|_2 \leq \sqrt{\mathbb{E}_\zeta \|\zeta\|_2^2} = \sqrt{2p_0 n}$.

The activation assumption implies $\|\sigma(v)\|_2 \leq \|v\|_2$. Together with Assumption 1 and the weight constraints, this yields:

$$\begin{aligned}
\|h_j^{(1)}\|_2 &\leq \Pi_{\text{Norm}} \left\| \widetilde{h}_j^{(1)} + W_2^{(1)} \sigma\left(W_1^{(1)} \widetilde{h}_j^{(1)}\right) \right\|_2 \\
&\leq \Pi_{\text{Norm}} \left(\|\widetilde{h}_j^{(1)}\|_2 + \|W_2^{(1)}\|_2 \left\| \sigma\left(W_1^{(1)} \widetilde{h}_j^{(1)}\right) \right\|_2 \right) \\
&\leq \Pi_{\text{Norm}} \left(1 + R_2 R_1\right) \|\widetilde{h}_j^{(1)}\|_2 \\
&\leq \Pi_{\text{Norm}}^2 (1 + R_1 R_2) \|x_j + \hat{h}_j^{(1)}\|_2.
\end{aligned}$$

Taking the sum of squared node norms and then applying the Frobenius triangle inequality gives:

$$\begin{aligned}
\|H^{(1)}\|_F &\leq \Pi_{\text{Norm}}^2 (1 + R_1 R_2) \|X + \hat{H}^{(1)}\|_F \\
&\leq \Pi_{\text{Norm}}^2 (1 + R_1 R_2) \left(\|X\|_F + \|\hat{H}^{(1)}\|_F \right).
\end{aligned} \tag{44}$$

We next bound the Frobenius norm of the attention output. Using the spectral norm constraints $\|O^{(1)}\|_2 \leq \widetilde{R}$ and $\|V^{(1,r_1)}\|_2 \leq R_0$, together with the norm identity for concatenated vectors, we obtain:

$$\begin{aligned}
\|\hat{H}^{(1)}\|_F^2 &= \sum_{j=1}^n \|\hat{h}_j^{(1)}\|_2^2 \\
&= \sum_{j=1}^n \left\| O^{(1)} \Big\|_{r_1=1}^{\Delta} \left(\sum_{v \in V} \alpha_{jv}^{(1,r_1)} V^{(1,r_1)} x_v \right) \right\|_2^2 \\
&\leq \widetilde{R}^2 \sum_{j=1}^n \left\| \Big\|_{r_1=1}^{\Delta} \left(\sum_{v \in V} \alpha_{jv}^{(1,r_1)} V^{(1,r_1)} x_v \right) \right\|_2^2 \\
&= \widetilde{R}^2 \sum_{j=1}^n \sum_{r_1=1}^{\Delta} \left\| V^{(1,r_1)} \sum_{v \in V} \alpha_{jv}^{(1,r_1)} x_v \right\|_2^2 \\
&\leq \widetilde{R}^2 R_0^2 \sum_{r_1=1}^{\Delta} \sum_{j=1}^n \left\| \sum_{v \in V} \alpha_{jv}^{(1,r_1)} x_v \right\|_2^2 \\
&= \widetilde{R}^2 R_0^2 \sum_{r_1=1}^{\Delta} \|A^{(1,r_1)} X\|_F^2 \\
&\leq \widetilde{R}^2 R_0^2 \|X\|_F^2 \sum_{r_1=1}^{\Delta} \|A^{(1,r_1)}\|_2^2,
\end{aligned} \tag{45}$$

where the last inequality uses $\|AX\|_F \leq \|A\|_2 \|X\|_F$. Taking square roots in Eq. (45) gives:

$$\|\hat{H}^{(1)}\|_F \leq \widetilde{R} R_0 \|X\|_F \left(\sum_{r_1=1}^{\Delta} \|A^{(1,r_1)}\|_2^2 \right)^{1/2}. \tag{46}$$

To express the bound in terms of effective structural complexity, we decompose each attention matrix as $A^{(1,r_1)} = \bar{A}^{(1,r_1)} + E^{(1,r_1)}$ for $r_1 = 1, \ldots, \Delta$, where $\bar{A}^{(1,r_1)}$ is the thresholded matrix defined in

