DRGFM: DYNAMIC RIEMANNIAN GRAPH FOUNDATION MODELS VIA LEARNABLE CURVATURE SPECTRUM

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph foundation models (GFMs) require structural knowledge that transfers across domains and remains useful as tasks evolve. Yet implicit structural encodings make it difficult to specify what should transfer and what should be preserved. Curvature and spectral analysis offer complementary descriptions of local geometry and multi-scale propagation, but predefined geometric operators cannot themselves adapt to learning objectives across domains. We propose DRGFM, a Dynamic Riemannian Graph Foundation Model that models structural knowledge through a learnable curvature spectrum. A learned curvature-weighted Laplacian provides graph-dependent spectral coordinates and filters, coupling mixed-curvature representations with spectral attention to accommodate geometric heterogeneity. This formulation further supports structural knowledge evolution: during continual learning, importance-weighted regularization constrains spectral-coordinate drift on bounded anchor subgraphs in frozen reference bases, while the current curvature estimator and backbone remain adaptable. A conditional local analysis relates this displacement to historical-loss drift. Experiments on few-shot classification, zero-shot link prediction, and temporal continual learning evaluate transfer and retention, while matched ablations support the contribution of learnable curvature beyond a fixed prior. Together, these results support a unified approach to learning transferable structural representations and preserving them during subsequent adaptation.

1 INTRODUCTION

Graph foundation models (GFMs) seek transferable representations across graph domains and tasks (Liu et al., 2025; 2023). Beyond node attributes, this requires structural knowledge: how connectivity organizes representations and governs information propagation. Graph encoders capture structure through message passing, positional encodings, and global attention (Ying et al., 2021; Rampášek et al., 2022; Kreuzer et al., 2021; Wu et al., 2021; Shirzad et al., 2023). These mechanisms can produce effective representations, but they leave the structural knowledge itself implicit in parameters and task-dependent signals. Consequently, it is difficult to identify what structural pattern has been transferred, or to decide which part should be preserved when the target graph has a different topology.

Graph geometry offers a way to make this knowledge explicit. Curvature summarizes local structural regimes: highly clustered regions and tree-like bottlenecks induce different neighborhood growth and therefore different propagation behavior (Ollivier, 2009; Forman, 2003; Ni et al., 2019). Spectral methods provide a complementary global coordinate system, in which graph-dependent modes and filter responses describe how information varies over the topology (Bruna et al., 2013; Defferrard et al., 2016; Saad, 2011). Yet these directions have not yielded a unified structural knowledge space for GFMs: curvature is often used as a fixed descriptor, spectral operators are commonly defined from a fixed graph geometry, and the resulting structural signal is not explicitly designed for cross-domain transfer.

This limitation is critical for GFMs. Source and target graphs can differ in density, community organization, and bottleneck structure (Wang et al., 2025; Zhao et al., 2024a; Sun et al., 2025), so

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.

a curvature signal or spectral operator fixed before adaptation may impose a structural bias that is appropriate for the source domain but unsuitable for the target. A useful structural representation should therefore be (i) expressive enough to describe heterogeneous graph geometry and its multi-scale propagation effects, (ii) adaptable when the graph domain changes, and (iii) equipped with coordinates in which structural knowledge can be compared and preserved. These desiderata motivate our central claim: a GFM should learn structural knowledge as a graph-dependent curvature spectrum—specifically, the spectrum of a curvature-weighted Laplacian—rather than treating curvature as a preprocessing feature or the spectrum as a fixed auxiliary object.

We introduce DRGFM, a Dynamic Riemannian Graph Foundation Model. A geometric prior and a learned residual define a curvature-weighted Laplacian whose spectral coordinates represent graph-dependent knowledge, and whose filters guide Transformer attention through Chebyshev spectral attention (Defferrard et al., 2016; Kreuzer et al., 2021). Here, dynamic denotes end-to-end adaptation of the curvature estimator and Laplacian during pre-training and continual backbone updates. Curvature-conditioned hyperbolic, Euclidean, and spherical components accommodate heterogeneous geometry without imposing a single constant-curvature space (Gu et al., 2018; Skopek et al., 2019; Chami et al., 2019; Bachmann et al., 2020). Together, this representation and spectral attention connect local geometric variation to multi-scale structural processing.

This representation also gives a principled account of structural knowledge evolution. As tasks arrive, adaptation must accommodate new patterns without erasing historical behavior (Kirkpatrick et al., 2017; Pan et al., 2020; Tian et al., 2024). DRGFM stores bounded anchor subgraphs and penalizes importance-weighted displacement of their spectral coordinates in frozen reference bases, while the current curvature estimator and Laplacian remain dynamic. This provides a stable functional reference while permitting weaker constraints on less important modes. Continual learning therefore preserves the structural representation learned during pre-training, rather than imposing uniform parameter rigidity. A conditional local analysis further relates spectral displacement on anchors to historical-loss drift under local spectral perturbation assumptions.

We evaluate transfer and retention through few-shot classification, zero-shot link prediction, temporal continual learning, and matched curvature ablations. Our contributions are:

- A curvature-spectral formulation of GFM structural knowledge, linking graph-dependent representation coordinates with learned propagation responses.
- DRGFM, a graph Transformer coupling a learnable curvature Laplacian with mixed-manifold representations and Chebyshev spectral attention.
- An importance-weighted anchor regularizer for structural knowledge evolution, supported by conditional local analysis and transfer, continual-learning, and ablation experiments.

2 Related Work

2.1 Graph foundation models and cross-domain transfer

Graph foundation models seek reusable node or graph representations through structural reconstruction, attribute masking, or contrastive pre-training (Yuan et al., 2026; Jiang et al., 2024; Wang et al., 2025; Zhao et al., 2024b;a; Sun et al., 2025; Yu et al., 2025; Hou et al., 2022; You et al., 2020; Veličković et al., 2018; Kipf & Welling, 2016b). A central design choice is how structural information is encoded: existing approaches typically distribute it across message-passing parameters, positional signals, or geometry-aware features (Zhang et al., 2024). Riemannian and product-manifold graph models introduce non-Euclidean representation spaces to reflect heterogeneous graph structure (Chami et al., 2019; Gu et al., 2018; Skopek et al., 2019), while curvature–spectral methods combine local curvature signals with graph spectral operators (Ye et al., 2019; Ni et al., 2019). These approaches motivate geometry as an inductive bias for transfer, but the curvature signal, spectral basis, or learned filter is commonly treated as fixed after pre-training (Grover et al., 2025). They therefore do not directly expose a graph-dependent functional coordinate system that can be reused for both cross-domain transfer and knowledge preservation. In contrast, DRGFM makes the curvature Laplacian and its spectral coordinates jointly useful across representation learning and continual adaptation.

2.2 CONTINUAL LEARNING ON GRAPHS

Regularization-based continual learning methods constrain parameter changes using estimates of parameter importance. Elastic Weight Consolidation (EWC) uses a diagonal Fisher approximation (Kirkpatrick et al., 2017), while Synaptic Intelligence and Memory Aware Synapses accumulate importance online or from parameter sensitivity (Zenke et al., 2017; Aljundi et al., 2018). These methods are simple and broadly applicable, but their coordinate system is the parameterization of a particular model: two parameter updates that have very different effects on graph structure can receive the same penalty, and parameters that compensate for one another are not represented naturally. Functional regularization methods such as FRCL and FROMP instead preserve predictions or function values on representative inputs (Titsias et al., 2019; Pan et al., 2020), and replay-based approaches retain examples or representations from earlier tasks (Chaudhry et al., 2019; Lopez-Paz & Ranzato, 2017). Graph-specific continual learning methods similarly combine replay, regularization, or domain-incremental training, but usually constrain node embeddings or model parameters without using a graph's geometric spectrum as a knowledge coordinate system (Tian et al., 2024; Sun et al., 2023). DRGFM is closest in spirit to functional regularization, while replacing generic representative outputs with curvature-spectral coordinates on bounded anchor subgraphs. This preserves important structural functions while leaving other spectral modes plastic.

