RIEMANNIAN CURVATURE-TRANSPORT SPECTRAL NETWORKS FOR HETEROPHILOUS GRAPH LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph neural networks (GNNs) on heterophilous graphs face two coupled challenges: neighborhood aggregation can mix semantically incompatible nodes, and a single latent geometry may fail to capture the coexistence of hierarchical, flat, and cyclic graph structure. Existing methods typically address only one of them: geometric GNNs improve representation fidelity in curved spaces but still rely on smoothing-based aggregation, while heterophily-aware spectral methods improve signal separation yet ignore the geometric mismatch between node representations that live in different tangent spaces. We propose **CTS-Net**, a Curvature-Transport Spectral Network that resolves these limitations by combining mixed-curvature representation learning with transport-aware and track-specific propagation. CTS-Net embeds nodes in a mixed-curvature product manifold to model heterogeneous graph geometry, parallel-transports neighbor messages to the receiver tangent space to align cross-node representations before aggregation, routes the aligned messages into smooth and contrastive tracks through a learned orthogonal operator to separate compatible from incompatible signals, and applies track-specific polynomial diffusion to learn distinct spectral responses . Across 14 node-classification benchmarks, CTS-Net achieves the strongest overall performance among the compared methods, with particularly consistent gains on heterophilous graphs, while remaining scalable in memory and runtime.

1 INTRODUCTION

Graph neural networks rely on a foundational assumption: neighbouring nodes share similar labels and thus benefit from aggregating each other's representations. This assumption fails systematically on *heterophilous graphs*, where structurally adjacent nodes frequently belong to different classes (Hongbin Pei, 2020; Zhu et al., 2020; Platonov et al., 2023). Web graphs link hub pages to semantically unrelated content; protein interaction networks connect binding partners with opposing functional roles; citation networks join papers that critique rather than extend one another. In each case, naive neighbourhood aggregation GNNs destroys precisely the discriminative signal (high-frequency signal) that classification requires.

Real-world graphs also contain geometric structure that is difficult to represent in a single Euclidean space. For example, hierarchical or tree-like relations, such as taxonomies and citation graphs, are often better captured by hyperbolic geometry (Nickel and Kiela, 2017; Chami et al., 2019). In contrast, densely connected communities with many short cycles are more naturally modeled by spherical geometry (Gu et al., 2019; Liu et al., 2019). Since many graphs contain both types of structure, forcing all nodes into one Euclidean latent space can distort the graph geometry and can also reinforce inappropriate smoothing during aggregation.

Several lines of work have tried to address these issues, but each leaves an important limitation. *Geometric GNNs* reduce representation distortion by embedding graphs in non-Euclidean spaces. For example, HGCN (Chami et al., 2019) uses the Poincaré ball, κ-GCN (Bachmann et al., 2020) extends this idea to constant-curvature spaces, mixed-curvature models (Gu et al., 2019) combine multiple geometric regimes, and Q-GCN (Bo Xiong, 2022) considers pseudo-Riemannian geometry. However, these methods still aggregate neighbors through smoothing: messages are mapped to a tangent space and averaged. Thus, compatible and incompatible neighbors can still be mixed in the same channel, so the basic heterophily problem remains.

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.

Heterophily-aware methods (Hongbin Pei, 2020; Zhu et al., 2020; Chien et al., 2021; Mingguo He, 2021) improve signal separation through signed propagation, adaptive diffusion, or multi-track routing, but they typically operate in a shared Euclidean latent space and therefore do not account for the geometric mismatch between tangent spaces at different manifold points. *Geometry-aware spectral methods* such as CUSP (Grover et al., 2025) incorporate curvature information into Laplacian-based diffusion but still use curvature primarily through scalar edge modulation rather than through explicit transport between tangent spaces. As a result, they improve edge weighting but do not directly address the geometric consistency of cross-node aggregation.

Motivated by these limitations, we propose **CTS-Net** (Curvature-Transport Spectral Network), a framework that treats message passing as geometric transport rather than scalar averaging. The key idea is that node representations on a curved manifold reside in distinct tangent spaces, so aggregating them without alignment can introduce geometric inconsistency that scalar reweighting alone may fail to resolve. CTS-Net addresses this by first transporting each neighbor's message to the receiver's tangent space using parallel transport along the connecting geodesic, ensuring that all messages are geometrically comparable before aggregation. The transported messages are then routed into two semantic tracks: a smooth track for compatible neighbors and a contrastive track for incompatible neighbors through an orthogonal operator parameterized by edge-level compatibility scores derived from local curvature and structural features. Because the routed representations still reside in receiver-specific tangent spaces, CTS-Net next transports them to a shared reference tangent space before multi-hop spectral propagation. This common reference construction makes the subsequent polynomial diffusion independent of arbitrary local tangent-frame choices and preserves the geometric consistency that motivates transport in the first place. Finally, each track undergoes independent polynomial diffusion, allowing the model to learn distinct spectral responses for smooth and contrastive content instead of forcing both through a single shared filter. We theoretically establish the geometric consistency of this shared-reference propagation, show that parallel transport and orthogonal routing introduce no additional pathwise norm contraction, and characterize the spectral capacity provided by independent track filters. Oversmoothing and oversquashing are further evaluated empirically. Moreover, CTS-Net delivers strong and consistent improvements across the 14 benchmarks considered in our evaluation protocol.