Section 3.4 and $E^{(1,r_1)}$ is the residual matrix with entries smaller than τ. By the triangle inequality of the spectral norm, we have:

$$\left\|A^{(1,r_1)}\right\|_2 \le \left\|\bar{A}^{(1,r_1)}\right\|_2 + \left\|E^{(1,r_1)}\right\|_2.$$

For the main component $\bar{A}^{(1,r_1)}$, since its entries are bounded by 1, $\|\bar{A}^{(1,r_1)}\|_2 \le \|\bar{A}^{(1,r_1)}\|_F \le \sqrt{\|\bar{A}^{(1,r_1)}\|_0}$. For the residual component $E^{(1,r_1)}$, since all entries satisfy $0 \le E_{ij}^{(1,r_1)} < \tau$, its spectral norm satisfies $\left\|E^{(1,r_1)}\right\|_2 \le \left\|E^{(1,r_1)}\right\|_F < n\tau = \mathcal{O}(\tau)$. Therefore, we have:

$$\begin{aligned}
&\left(\sum_{r_1=1}^{\Delta} \|A^{(1,r_1)}\|_2^2\right)^{1/2} \\
\le &\left[\sum_{r_1=1}^{\Delta} \left(\sqrt{\|\bar{A}^{(1,r_1)}\|_0} + \mathcal{O}(\tau)\right)^2\right]^{1/2} \\
\le &\left(\sum_{r_1=1}^{\Delta} \|\bar{A}^{(1,r_1)}\|_0\right)^{1/2} + \mathcal{O}(\tau) \\
\le &\sqrt{\eta^{(1)}_{\mathrm{MHT},\tau}} + \mathcal{O}(\tau).
\end{aligned} \tag{47}$$

Combining Eqs. (46) and (47) yields:

$$\|\hat{H}^{(1)}\|_F \le \widetilde{R} R_0 \|X\|_F \left(\sqrt{\eta^{(1)}_{\mathrm{MHT},\tau}} + \mathcal{O}(\tau)\right). \tag{48}$$

Substituting Eq. (48) into Eq. (44), we obtain:

$$\|H^{(1)}\|_F \le \Pi^2_{\mathrm{Norm}}(1 + R_1 R_2)\|X\|_F \left[1 + R_0 \widetilde{R}\left(\sqrt{\eta^{(1)}_{\mathrm{MHT},\tau}} + \mathcal{O}(\tau)\right)\right].$$

Using $\|x_j\|_2 \le B$ for all nodes, we have $\|X\|_F \le B\sqrt{n}$. Therefore, we can obtain:

$$\sup_{f \in \mathcal{F}_{\mathrm{GT}}} \|H^{(1)}\|_F \le \Pi^2_{\mathrm{Norm}}(1 + R_1 R_2) B\sqrt{n} \left[1 + R_0 \widetilde{R}\left(\sqrt{\eta^{(1)}_{\mathrm{MHT},\tau}} + \mathcal{O}(\tau)\right)\right]. \tag{49}$$

Substituting Eq. (49) into Eq. (43) gives:

$$\mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GT}}) \le Q_{m,n} R_{\mathrm{out}} \sqrt{2p_0 n}\, \Pi^2_{\mathrm{Norm}}(1 + R_1 R_2) B\sqrt{n} \left[1 + R_0 \widetilde{R}\left(\sqrt{\eta^{(1)}_{\mathrm{MHT},\tau}} + \mathcal{O}(\tau)\right)\right].$$

Using $Q_{m,n} = \frac{1}{m} + \frac{1}{n-m} = \frac{n}{m(n-m)}$ and $p_0 = \frac{m(n-m)}{n^2}$, we have:

$$Q_{m,n}\sqrt{2p_0 n}\sqrt{n} = \frac{n}{m(n-m)} \sqrt{\frac{2m(n-m)}{n}} \sqrt{n} = n\sqrt{\frac{2}{m(n-m)}}.$$