2.3 RIEMANNIAN GEOMETRY AND GRAPH SPECTRAL OPERATORS

Ollivier–Ricci curvature describes graph edges through neighborhood transport (Ollivier, 2009), whereas Forman curvature provides a computationally efficient combinatorial approximation (Forman, 2003). Both characterize bottlenecks, community structure, and local geometric heterogeneity (Ni et al., 2019; Ye et al., 2019; Alon & Yahav, 2021). Product and mixed-curvature manifolds extend this perspective by representing different structural regimes in hyperbolic, Euclidean, or spherical components (Gu et al., 2018; Skopek et al., 2019; Chami et al., 2019). Separately, spectral graph methods define propagation through functions of a graph Laplacian (Skopek et al., 2019; Bruna et al., 2013; Kipf & Welling, 2016a; Levie et al., 2018). Chebyshev polynomial filters avoid explicit eigendecomposition and have linear dependence on the number of edges for a fixed polynomial order (Defferrard et al., 2016); Lanczos methods provide a complementary way to extract a small number of informative spectral modes (Saad, 2011). DRGFM combines these tools in a learnable curvature Laplacian: Forman curvature supplies a structural prior, a feature-dependent residual adapts it, and polynomial or Lanczos approximations make spectral processing feasible. The resulting spectrum is both a representation-learning interface and the coordinate system used by the continual-learning regularizer.

3 PROBLEM SETUP

3.1 GRAPH SEQUENCE AND LEARNING TRACKS

Let $\mathcal{G} = \{G_t\}_{t=1}^{T}$ be a sequence of attributed graphs, where $G_t = (V_t, E_t, X_t)$, $X_t \in \mathbb{R}^{|V_t| \times d_x}$, and E_t may change across time. Each graph is associated with a task $\mathcal{T}_t$ and a task-specific label space. The model is written as

$$f_{\theta,t}(G_t) = \psi_t(g_\theta(G_t; \phi)), \tag{1}$$

where g_θ is the shared Riemannian graph Transformer, ϕ parameterizes the curvature estimator, and ψ_t is a task head. The shared parameters are pre-trained on a collection of source graphs and subsequently adapted to target tasks.

We distinguish two evaluation tracks. In cross-domain transfer, a target graph or task is observed independently and the model is adapted with its support labels, without requiring performance preservation on previous targets. In continual learning, tasks arrive sequentially as $\mathcal{T}_1 \to \cdots \to \mathcal{T}_T$; after learning task t, the model should fit $\mathcal{T}_t$ while retaining performance on earlier tasks. The corresponding objectives are, respectively, target-task risk minimization and sequential risk minimization with a functional stability constraint.

Information visibility. Pre-training uses source-graph structure, node features, and explicitly declared source-domain supervision. Supervised few-shot adaptation uses support labels. Zero-shot link prediction uses no target edge labels or parameter updates; held-out test edges do not participate in message passing, curvature estimation, or negative sampling.

3.2 CURVATURE LAPLACIAN

For an undirected graph, let $\bar{k}_{uv}$ denote the estimated curvature of edge (u, v). We define curvature-weighted adjacency and a symmetric normalized curvature Laplacian by

$$W_\kappa[u, v] = \exp(\beta \bar{k}_{uv}) \mathbb{1}[(u, v) \in E], \qquad L_\kappa = I - D_\kappa^{-1/2} W_\kappa D_\kappa^{-1/2}, \tag{2}$$

where $D_\kappa[u, u] = \sum_v W_\kappa[u, v]$ and β is a learnable temperature. Since W_κ is symmetric and nonnegative, L_κ is positive semidefinite and its spectrum lies in $[0, 2]$. We write

$$L_\kappa = U_\kappa \Lambda_\kappa U_\kappa^\top, \tag{3}$$

where U_κ is an orthonormal eigenbasis and $\Lambda_\kappa = \mathrm{diag}(\lambda_1, \ldots, \lambda_{|V|})$. For a normalized Laplacian, Chebyshev filters use the shifted operator $\widetilde{L}_\kappa = L_\kappa - I$, whose eigenvalues lie in $[-1, 1]$. We do not require an explicit full eigendecomposition during training.

3.3 SPECTRAL COORDINATES AND ASSUMPTIONS

For a node representation $H \in \mathbb{R}^{|V| \times d}$, its curvature-spectral coordinates are

$$\widehat{H} = U_\kappa^\top H. \tag{4}$$

A spectral filter in layer ℓ is represented by a polynomial response

$$g_\theta^{(\ell)}(\lambda) = \sum_{k=0}^{K} \alpha_k^{(\ell)} T_k(\widetilde{\lambda}), \tag{5}$$

where T_k is the k-th Chebyshev polynomial. In practice, we retain $m \ll |V|$ modes using a Lanczos approximation when explicit spectral coordinates are needed. Let $\rho_i \geq 0$ denote the importance of mode i, estimated from the energy or sensitivity of anchor spectral coordinates. The weighted spectral displacement used later is $\sum_i \rho_i \|\widehat{H}_i - \widehat{H}_i^*\|_2^2$.

Our formulation makes three assumptions explicit. First, it uses a transductive encoder: all nodes of a target graph, including unlabeled test nodes, may participate in message passing and normalization, although their labels are not used for optimization. Second, explicit spectral operations use a top-m approximation; the omitted tail is controlled by the retained-mode approximation and is not assumed to vanish. Third, continual learning stores a bounded set of anchor subgraphs, their node indices and edge structure, together with frozen spectral bases and snapshot coordinates; node features are recomputed during later regularization passes. Thus, the proposed functional regularization is a memory-based method rather than a strict no-replay setting. These assumptions delimit the claims made by the subsequent analysis.

4 METHOD

DRGFM is organized around a learnable curvature-spectral interface (Figure 1). Given a graph and node features, the curvature module combines a Forman prior with a feature-dependent residual to construct a weighted Laplacian and its spectral coordinates. These coordinates condition a mixed-curvature representation space and provide independent structural probes for the per-head Chebyshev attention bias. The resulting backbone is trained with complementary reconstruction, link-prediction, and spectral-consistency objectives; for sequential tasks, frozen anchor coordinates impose stability only on important modes while leaving the remaining spectrum adaptable. Sections 4.1–4.6 describe the curvature, representation, modulation, attention, pre-training, and anchor-regularization components.

4.1 CURVATURE ESTIMATION AND LEARNABLE LAPLACIAN

We predict edge curvature by combining a differentiable geometric prior with a learned residual. For edge (u, v), the Forman curvature k_{uv}^F is computed directly from the degree and triangles (Forman, 2003):

$$k_{uv}^F = 4 - d_u - d_v + 3|\triangle_{uv}|, \tag{6}$$

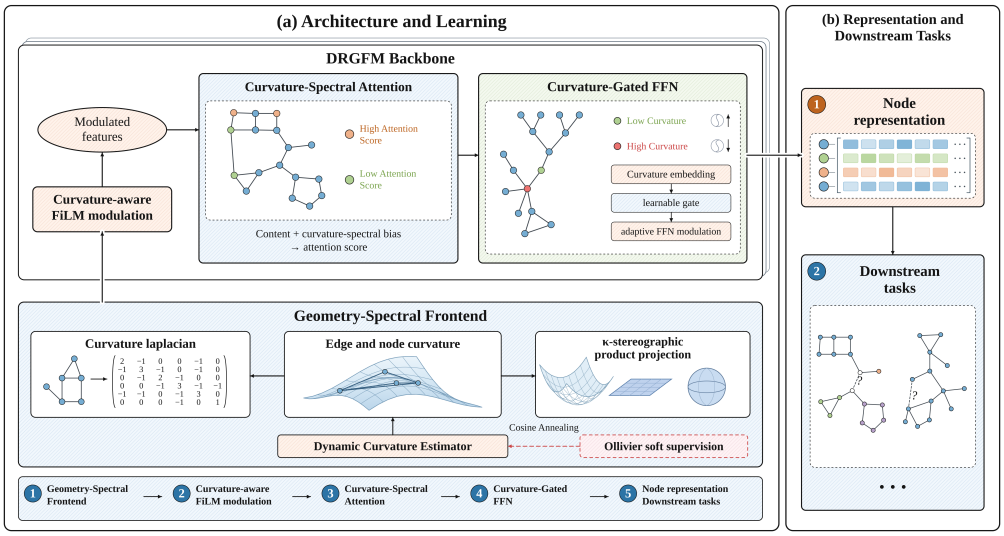

Figure 1: Overview of DRGFM. The learnable curvature spectrum provides a shared interface for structural representation learning, cross-domain transfer, and continual knowledge evolution.

where $|\triangle_{uv}|$ denotes the number of triangles containing (u, v). This provides a combinatorial approximation to Ricci curvature that is computationally efficient and preserves the sign of local geometry (negative for tree-like bridges, positive for high-clustering cores). A learned residual network Δ_ϕ then adapts this prior using node features and local aggregates:

$$\bar{k}_{uv} = k^F_{uv} + s_\phi \Delta_\phi\big([x_u; x_v; \overline{x}_{N(u)}; \overline{x}_{N(v)}]\big), \tag{7}$$

where $[\cdot;\cdot]$ denotes concatenation, $\overline{x}_{N(v)}$ is a mean-pooled neighborhood feature, and s_ϕ is a learnable scale initialized near zero. During early training, an Ollivier–Ricci warm-up loss (Ollivier, 2009) encourages $\bar{k}_{uv}$ to remain close to a cached approximate Ollivier curvature:

$$\mathcal{L}_{\text{warm}}(t) = \alpha(t) \sum_{(u,v)\in E} (\bar{k}_{uv} - \hat{\kappa}^{\text{Oll}}_{uv})^2, \quad \alpha(t) \to 0 \text{ (cosine)}. \tag{8}$$

The curvature-weighted adjacency and normalized Laplacian are then

$$W_\kappa[u, v] = \exp\big(\text{clip}(\beta \bar{k}_{uv}, -15, 15)\big)\mathbb{1}[(u, v) \in E], \quad L_\kappa = I - D_\kappa^{-1/2} W_\kappa D_\kappa^{-1/2}, \tag{9}$$

where $D_\kappa[u, u] = \sum_v W_\kappa[u, v]$ and β is a learnable temperature. The clipping prevents numerical overflow; the temperature is initialized at $\beta = 1$ and optimized end to end. Node curvature is defined as the average edge curvature of incident edges: $\kappa_v = |N(v)|^{-1} \sum_{u\in N(v)} \bar{k}_{uv}$.

4.2 Mixed-curvature representation space

Node representations reside in a product manifold $\mathcal{M} = \mathcal{M}^{d_h}_{\kappa_h} \times \mathbb{R}^{d_e} \times \mathcal{M}^{d_s}_{\kappa_s}$, where $\mathcal{M}^d_\kappa$ is a d-dimensional manifold of constant curvature κ under the κ-stereographic model (Skopek et al., 2019; Gu et al., 2018; Chami et al., 2019; Bachmann et al., 2020). Curvature-driven component routing assigns a mixture weight to each component via

$$(w^h_v, w^e_v, w^s_v) = \text{softmax}(-\kappa_v, 0, +\kappa_v). \tag{10}$$

When $\kappa_v < 0$ (bridge), the hyperbolic component is emphasized; when $\kappa_v > 0$ (core), the spherical component is emphasized; near zero, all components contribute equally. All linear operations (attention query/key/value projections, layer normalization, and feed-forward networks) are performed in the shared tangent space at the origin. Node representations are mapped to the tangent space via $\log_0$, processed through standard Euclidean operations with component-wise scaling by routing weights, and mapped back to the manifold via $\exp_0$. This ensures differentiability and avoids separate implementations for each manifold type.

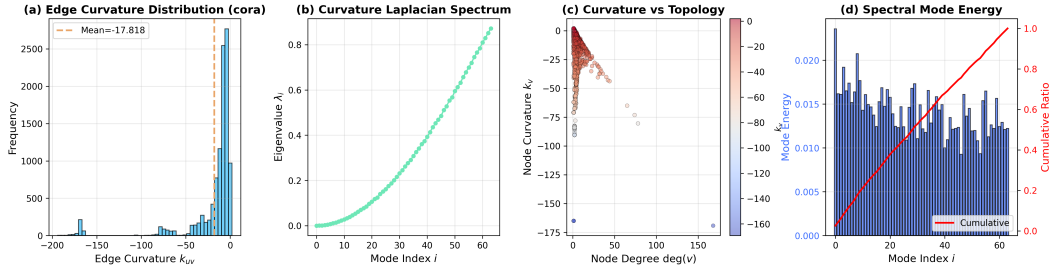

Figure 2: Curvature-spectrum analysis on Cora. The four views summarize the structural signal used by DRGFM: edge-curvature heterogeneity, the spectrum of the curvature Laplacian, the relationship between node curvature and topological degree, and the energy carried by spectral modes.

4.3 Curvature-conditioned modulation and gating

Input node features are modulated by curvature-dependent affine parameters. Let $\gamma(\kappa_v)$ and $\delta(\kappa_v)$ be outputs of a learned sinusoidal embedding network. The initial node representation is

$$h_v^{(0)} = \gamma(\kappa_v) \odot (W_x x_v + b_x) + \delta(\kappa_v), \tag{11}$$

which allows the encoder to adjust feature scales and shifts based on local geometry. Similarly, the feed-forward block uses curvature-dependent gating:

$$\text{gate}_v = \sigma\big(W_g[\rho(\kappa_v); h_v]\big), \quad h'_v = h_v + \text{FFN}(h_v) \odot \text{gate}_v, \tag{12}$$

where $\rho(\kappa_v)$ is a learned embedding of κ_v and σ is the sigmoid. The design intent is that positive-curvature nodes (susceptible to over-smoothing) receive smaller gates, while negative-curvature nodes (bottlenecks) receive larger gates (Li et al., 2018; Oono & Suzuki, 2020; Alon & Yahav, 2021). Whether this behavior emerges in practice is evaluated in the ablations.