Therefore, we obtain:

$$\mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GT}}) \le \Pi^2_{\mathrm{Norm}} B R_{\mathrm{out}} (1 + R_1 R_2) n \sqrt{\frac{2}{m(n-m)}} \left[1 + R_0 \widetilde{R}\left(\sqrt{\eta^{(1)}_{\mathrm{MHT},\tau}} + \mathcal{O}(\tau)\right)\right]. \tag{50}$$

Let $\ell \circ \mathcal{F}_{\mathrm{GT}}$ denote the corresponding loss-composed hypothesis class. Since $\ell(\cdot, y)$ is L-Lipschitz, the contraction inequality gives:

$$\mathfrak{R}_{m,n}(\ell \circ \mathcal{F}_{\mathrm{GT}}) \le L\, \mathfrak{R}_{m,n}(\mathcal{F}_{\mathrm{GT}}). \tag{51}$$

Finally, combining Eq. (50) and Eq. (51) with Eq. (2), we conclude that for $0 < \delta < 1$, with probability at least $1-\delta$ for all $f \in \mathcal{F}_{\mathrm{GT}}$ the following holds:

$$\epsilon(f) \leq \hat{\epsilon}(f) + \Pi_{\mathrm{Norm}}^2 LBR_{\mathrm{out}}(1+R_1R_2)n\sqrt{\frac{2}{m(n-m)}}\left[1 + R_0\widetilde{R}\left(\sqrt{\eta_{\mathrm{MHT},\tau}^{(1)}} + \mathcal{O}(\tau)\right)\right] + c_1\frac{n\sqrt{\min\{m,n-m\}}}{m(n-m)} + c_2\sqrt{\frac{n}{m(n-m)}\log\frac{1}{\delta}}.$$

This completes the proof. □

While the formal derivation above is presented for a single-layer GT for mathematical tractability, the same bounding strategy can in principle be extended to multi-layer architectures by applying the argument recursively across layers.

Theorem 2. *For each layer k, assume that the propagation weights are nonnegative and that there exists $\omega_{\min}^{(k)} > 0$ such that $\omega_{vu}^{(k)} \geq \omega_{\min}^{(k)}$ for every active propagation relation with $\omega_{vu}^{(k)} \neq 0$. Let $p_j^{(k)} = \frac{\mathrm{vol}^{(k)}(C_j)}{\mathrm{vol}^{(k)}(V)}$ and $p_{\max}^{(k)} = \max_{1\leq j\leq c} p_j^{(k)}$, with $0 < p_j^{(k)} \leq p_{\max}^{(k)} < 1$, and define $\xi^{(k)} = \frac{\mathrm{vol}^{(k)}(V)}{\omega_{\min}^{(k)}\log\left(1/p_{\max}^{(k)}\right)}$. Then, the following upper bound holds for the cross-class structural complexity:*

$$\eta_{\mathrm{cross}}^{(k)} \leq \xi^{(k)}\left[-\sum_{j=1}^{c}\frac{g_{C_j}^{(k)}}{\mathrm{vol}^{(k)}(V)}\log\frac{\mathrm{vol}^{(k)}(C_j)}{\mathrm{vol}^{(k)}(V)}\right].$$

Proof. Let $G_{\mathrm{cross}}^{(k)} = \sum_{j=1}^{c} g_{C_j}^{(k)}$. By the definition of $g_{C_j}^{(k)}$, we have:

$$\begin{aligned} G_{\mathrm{cross}}^{(k)} &= \sum_{j=1}^{c}\sum_{v\in V} P_{vj}\sum_{u\in\mathcal{N}(v)}\omega_{vu}^{(k)}(1-P_{uj}) \\ &= \sum_{v,u\in V}\omega_{vu}^{(k)}\sum_{j=1}^{c}P_{vj}(1-P_{uj}) \\ &= \sum_{v,u\in V}\omega_{vu}^{(k)}\left(1-P_v^\top P_u\right), \end{aligned}$$

where the last equality follows from $\sum_{j=1}^{c} P_{vj} = 1$. Since $P_{vj}, P_{uj} \in [0,1]$ and $\sum_{j=1}^{c} P_{vj} = 1$, we have:

$$P_v^\top P_u = \sum_{j=1}^{c} P_{vj}P_{uj} \leq \sum_{j=1}^{c} P_{vj} = 1.$$

Therefore, we have $0 \leq 1 - P_v^\top P_u \leq 1$. Moreover, by the minimum-weight assumption, it follows that:

$$\omega_{vu}^{(k)} \geq \omega_{\min}^{(k)}\mathbf{1}\left(\omega_{vu}^{(k)} \neq 0\right).$$