4.4 Curvature-spectral attention with independent structural probes

Standard graph attention computes pairwise scores from query-key products (Veličković et al., 2017). We augment this with a bounded curvature-spectral bias that reflects multi-scale structural patterns. Rather than filtering a single content-driven signal, we use multiple independent structural probes: a constant channel, node curvature, weighted degree, and local geometric statistics. Let $S \in \mathbb{R}^{|V| \times R}$ denote the probe matrix with R channels. For each attention head h and probe channel r, a Chebyshev polynomial filters the corresponding structural signal:

$$F_{h,r} = \sum_{k=0}^{K} \alpha_{h,r,k} T_k(\widetilde{L}_\kappa)\, S_r, \tag{13}$$

where $\alpha_{h,r,k}$ are learnable per-head per-channel coefficients. The filtered signals are row-normalized and combined into pairwise cosine similarities:

$$\bar{F}_{h,v} = F_{h,v} \,/\, \max(\|F_{h,v}\|_2, \epsilon), \quad B_{uv}^{(h)} = b_h \cdot \langle \bar{F}_{h,u}, \bar{F}_{h,v} \rangle, \quad |b_h| \le b_{\max}, \tag{14}$$

where b_h is a per-head learnable scale with bounded magnitude. The attention score combines content and structure:

$$A_{uv}^{(h)} = \frac{\langle \log_0(h_u) W_Q^{(h)}, \log_0(h_v) W_K^{(h)} \rangle}{\sqrt{d_h}} + B_{uv}^{(h)}. \tag{15}$$

This pairwise formulation ensures that structural bias is not canceled by row-wise softmax normalization. Independent probe channels allow different heads to emphasize different structural scales (e.g., local clustering vs. global topology).

4.5 Pre-training objectives

The model is pre-trained on source graphs with

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{rec}} + \lambda_F \mathcal{L}_F + \lambda_{\text{link}} \mathcal{L}_{\text{link}} + \lambda_{\text{spec}} \mathcal{L}_{\text{spec}} + \lambda_{\text{node}} \mathcal{L}_{\text{node}} + \lambda_{\text{proto}} \mathcal{L}_{\text{proto}} + \alpha_{\text{warm}}(t) \mathcal{L}_O. \tag{16}$$

Here, the first five terms provide masked reconstruction, Forman curvature anchoring, link prediction, spectral consistency, and node-level invariance. $\mathcal{L}_{\text{proto}}$ is optional source-domain class-text supervision, and $\mathcal{L}_O$ is the decaying Ollivier warm-up.

4.6 ANCHOR SPECTRAL-COORDINATE FUNCTIONAL REGULARIZATION

During continual learning, we store a bounded set of anchor subgraphs $\{G_a^s\}$ from earlier tasks, each with a few hundred nodes. For each anchor, the curvature Laplacian L_κ^s and its top-m eigenmodes (U_a^s, Λ_a^s) are cached at the task snapshot θ_s^*. Spectral-mode importance is estimated as the energy of the anchor's spectral coordinates:

$$e_{s,i} = \|\widehat{H}_{s,i}^*\|_2^2 \,/\, d, \quad \bar{\rho}_{s,i} = e_{s,i} \,/\, \max\Big(\sum_j e_{s,j}, \epsilon\Big). \tag{17}$$

At training step t, the model's current spectral coordinates on anchor s are $\widehat{H}_s(\theta) = (U_a^s)^\top H_\theta(G_a^s)$, where H_θ is the embedding produced by the backbone. The functional regularization penalizes drift on important modes:

$$\mathcal{L}_t(\theta) = \mathcal{L}_{\text{task},t}(\theta) + \frac{\lambda_{\text{spec}}}{2|\mathcal{A}|} \sum_{s\in\mathcal{A}} \sum_i \bar{\rho}_{s,i} \, \frac{\|\widehat{H}_{s,i}(\theta) - \widehat{H}_{s,i}^*\|_2^2}{d}, \tag{18}$$

where $\mathcal{A}$ is the set of anchors and d normalizes across feature dimensions. This constrains the model to preserve the spectral representation on important structural frequencies while allowing adaptation on less important or newly emerging modes. Because $\widehat{H}$ is the output of the full backbone, gradients from the regularizer reach all layers, unlike a filter-response penalty that only affects the top attention-bias parameters.

5 ANALYSIS

Our regularizer admits a local functional interpretation. Fix an anchor graph, its cached spectral basis, and mode weights, and let $q(\theta) = \mathrm{vec}(Q(\theta))$ collect the anchor spectral coordinates produced by the current backbone. With $R = \mathrm{diag}(\rho) \otimes I_d$, the spectral penalty is

$$\Phi(\theta) = \frac{1}{2}\big(q(\theta) - q^*\big)^\top R\big(q(\theta) - q^*\big). \tag{19}$$

At a task snapshot where $q(\theta^*) = q^*$, let $J_q = \partial q/\partial\theta$. The local Hessian induced by the penalty is

$$\nabla^2\Phi(\theta^*) = J_q^\top R J_q. \tag{20}$$

Thus, the penalty couples parameter directions according to their effect on graph- and spectrum-dependent functions. Parameter changes that compensate in representation space can be penalized jointly, while directions with little effect on important anchor modes receive a weaker constraint.

Suppose the historical-task loss, expressed as a function of the anchor coordinates, is twice differentiable near q^*, has zero gradient at q^*, and satisfies $\nabla_q^2\mathcal{L}_s(q) \preceq \mu R$ in that neighborhood. Taylor's theorem gives

$$\mathcal{L}_s(q) - \mathcal{L}_s(q^*) \le \frac{\mu}{2}(q - q^*)^\top R(q - q^*). \tag{21}$$

This local bound connects historical-loss drift to the weighted spectral displacement minimized by our regularizer. It depends on the smoothness condition, frozen anchor basis, and retained modes; it is not an unconditional guarantee for an arbitrary nonlinear Transformer or task stream.

6 EXPERIMENTS

6.1 EXPERIMENTAL PROTOCOL

We evaluate two transfer tasks, a lightweight continual-learning study, and targeted ablations. The primary pool contains `amazon-computers`, `amazon-photo` (McAuley et al., 2015), `flickr` (Zeng et al., 2019), `wiki-cs` (Mernyei & Cangea, 2020), and `coauthor-cs` (Shchur et al., 2018); the four target graphs are unseen during pre-training. All methods use the same splits, support budget, episode protocol, and metrics. Baseline descriptions are provided in Appendix E.

For few-shot node classification, we use transductive episodic N-way K-shot with $N = \min(5, C)$ and $K \in \{1, 5\}$ (Snell et al., 2017; Finn et al., 2017). Each episode samples five classes when

Table 1: Few-shot node classification accuracy on unseen target graphs. Results are mean ± standard deviation over 100 episodes; bold and underline denote the best and second-best means, respectively.

	facebook		github		lastfm_asia		pubmed	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
SAMGPT	0.3759 ± 0.1023	<u>0.6111</u> ± 0.0676	0.5670 ± 0.1171	0.6283 ± 0.1049	0.1503 ± 0.0624	0.2956 ± 0.0758	0.4799 ± 0.1085	0.5823 ± 0.0585
RiemannGFM	0.3600 ± 0.0304	0.5600 ± 0.0075	0.6173 ± 0.1799	<u>0.7628</u> ± 0.0412	0.2526 ± 0.0926	0.4565 ± 0.0491	0.4166 ± 0.0723	0.5994 ± 0.1040
GCOPE	**0.4852** ± 0.0125	0.5922 ± 0.0252	0.6240 ± 0.0141	0.7344 ± 0.0241	<u>0.3208</u> ± 0.0230	**0.6698** ± 0.0067	0.3831 ± 0.0115	0.6230 ± 0.0087
GraphAny	0.4365 ± 0.0183	**0.6122** ± 0.0285	<u>0.6869</u> ± 0.0604	0.6992 ± 0.0507	0.2596 ± 0.0463	0.4269 ± 0.0455	<u>0.5570</u> ± 0.0201	**0.6817** ± 0.0124
MDGFM	0.3701 ± 0.0721	0.5545 ± 0.0690	0.5838 ± 0.1233	0.7064 ± 0.0941	0.1915 ± 0.0447	0.3249 ± 0.0446	0.4379 ± 0.0976	0.5955 ± 0.0689
RAGraph	0.2826 ± 0.0316	0.3033 ± 0.0608	0.4478 ± 0.1244	0.5916 ± 0.1349	0.1200 ± 0.0400	0.2500 ± 0.0311	0.3008 ± 0.0824	0.5200 ± 0.0932
GRAVER	0.3284 ± 0.0668	0.4509 ± 0.0696	0.5932 ± 0.1642	0.6774 ± 0.1138	0.2042 ± 0.0576	0.3657 ± 0.0642	0.4318 ± 0.1030	0.5421 ± 0.1069
QGT	0.2944 ± 0.0387	0.5000 ± 0.0492	0.6572 ± 0.1087	0.7041 ± 0.0790	0.1400 ± 0.0167	0.2800 ± 0.0247	0.4561 ± 0.0944	0.5686 ± 0.0674
DRGFM	<u>0.4671</u> ± 0.0976	0.5936 ± 0.0558	**0.7007** ± 0.1950	**0.7649** ± 0.1195	**0.3290** ± 0.0847	<u>0.6307</u> ± 0.0916	**0.5782** ± 0.0731	<u>0.6709</u> ± 0.0623