Multiplying both sides by the nonnegative quantity $1 - P_v^\top P_u$ and summing over all (v,u) gives:

$$\begin{aligned} G_{\mathrm{cross}}^{(k)} &= \sum_{v,u\in V}\omega_{vu}^{(k)}\left(1-P_v^\top P_u\right) \\ &\geq \omega_{\min}^{(k)}\sum_{v,u\in V}\mathbf{1}\left(\omega_{vu}^{(k)} \neq 0\right)\left(1-P_v^\top P_u\right) \\ &= \omega_{\min}^{(k)}\eta_{\mathrm{cross}}^{(k)}, \end{aligned}$$

which immediately implies that:

$$\eta_{\mathrm{cross}}^{(k)} \leq \frac{G_{\mathrm{cross}}^{(k)}}{\omega_{\min}^{(k)}}. \tag{52}$$

Algorithm 1: Training algorithm with SER

Input: A graph $G = (V, E)$ with node feature matrix X and adjacency matrix A, ground-truth labels $Y_{\mathcal{S}}$ for labeled nodes, a randomly initialized GNN model f_θ, regularization coefficient λ

Output: Optimized model parameters θ

while *not converged* **do**
- $H = f_\theta(X, A)$
- Construct class assignment matrix P
- Compute aggregation weights $\Omega^{(k)} = (\omega_{vu}^{(k)})$ for each layer
- Compute class-level structural entropy $\mathcal{L}_{\mathrm{SE}}^{(k)}$ through Eq.(7)
- Compute total loss $\mathcal{L}_{\mathrm{total}}$ through Eq.(8)
- Compute gradients of $\mathcal{L}_{\mathrm{total}}$ with respect to θ
- Update model parameters θ using gradient descent

end

Next, using $p_j^{(k)} = \mathrm{vol}^{(k)}(C_j)/\mathrm{vol}^{(k)}(V)$, we have:

$$-\sum_{j=1}^{c} \frac{g_{C_j}^{(k)}}{\mathrm{vol}^{(k)}(V)} \log \frac{\mathrm{vol}^{(k)}(C_j)}{\mathrm{vol}^{(k)}(V)} = \sum_{j=1}^{c} \frac{g_{C_j}^{(k)}}{\mathrm{vol}^{(k)}(V)} \log\left(\frac{1}{p_j^{(k)}}\right). \tag{53}$$

Since $p_j^{(k)} \le p_{\max}^{(k)} < 1$, it holds that $\log(1/p_j^{(k)}) \ge \log(1/p_{\max}^{(k)}) > 0$. Substituting this lower bound into Eq. (53) yields:

$$\begin{aligned}
&-\sum_{j=1}^{c} \frac{g_{C_j}^{(k)}}{\mathrm{vol}^{(k)}(V)} \log \frac{\mathrm{vol}^{(k)}(C_j)}{\mathrm{vol}^{(k)}(V)} \\
&\ge \frac{\log(1/p_{\max}^{(k)})}{\mathrm{vol}^{(k)}(V)} \sum_{j=1}^{c} g_{C_j}^{(k)} \\
&= \frac{\log(1/p_{\max}^{(k)})}{\mathrm{vol}^{(k)}(V)} G_{\mathrm{cross}}^{(k)}.
\end{aligned}$$

Rearranging the inequality, we obtain:

$$G_{\mathrm{cross}}^{(k)} \le \frac{\mathrm{vol}^{(k)}(V)}{\log(1/p_{\max}^{(k)})} \left[-\sum_{j=1}^{c} \frac{g_{C_j}^{(k)}}{\mathrm{vol}^{(k)}(V)} \log \frac{\mathrm{vol}^{(k)}(C_j)}{\mathrm{vol}^{(k)}(V)} \right]. \tag{54}$$

Combining Eq. (54) with Eq. (52) yields:

$$\eta_{\mathrm{cross}}^{(k)} \le \xi^{(k)} \left[-\sum_{j=1}^{c} \frac{g_{C_j}^{(k)}}{\mathrm{vol}^{(k)}(V)} \log \frac{\mathrm{vol}^{(k)}(C_j)}{\mathrm{vol}^{(k)}(V)} \right].$$

□

E Training Algorithm of Structural Entropy Regularization

The overall optimization procedure of SER is summarized in Algorithm 1.

F DATASETS AND EXPERIMENTAL DETAILS

F.1 Datasets

We evaluate our method on seven widely used benchmark datasets for semi-supervised node classification with random splits. These include three citation networks (Sen et al., 2008) (Cora, CiteSeer,

Table 4: Statistics of the datasets.

Dataset	# Nodes	# Edges	# Classes	# Features
Cora	2,708	5,429	7	1,433
CiteSeer	3,327	4,732	6	3,703
PubMed	19,717	44,338	3	500
CS	18,333	81,894	15	6,805
Physics	34,493	247,962	5	8,415
Computers	13,381	245,778	10	767
Photo	7,487	119,043	8	745

and PubMed), two co-authorship networks (Shchur et al., 2018) (CS and Physics), and two Amazon co-purchase networks (Shchur et al., 2018) (Computers and Photo). The statistical summary of these datasets, including the numbers of nodes, edges, classes, and features, is reported in Table 4, followed by detailed information about each dataset.

- **Citation networks.** Cora, CiteSeer, and PubMed are citation graphs in which nodes represent scientific publications and edges denote citation relationships. Each node is associated with a sparse bag-of-words feature vector constructed from the document content, and class labels correspond to research topics.
- **Co-author networks.** CS and Physics are co-authorship graphs constructed from the Microsoft Academic database. Nodes represent authors, and edges indicate co-authorship relations. Node features are derived from the keywords of published papers, and labels correspond to the primary research area of each author.
- **Amazon co-purchase networks.** Computers and Photo are segments of the Amazon co-purchase network. Nodes correspond to products, and edges connect items that are frequently purchased together. Node features are bag-of-words representations extracted from product reviews, and labels indicate product categories.

All datasets are treated as undirected graphs. For the random-split experiments, following the data-splitting protocol of Shchur et al. (2018); Yang et al. (2021), we use 20 nodes per class for training, 30 nodes per class for validation, and all remaining nodes for testing. Each configuration is evaluated on five randomly generated splits with 10 random seeds per split. We report the mean classification accuracy and standard deviation over the resulting 50 runs.

F.2 Experimental Setup

Backbone Architectures. We evaluate the proposed structure-aware regularization on three representative message-passing architectures: GCN, GAT, and GT. For each backbone, we compare the vanilla model (without structural regularization) and the regularized version obtained by adding the proposed structural entropy term. We adopt a standard two-layer GCN with symmetric normalization, a two-layer GAT with eight attention heads in the first layer and a single head in the output layer, and a two-layer GT with four attention heads per layer. GCN employs ReLU activation, GAT uses ELU activation, and GT applies ReLU within its feed-forward sublayers. Dropout is applied in all three architectures according to their respective backbone configurations.

Optimization and Model Selection. For fair comparison, the same backbone configuration is used for the vanilla and regularized variants. All models are optimized using Adam with a learning rate of 0.01 and a weight decay of 5×10^{-4}. We apply early stopping based on validation accuracy with a patience of 200 epochs. The regularization coefficient λ is selected from the range $[0, 1]$ based on validation performance. Validation performance is used for hyperparameter selection, early stopping, and checkpoint selection, whereas test performance is never used for training or model selection.

SER Implementation. Ground-truth labels used in the training objective are restricted to the training nodes. In the class-assignment matrix used by SER, training nodes use their ground-truth one-hot labels, whereas all remaining nodes use model posteriors. Since the class assignments of non-training nodes are derived from the current model posteriors, they may be unreliable during the early

stage of training. To reduce the influence of such potentially unreliable assignments, we linearly warm up the SER coefficient over the first 10 epochs. Specifically, the effective regularization coefficient is gradually increased from 0 to λ, after which it remains fixed at λ. This allows the structural regularization to be introduced progressively as the model predictions become more informative.