Table 2: Zero-shot link prediction on unseen target graphs under a strict held-out-edge protocol. We report ROC-AUC and AP; bold and underline denote the best and second-best means, respectively.

	facebook		github		lastfm_asia		pubmed	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP
SAMGPT	0.8120 ± 0.0014	0.8198 ± 0.0017	0.7000 ± 0.0022	0.7100 ± 0.0022	0.7500 ± 0.0081	0.7600 ± 0.0061	<u>0.9229</u> ± 0.0028	0.8942 ± 0.0046
RiemannGFM	0.9529 ± 0.0010	0.9500 ± 0.0012	0.8888 ± 0.0015	0.8796 ± 0.0018	0.9106 ± 0.0040	0.8918 ± 0.0045	**0.9343** ± 0.0030	<u>0.9216</u> ± 0.0035
GCOPE	0.9642 ± 0.0009	0.9620 ± 0.0008	0.8931 ± 0.0015	0.9121 ± 0.0013	0.9444 ± 0.0021	0.9445 ± 0.0021	0.9000 ± 0.0068	0.9200 ± 0.0085
GraphAny	0.8321 ± 0.0020	0.8156 ± 0.0056	0.6252 ± 0.0123	0.5819 ± 0.0122	0.7985 ± 0.0869	0.7619 ± 0.0913	0.7867 ± 0.0079	0.7771 ± 0.0080
MDGFM	0.9784 ± 0.0001	<u>0.9817</u> ± 0.0002	0.7838 ± 0.0005	0.7415 ± 0.0007	0.9357 ± 0.0007	0.9331 ± 0.0011	0.7487 ± 0.0008	0.7647 ± 0.0010
RAGraph	0.9252 ± 0.0003	0.9442 ± 0.0004	0.5683 ± 0.0010	0.5866 ± 0.0014	0.8580 ± 0.0048	0.8901 ± 0.0033	0.6500 ± 0.0044	0.6800 ± 0.0017
GRAVER	<u>0.9815</u> ± 0.0004	<u>0.9817</u> ± 0.0005	<u>0.9470</u> ± 0.0007	<u>0.9447</u> ± 0.0008	<u>0.9548</u> ± 0.0009	<u>0.9545</u> ± 0.0012	0.9183 ± 0.0030	0.9207 ± 0.0029
QGT	0.9360 ± 0.0006	0.9434 ± 0.0006	0.9070 ± 0.0008	0.9086 ± 0.0007	0.8111 ± 0.0041	0.8130 ± 0.0043	0.9123 ± 0.0024	0.9170 ± 0.0024
DRGFM	**0.9820** ± 0.0062	**0.9848** ± 0.0051	**0.9546** ± 0.0082	**0.9565** ± 0.0076	**0.9751** ± 0.0065	**0.9844** ± 0.0063	0.9217 ± 0.0039	**0.9226** ± 0.0043

$C > 5$, selects K support nodes per class, and uses the remaining nodes as queries. Support and query nodes share message passing, but only support labels are used. DRGFM freezes the backbone and trains an adapter plus linear probe; results average 100 episodes.

For zero-shot link prediction, the pretrained model is evaluated without labels or updates (Zhang & Chen, 2018). We hold out 15% of edges, exclude them from representation computation, curvature estimation, and negative sampling, and report ROC-AUC and AP over 10 negative-sampling runs (Mikolov et al., 2013). All results are produced by the code provided in the supplementary material.

6.2 Primary transfer results

Few-shot node classification. Table 1 reports accuracy on four unseen graphs. DRGFM leads on github (Rozemberczki et al., 2021) in both shot regimes and on lastfm_asia (Rozemberczki & Sarkar, 2020) and pubmed (Namata et al., 2012) in 1-shot, while other methods lead on several 5-shot entries. The variation across domains motivates the narrower claim that the curvature-spectral representation transfers under limited supervision, rather than uniform superiority.

Zero-shot link prediction. Table 2 evaluates the same graphs under held-out edges. DRGFM obtains the best AP on all four graphs and the best AUC on three; on pubmed it ranks third in AUC while retaining the best AP. These results measure transfer of link-relevant structural representations without access to test edges.

6.3 Continual learning

We use the Elliptic temporal graph in a five-task domain-incremental stream (Weber et al., 2019). Pre-training uses `wiki-cs`, `amazon-photo`, and `amazon-computers`; the label space is fixed while graph structure and transaction distributions evolve. After each task, DRGFM stores a 1024-node anchor, extracts 32 Lanczos modes, and regularizes later updates in the frozen basis. We report AA, BWT, Forgetting, and FWT (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019).

Table 3 compares parameter regularization, experience replay, output-space functional regularization, and DRGFM's spectral-domain functional regularization. DRGFM obtains the highest AA (0.939), positive BWT (0.005), and lowest Forgetting (0.015) among the listed methods. FROMP is the closest comparator in AA and forgetting, whereas MAS and SI obtain the strongest FWT values in this table. This pattern is consistent with the intended role of the spectral anchors: the main gain is preservation of historical performance, while forward transfer remains a separate property rather than an assumed consequence of reducing forgetting.

Table 3: Continual learning on the Elliptic temporal graph under the five-task domain-incremental stream. Higher is better for AA, BWT, and FWT, while lower is better for Forgetting.

Method	Category	AA↑	BWT↑	Forgetting↓	FWT↑
MAS	Parameter Regularization	0.878	-0.063	0.089	0.325
SI	Parameter Regularization	0.881	-0.060	0.089	**0.330**
EWC	Parameter Regularization	0.916	-0.028	0.071	0.270
ER-GNN	Experience Replay	0.912	-0.026	0.043	0.256
FROMP	Output-space Function Regularization	0.929	-0.008	0.034	0.321
Ours	Spectral-domain Function Regularization	**0.939**	**0.005**	**0.015**	0.286

Table 4: Ablation study on the curvature-spectral interface for node classification.

	facebook		github		lastfm_asia		pubmed	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
DRGFM	**0.4671** ± 0.0976	**0.5936** ± 0.0558	**0.7007** ± 0.1950	**0.7649** ± 0.1195	**0.3290** ± 0.0847	**0.6307** ± 0.0916	**0.5782** ± 0.0731	**0.6709** ± 0.0623
w/o learnable curvature	0.4050 ± 0.0860	0.5379 ± 0.0525	0.6517 ± 0.1750	0.7043 ± 0.1050	0.3218 ± 0.0850	0.5861 ± 0.1030	0.5137 ± 0.0760	0.6016 ± 0.0770
w/o curvature	0.3192 ± 0.0755	0.4544 ± 0.0496	0.5919 ± 0.1537	0.6134 ± 0.0925	0.3109 ± 0.0857	0.5034 ± 0.1148	0.3765 ± 0.0799	0.4400 ± 0.0931

Table 5: Ablation study on the curvature-spectral interface for link prediction.

	facebook		github		lastfm_asia		pubmed	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP
DRGFM	**0.9820** ± 0.0062	**0.9848** ± 0.0051	**0.9546** ± 0.0082	**0.9565** ± 0.0076	**0.9751** ± 0.0065	**0.9844** ± 0.0063	**0.9217** ± 0.0039	**0.9226** ± 0.0043
w/o learnable curvature	0.9315 ± 0.0077	0.9435 ± 0.0062	0.8407 ± 0.0068	0.8414 ± 0.0059	0.9152 ± 0.0066	0.9321 ± 0.0064	0.8876 ± 0.0055	0.8968 ± 0.0057
w/o curvature	0.8135 ± 0.0092	0.8373 ± 0.0072	0.4991 ± 0.0054	0.5137 ± 0.0043	0.7755 ± 0.0067	0.7975 ± 0.0064	0.8215 ± 0.0070	0.8421 ± 0.0071

6.4 Ablation study

We isolate the curvature-spectral interface while keeping the remaining experimental settings fixed. In the *w/o learnable curvature* condition, the Forman-plus-residual estimator is replaced by fixed Forman curvature; the curvature-weighted Laplacian and independent structural spectral channels are retained. In the *w/o curvature* condition, curvature is removed, the standard normalized Laplacian is used, and the spectral bias is disabled, yielding a conventional graph encoder. The two ablations therefore distinguish the contribution of learning curvature from the contribution of curvature itself.

Tables 4 and 5 show the corresponding node-classification and link-prediction results on the same four target graphs as the primary transfer tables. The full DRGFM configuration is the best entry in every displayed cell, while fixed Forman curvature is consistently intermediate and the no-curvature variant is lower throughout the reported datasets. The same ordering appears for both few-shot node classification and zero-shot link prediction. These results support the narrower claim that the learnable curvature interface contributes beyond a fixed structural prior under the matched ablation protocol; they do not by themselves establish which individual submodule accounts for every gain.

7 Conclusion

DRGFM provides a single structural interface across pre-training, cross-domain adaptation, and continual learning. The curvature estimator turns local graph geometry into a learnable signal; the curvature-weighted Laplacian turns that signal into a reusable spectral operator; and the anchor regularizer protects important spectral coordinates while leaving other modes available for adaptation. The transfer and ablation results support the contribution of this interface under the matched protocols used here, while the conditional analysis explains why reducing important spectral-coordinate drift can control local historical-task loss drift. Future work should test larger task streams, inductive settings, broader graph families, and explicit trade-offs between anchor memory, spectral approximation, and adaptation cost.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *European conference on computer vision*, pp. 144–161. Springer, 2018.

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=i80OPhOCVH2.

Gregor Bachmann, Gary Bécigneul, and Octavian Ganea. Constant curvature graph convolutional networks. In *International conference on machine learning*, pp. 486–496. PMLR, 2020.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Forman. Bochner's method for cell complexes and combinatorial ricci curvature. *Discrete & Computational Geometry*, 29(3):323–374, 2003.

Karish Grover, Haiyang Yu, Qi Zhu, Han Xie, Vassilis Ioannidis, Christos Faloutsos, et al. Spectro-riemannian graph neural networks. In *International Conference on Learning Representations*, volume 2025, pp. 57833–57859, 2025.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International conference on learning representations*, 2018.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 594–604, 2022.

Xinke Jiang, Rihong Qiu, Yongxin Xu, Wentao Zhang, Yichen Zhu, Ruizhe Zhang, Yuchen Fang, Xu Chu, Junfeng Zhao, and Yasha Wang. Ragraph: A general retrieval-augmented graph learning framework. *Advances in Neural Information Processing Systems*, 37:29948–29985, 2024.

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

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016b.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. *Advances in neural information processing systems*, 34:21618–21629, 2021.

Viet Quan Le and Cuong Viet Ta. Pseudo-riemannian graph transformer. *Advances in Neural Information Processing Systems*, 38:15951–15983, 2026.

Ron Levie, Federico Monti, Xavier Bresson, and Michael M Bronstein. Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing*, 67(1):97–109, 2018.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S Yu, et al. Towards graph foundation models: A survey and beyond. *arXiv preprint arXiv:2310.11829*, 3, 2023.

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S Yu, et al. Graph foundation models: Concepts, opportunities and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):5023–5044, 2025.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794, 2015.

Péter Mernyei and Cătălina Cangea. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Galileo Namata, Ben London, Lise Getoor, Bert Huang, and U Edu. Query-driven active surveying for collective classification. In *10th international workshop on mining and learning with graphs*, volume 8, pp. 1, 2012.

Chien-Chun Ni, Yu-Yao Lin, Feng Luo, and Jie Gao. Community detection on networks with ricci flow. *Scientific reports*, 9(1):9984, 2019.

Yann Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=S1ldO2EFPr`.

Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard Turner, and Mohammad Emtiyaz Khan. Continual deep learning by functional regularisation of memorable past. *Advances in neural information processing systems*, 33:4453–4464, 2020.

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

Benedek Rozemberczki and Rik Sarkar. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 1325–1334, 2020.

Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014, 2021.

Yousef Saad. *Numerical methods for large eigenvalue problems: revised edition*. SIAM, 2011.

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

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

Ondrej Skopek, Octavian-Eugen Ganea, and Gary Bécigneul. Mixed-curvature variational autoencoders. *arXiv preprint arXiv:1911.08411*, 2019.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Li Sun, Junda Ye, Hao Peng, Feiyang Wang, and Philip S Yu. Self-supervised continual graph learning in adaptive riemannian spaces. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 4633–4642, 2023.

Li Sun, Zhenhao Huang, Suyang Zhou, Qiqi Wan, Hao Peng, and Philip Yu. Riemanngfm: Learning a graph foundation model from riemannian geometry. In *Proceedings of the ACM on Web Conference 2025*, pp. 1154–1165, 2025.

Zonggui Tian, Du Zhang, and Hong-Ning Dai. Continual learning on graphs: A survey. *arXiv preprint arXiv:2402.06330*, 2024.

Michalis K Titsias, Jonathan Schwarz, Alexander G de G Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. *arXiv preprint arXiv:1901.11356*, 2019.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.

Shuo Wang, Bokui Wang, Zhixiang Shen, Boyan Deng, and Zhao Kang. Multi-domain graph foundation models: Robust knowledge transfer via topology alignment. *arXiv preprint arXiv:2502.02017*, 2025.

Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I Weidele, Claudio Bellei, Tom Robinson, and Charles E Leiserson. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*, 2019.

Zhanghao Wu, Paras Jain, Matthew Wright, Azalia Mirhoseini, Joseph E Gonzalez, and Ion Stoica. Representing long-range context for graph neural networks with global attention. *Advances in neural information processing systems*, 34:13266–13279, 2021.

Ze Ye, Kin Sum Liu, Tengfei Ma, Jie Gao, and Chao Chen. Curvature graph network. In *International conference on learning representations*, 2019.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823, 2020.

Xingtong Yu, Zechuan Gong, Chang Zhou, Yuan Fang, and Hui Zhang. Samgpt: Text-free graph foundation model for multi-domain pre-training and cross-domain adaptation. In *Proceedings of the ACM on Web Conference 2025*, pp. 1142–1153, 2025.