For the GCN backbone, whose propagation structure is fixed by the input graph, we associate each candidate message-passing relation (v, u) with a learnable edge logit s_{vu} and construct a continuous edge mask as $M_{vu} = \mathrm{sigmoid}\,(s_{vu}/\rho_m)$, where ρ_m denotes the mask temperature. The edge logits are initialized to zero, such that all candidate relations start from the same mask value of 0.5. We use the standard sigmoid parameterization with $\rho_m = 1$ and keep this temperature fixed across all experiments. The mask is optimized jointly with the model parameters and is used to reweight candidate message-passing relations during training.

Compared Regularization Methods. We compare SER with several established regularization methods for GNNs. Detailed information about the compared methods is provided as follows:

- Laplacian regularization (LapR) (Ando & Zhang, 2006) enforces spatial smoothness by penalizing the distance between representations of adjacent nodes to prevent overfitting and improve semi-supervised generalization.
- Confidence penalty (CP) (Pereyra et al., 2017) adds the negative entropy of the network outputs to the classification loss to penalize overconfident predictions and improve generalization.
- P-reg (Yang et al., 2021) imposes a variation of Laplacian-based regularization that possesses the equivalent power of an infinite-depth GCN to capture long-range structural information.
- NASA (Bo et al., 2022) replaces immediate neighbors with remote neighbors to generate graph augmentations that balance consistency and diversity via neighbor-constrained regularization.
- R-reg (Ji et al., 2025) introduces the concept of graph distributional signals to encode smoothness and non-uniformity of model outputs as a plug-in regularization term.

In addition to the established regularization methods above, we construct an L_0-regularized baseline (L_0-reg) to examine whether indiscriminately reducing message-passing relations is sufficient for improving generalization. We implement this baseline using the Hard-Concrete relaxation, where each candidate propagation relation is associated with a learnable stochastic gate. The normalized expected proportion of active gates is optimized jointly with the classification loss. Thus, this baseline provides a differentiable approximation to the normalized expected active-relation count rather than directly optimizing the exact support size $\|\Omega\|_0$. Although the gates are learned jointly with the prediction objective, its sparsity penalty does not distinguish between class-consistent and cross-class propagation relations, in contrast to SER.

G Additional Analysis of Regularization Strategies

To provide a more comprehensive analysis of different regularization strategies, we further discuss the performance characteristics of existing methods. As shown in Table 1, the effectiveness of different regularization strategies varies significantly across datasets and architectures, reflecting their distinct mechanisms and inductive biases.

Methods such as CP and P-reg provide moderate improvements on several datasets but show limited robustness, especially on more complex graphs. Their inconsistent performance across different architectures suggests that prediction-level or fixed propagation constraints are insufficient to manage the dynamic structural complexity of GNNs. The LapR achieves consistent gains on relatively homophilous datasets by enforcing local smoothness, but its performance tends to degrade on denser graphs due to over-smoothing effects. NASA, which leverages neighborhood-based augmentation, exhibits competitive performance in some cases but suffers from instability, particularly on transformer-based architectures such as GT. The recently proposed R-reg, which models graph distributional signals, shows competitive results but exhibits instability on attention-based models like

GAT and GT. This suggests that enforcing distributional smoothness can conflict with the anisotropic nature of attention mechanisms, leading to inconsistent generalization.

In contrast, the proposed SER consistently outperforms both the vanilla baselines and competing regularization methods across most datasets and model architectures. The improvements are particularly pronounced on datasets with denser connectivity, such as Physics and Computers, where excessive cross-class message passing is more likely to occur. From a model-specific perspective, for GCN, SER achieves the best or near-best performance across all datasets, demonstrating that explicitly controlling effective connectivity effectively mitigates over-smoothing and structural overfitting. For GAT, the improvements indicate that suppressing unnecessary cross-class attention weights enhances the discriminative power of learned attention coefficients. For GT, despite its global self-attention mechanism, SER still yields consistent gains over both the baseline and other regularization methods, confirming that even globally connected architectures benefit from explicitly regulating effective edge participation.