Haonan Yuan, Qingyun Sun, Junhua Shi, Xingcheng Fu, Bryan Hooi, Jianxin Li, and Philip S Yu. Graver: Generative graph vocabularies for robust graph foundation models fine-tuning. *Advances in Neural Information Processing Systems*, 38:16639–16684, 2026.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. Graphsaint: Graph sampling based inductive learning method. *arXiv preprint arXiv:1907.04931*, 2019.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. Pmlr, 2017.

Mengmei Zhang, Mingwei Sun, Peng Wang, Shen Fan, Yanhu Mo, Xiaoxiao Xu, Hong Liu, Cheng Yang, and Chuan Shi. Graphtranslator: Aligning graph model to large language model for open-ended tasks. In *Proceedings of the ACM Web Conference 2024*, pp. 1003–1014, 2024.

Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. *Advances in neural information processing systems*, 31, 2018.

Haihong Zhao, Aochuan Chen, Xiangguo Sun, Hong Cheng, and Jia Li. All in one and one for all: A simple yet effective method towards cross-domain graph pretraining. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4443–4454, 2024a.

Jianan Zhao, Hesham Mostafa, Michael Galkin, Michael Bronstein, Zhaocheng Zhu, and Jian Tang. Graphany: A foundation model for node classification on any graph. *arXiv preprint arXiv:2405.20445*, 29, 2024b.

A Proofs for the Functional Analysis

This appendix provides proofs for the local functional analysis, implementation details, and the complete evaluation protocols. Claims involving spectral truncation are stated conditionally because the retained Lanczos modes provide an approximation rather than an exact full-spectrum representation.

A.1 Quadratic form and local parameter geometry

Fix an anchor subgraph, its frozen spectral basis, and nonnegative mode weights. Let $q(\theta) = \mathrm{vec}(Q(\theta))$ collect the current anchor spectral coordinates and let

$$R = \mathrm{diag}(\rho) \otimes I_d, \qquad \Phi(\theta) = \frac{1}{2}(q(\theta) - q^*)^\top R(q(\theta) - q^*). \tag{22}$$

At a snapshot θ^* satisfying $q(\theta^*) = q^*$, define $J_q = \partial q / \partial \theta|_{\theta^*}$. A first-order expansion gives

$$q(\theta^* + \Delta\theta) - q^* = J_q \Delta\theta + O(\|\Delta\theta\|^2). \tag{23}$$

Substitution into Φ yields

$$\Phi(\theta^* + \Delta\theta) = \frac{1}{2}\Delta\theta^\top J_q^\top R J_q \Delta\theta + O(\|\Delta\theta\|^3), \tag{24}$$

so

$$\nabla^2 \Phi(\theta^*) = J_q^\top R J_q. \tag{25}$$

This matrix is positive semidefinite. It is generally non-diagonal because several parameter directions can jointly affect the same spectral coordinates. Directions in the null space of J_q are not penalized to second order, which formalizes the distinction between preserving a function and freezing all parameters.

A.2 Conditional historical-loss bound

Let $\ell_s(q)$ denote the historical-task loss expressed as a function of the anchor coordinates. Assume that ℓ_s is twice continuously differentiable in a neighborhood $\mathcal{N}$ of q^*, that $\nabla \ell_s(q^*) = 0$, and that

$$\nabla^2 \ell_s(q) \preceq \mu R, \qquad q \in \mathcal{N}. \tag{26}$$

For $\Delta q = q - q^*$, Taylor's theorem with an integral remainder gives

$$\ell_s(q) - \ell_s(q^*) = \int_0^1 (1-a)\, \Delta q^\top \nabla^2 \ell_s(q^* + a\Delta q)\, \Delta q \, da. \tag{27}$$

Using equation 26 and $\int_0^1 (1-a)da = 1/2$ gives

$$\ell_s(q) - \ell_s(q^*) \le \frac{\mu}{2} \Delta q^\top R \Delta q. \tag{28}$$

The right-hand side is μ times the quantity minimized by the spectral-coordinate regularizer. The conclusion is local and conditional: it requires the historical loss to be stationary at the snapshot and its Hessian to be dominated by the chosen mode-weight matrix. It does not imply a global guarantee for an arbitrary nonlinear backbone or an evolving spectral basis.

B Properties of the Curvature Laplacian

B.1 Positive semidefiniteness

For the symmetric nonnegative adjacency W_κ used in the paper, we assume zero-degree nodes are removed or handled by the standard convention $D_\kappa^{-1/2}[u,u] = 0$. Define $D_\kappa[u,u] = \sum_v W_\kappa[u,v]$ and

$$L_\kappa = I - D_\kappa^{-1/2} W_\kappa D_\kappa^{-1/2}. \tag{29}$$

Let $S = D_\kappa^{-1/2} W_\kappa D_\kappa^{-1/2}$. Since S is similar to the random-walk matrix $D_\kappa^{-1} W_\kappa$, its eigenvalues lie in $[-1,1]$ for an undirected graph. Therefore the eigenvalues of $L_\kappa = I - S$ lie in $[0,2]$, and L_κ is positive semidefinite. Equivalently, with the combinatorial Laplacian $D_\kappa - W_\kappa$,

$$x^\top (D_\kappa - W_\kappa) x = \frac{1}{2} \sum_{u,v} W_\kappa[u,v](x_u - x_v)^2 \geq 0. \tag{30}$$

This also justifies the shifted operator $\widetilde{L}_\kappa = L_\kappa - I$, whose spectrum lies in $[-1,1]$ and is therefore suitable for Chebyshev recurrence.

B.2 Polynomial filter stability

For $\widetilde{\lambda} \in [-1,1]$, the Chebyshev polynomial satisfies $|T_k(\widetilde{\lambda})| \leq 1$. Hence a filter

$$g(\widetilde{L}_\kappa) = \sum_{k=0}^{K} \alpha_k T_k(\widetilde{L}_\kappa) \tag{31}$$

obeys

$$\|g(\widetilde{L}_\kappa)\|_2 \leq \sum_{k=0}^{K} |\alpha_k|. \tag{32}$$

The bound is independent of the graph size. In implementation, the recurrence applies sparse matrix–dense matrix products with L_κ, avoiding a full eigendecomposition during ordinary message passing. Lanczos is used only when explicit anchor modes are required.

C Spectral Truncation

Let Δq_i denote the displacement of mode i and define the full and retained weighted displacements by

$$\Omega = \sum_{i=1}^{N} \rho_i \|\Delta q_i\|_2^2, \qquad \Omega_m = \sum_{i=1}^{m} \rho_i \|\Delta q_i\|_2^2. \tag{33}$$

If $\|\Delta q_i\|_2 \leq G$ and the omitted-mode weights satisfy $\rho_i \leq \bar{\rho}\, \varphi(\lambda_i)$ for a non-increasing function φ, then

$$0 \leq \Omega - \Omega_m \leq G^2 \bar{\rho} \sum_{i>m} \varphi(\lambda_i). \tag{34}$$

For the coarser bound $\sum_{i>m} \varphi(\lambda_i) \leq (N-m)\varphi(\lambda_{m+1})$, the right-hand side becomes $G^2\bar{\rho}(N-m)\varphi(\lambda_{m+1})$. This result does not assume that the spectral tail is zero; it states the additional condition required for a small tail to imply a small approximation error. The main experiments use $m = 32$ anchor modes and should therefore be interpreted under this finite-mode approximation.

D Evaluation Details

D.1 Few-shot node classification

We use a transductive episodic protocol with $N = \min(5, C)$ and $K \in \{1, 5\}$, where C is the number of target classes. When $C > 5$, each episode samples five classes; otherwise all classes are used. We sample K support nodes per selected class from the test split, and use the remaining nodes

of those classes as queries. Support and query nodes are included in the same graph for message passing, but only support labels are used to fit the adapter and linear probe. We report the mean and standard deviation over 100 independent episodes.

D.2 Held-out link prediction

For each target graph, edges are split into 85% training and 15% test edges. Node representations, curvature estimation, and message passing use only training edges. Test edges are used only for evaluation and are excluded from negative sampling. Negative pairs are sampled from non-existent pairs under the training graph, with the number of negatives matched to the positive test edges. We report ROC-AUC and Average Precision over 10 negative-sampling runs.

D.3 Continual learning

The temporal stream contains five binary node-classification tasks: `elliptic-1-8`, `elliptic-9-16`, `elliptic-17-24`, `elliptic-25-32`, and `elliptic-33-40`. The label space is fixed while transaction distributions and graph structure change. After each task, we store a connected anchor subgraph with at most 1024 nodes, cache its curvature-spectral basis and snapshot coordinates, and retain 32 Lanczos modes for subsequent regularization.

Let $A_{i,j}$ denote accuracy on task i after training task j. The reported metrics are

$$\mathrm{AA} = \frac{1}{T}\sum_{i=1}^{T} A_{i,T}, \tag{35}$$

$$\mathrm{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1}(A_{i,T} - A_{i,i}), \tag{36}$$

$$\mathrm{FWT} = \frac{1}{T-1}\sum_{i=2}^{T}(A_{i,i-1} - A_{i,\mathrm{random}}). \tag{37}$$

Forgetting is the average historical performance drop, computed from the same accuracy matrix using the maximum pre-final accuracy for each previous task. Positive BWT indicates improvement on previous tasks; negative BWT indicates forgetting. FWT uses the random-initialization baseline under the same training budget.

E Baseline Methods

We describe the transfer baselines in Tables 1 and 2, followed by the continual-learning baselines in Table 3. The descriptions summarize the original methods; the results in this paper use the evaluation protocols specified above. A method's original task scope should be distinguished from its use in our experiments, particularly when a node-classification method is evaluated for link prediction.

E.1 Cross-domain transfer baselines

SAMGPT. SAMGPT is a text-free graph foundation model for multi-domain pre-training and cross-domain adaptation (Yu et al., 2025). It addresses discrepancies between graph domains through feature alignment and structure-aware prompting, and transfers pre-trained knowledge through prompts. It provides a comparison with prompt-based adaptation without relying on textual descriptions of graphs.

RiemannGFM. RiemannGFM learns transferable graph representations from Riemannian geometry (Sun et al., 2025). It uses non-Euclidean geometric modeling to represent structural patterns across graphs. This is a geometry-aware comparator for DRGFM: the comparison concerns the complete models, whereas the matched curvature ablations separately examine the contribution of learning curvature.

GCOPE. GCOPE is a cross-domain graph pre-training framework that combines graphs into a unified pre-training structure using coordinator nodes (Zhao et al., 2024a). These nodes facilitate information exchange across source graphs despite domain differences. It offers a graph-composition approach to sharing knowledge, in contrast to DRGFM's curvature-dependent spectral operator.

GraphAny. GraphAny targets inductive node classification across graphs with different feature and label spaces (Zhao et al., 2024b). It combines analytical linear graph models with learned attention to transfer classification behavior across graphs. Its original formulation is classification-oriented; the link-prediction results here should be interpreted under our held-out-edge protocol rather than as its original benchmark setting.

MDGFM. MDGFM studies robust multi-domain knowledge transfer through topology alignment (Wang et al., 2025). It addresses structural discrepancies between source and target domains, making it a relevant comparison for cross-domain structural transfer. Unlike the continual-learning comparison, these transfer experiments do not require retaining performance on previously visited targets.

RAGraph. RAGraph is a retrieval-augmented graph learning framework (Jiang et al., 2024). It retrieves relevant graph knowledge and incorporates the retrieved information into downstream prediction. This provides a retrieval-based alternative to encoding transferable structural information within a shared backbone.

GRAVER. GRAVER uses generative graph vocabularies for robust graph foundation model fine-tuning (Yuan et al., 2026). It organizes reusable graph knowledge through a learned vocabulary to support downstream adaptation. It therefore represents a vocabulary-based transfer strategy, complementary to the geometric and prompting baselines.

QGT. QGT (Q-GT) is a graph Transformer built in pseudo-Riemannian geometry (Le & Ta, 2026). It introduces a diffeomorphic framework that decomposes a pseudo-Riemannian manifold into the product of spherical and hyperbolic spaces, enabling linear attention, residual connection, and layer normalization while preserving hierarchical graph structure. It further develops a lightweight space-searching algorithm based on Gaussian sectional curvature to select the optimal pseudo-Riemannian manifold. This provides a geometry-aware Transformer alternative to Euclidean and Riemannian graph models.

E.2 Continual-learning baselines

EWC. Elastic Weight Consolidation estimates parameter importance using a diagonal Fisher approximation and penalizes deviations from historical parameter values (Kirkpatrick et al., 2017). It provides a parameter-space stability baseline, unlike DRGFM's penalty on anchor spectral coordinates.

SI. Synaptic Intelligence accumulates parameter-importance estimates along the optimization trajectory (Zenke et al., 2017). Parameters that contributed more to reducing earlier-task losses receive stronger constraints during subsequent learning. This contrasts with the snapshot-based importance estimate used by EWC.

MAS. Memory Aware Synapses estimates importance through the sensitivity of model outputs to parameter changes (Aljundi et al., 2018). Its importance estimation does not require task labels, but the resulting regularization still constrains parameters rather than preserving spectral-coordinate values directly.

ER-GNN. ER-GNN is the graph experience-replay comparator in Table 3. It retains historical examples and revisits them during learning on later tasks. The distinction from DRGFM is the retention objective: replay rehearses historical examples, whereas DRGFM preserves weighted spectral coordinates on stored anchor graphs. Both approaches require historical memory; the comparison does not establish an equal-memory advantage unless their storage budgets are matched.

FROMP. Functional Regularisation of Memorable Past preserves predictive functions on selected memorable inputs (Pan et al., 2020). It is the closest functional-regularization comparator: FROMP constrains predictive behavior, whereas DRGFM constrains intermediate representations in frozen curvature-spectral bases. The comparison tests these choices of functional target, not a general superiority of functional over parameter regularization.

F Implementation Notes

The curvature estimator combines the closed-form Forman prior with a feature-dependent residual. An Ollivier–Ricci warm-up is applied only in the early training phase and its weight is cosine-annealed to zero. The curvature-weighted adjacency uses clipped exponential weights to avoid numerical overflow. The mixed-curvature representation performs linear operations in the shared tangent space and maps representations back through the corresponding exponential map.

Continual-learning anchors store node indices, edge structure, the frozen spectral basis, snapshot coordinates, and mode weights. Input features are recomputed during later regularization passes; the stored anchor therefore defines a structural and functional reference rather than a permanently frozen feature tensor. This makes the method memory-based, not a strict no-replay method.

All transfer targets are unseen during pre-training. The few-shot protocol uses support labels only, while held-out link prediction uses no target edge labels or parameter updates. These visibility rules are part of the reported evaluation protocol and are required when reproducing the results.

G Additional Discussion

The analysis is intentionally expressed in terms of anchor spectral coordinates rather than a global RKHS or an exact full-spectrum representation. This matches the implemented regularizer, which uses a frozen basis on bounded anchor subgraphs and a finite number of Lanczos modes. Stronger statements about global filter generalization, optimal basis alignment, or test-time proximal convergence require additional assumptions and are outside the claims made in this paper.