Our contributions are fourfold. First, we introduce a geometric message-passing framework that combines mixed-curvature embeddings with parallel transport for cross-node aggregation and a shared reference tangent space for geometrically consistent multi-hop propagation. Second, we propose an orthogonal routing mechanism that softly redistributes transported messages between complementary smooth and contrastive tracks without relying on sign flipping or discarding the combined signal. Third, we establish frame-consistent spectral propagation, characterize the non-contractive role of transport and routing along individual paths, and analyze the spectral expressivity provided by independent track filters. Fourth, we demonstrate strong empirical performance across 14 node-classification benchmarks together with controlled ablations, mechanism analysis, and scalability experiments.

2 Related Work

Geometric and Heterophily-Aware GNNs. Geometric GNNs embed nodes in curved spaces to reduce representation distortion: HGCN (Chami et al., 2019) uses the Poincaré ball, κ-GCN (Bachmann et al., 2020) generalises to constant-curvature spaces, and product manifolds (Gu et al., 2019) allow mixed geometric regimes. However, these retain smoothing-based aggregation that mixes incompatible neighbours. Heterophily-aware methods address this through flexible spectral filters (Chien et al., 2021; Mingguo He, 2021), signed propagation (Qi, 2024), or multi-track routing (Pei et al., 2024), but treat node states as vectors in a shared Euclidean space, ignoring tangent-space geometry. CTS-Net combines both perspectives: geometry defines *how* messages are transported, while routing determines *where* they aggregate.

Curvature-Aware Spectral Methods. CUSP (Grover et al., 2025) incorporates Ollivier–Ricci curvature (Ollivier, 2009) into spectral diffusion, and discrete curvature has been linked to oversquashing and oversmoothing (Topping et al., 2022; Nguyen et al., 2023). These approaches apply curvature as scalar edge weights on the standard Laplacian rather than transporting messages between tangent spaces. CTS-Net instead uses curvature within an edge descriptor that parameterises an or-

thogonal routing operator, replacing scalar reweighting with operator-valued transport that preserves message norms while separating compatible from incompatible content.

3 BACKGROUND

Heterophilous Graphs. Let $G = (V, E, X)$ be a graph with n nodes, edge set E, and features $X \in \mathbb{R}^{n \times d_0}$. The *homophily ratio* $h = |\{(i,j) \in E : y_i = y_j\}|/|E|$ measures the fraction of edges connecting same-label nodes. Standard GNNs aggregate neighbour representations as $h_i^{(\ell+1)} = \sigma(\sum_{j \in \mathcal{N}(i)} \omega_{ij} h_j^{(\ell)} W^{(\ell)})$ with $\omega_{ij} \geq 0$. When $h \approx 1$, this smoothing reinforces class structure; when $h \ll 1$ (heterophily), neighbours with different labels are averaged together and discriminative information is destroyed.

Riemannian Manifolds. A Riemannian manifold $(\mathcal{M}, g)$ is a smooth space where each point $z \in \mathcal{M}$ has an associated tangent space $T_z\mathcal{M}$ equipped with an inner product $g_z(\cdot, \cdot)$. The exponential map $\exp_z : T_z\mathcal{M} \to \mathcal{M}$ sends a tangent vector to the manifold along a geodesic, while the logarithmic map $\log_z$ locally inverts this operation. Unlike Euclidean space, tangent spaces at different points are different vector spaces, so vectors based at z_i and z_j cannot be compared or added directly.

Product Manifolds. Real graphs often contain multiple geometric regimes at once: hierarchical regions are naturally modeled in hyperbolic space, locally flat regions in Euclidean space, and highly cyclic communities in spherical space. A product manifold $\mathcal{M} = \mathbb{H}^{d_h}_{\kappa_h} \times \mathbb{R}^{d_e} \times \mathbb{S}^{d_s}_{\kappa_s}$ combines these geometries in a single latent space. Its tangent space decomposes component-wise as $T_z\mathcal{M} \cong T_{z^{(h)}}\mathbb{H} \oplus \mathbb{R}^{d_e} \oplus T_{z^{(s)}}\mathbb{S}$, and parallel transport acts separately on each factor.

Oversmoothing and Oversquashing. *Oversmoothing* refers to the tendency of repeated message passing to make node representations increasingly similar, eventually reducing their discriminative power. *Oversquashing* refers to the compression of information from many distant nodes into fixed-size representations, so that long-range dependencies are severely attenuated when propagated through narrow graph bottlenecks. These two phenomena are especially harmful on heterophilous graphs, where useful signal often lies in contrastive interactions and long-range structure.

4 METHODOLOGY

CTS-Net propagates messages in four stages. It first embeds nodes in a mixed-curvature product manifold $\mathcal{M}$ to model hierarchical, flat, and cyclic graph structure in a unified latent space. It then parallel-transports neighbor states to the receiver tangent space so that cross-node messages are geometrically aligned before aggregation. The aligned messages are routed into smooth and contrastive tracks, mapped to a shared reference tangent space, and then propagated with independent polynomial diffusion filters.

4.0.1 MIXED-CURVATURE PRODUCT MANIFOLD

Real-world graphs can simultaneously exhibit hierarchical, locally flat, and cyclic structure. We therefore represent nodes in the product manifold

$$\mathcal{M} = \mathbb{H}^{d_h}_{\kappa_h} \times \mathbb{R}^{d_e} \times \mathbb{S}^{d_s}_{\kappa_s}, \qquad \kappa_h < 0, \quad \kappa_s > 0, \tag{1}$$

where the hyperbolic, Euclidean, and spherical factors model tree-like, approximately flat, and cyclic structure, respectively (Gu et al., 2019).

Given $x_i \in \mathbb{R}^{d_0}$, separate learnable projections produce $z_i^{(h)} = \exp^{\mathbb{H}}_{o_h}(W_h x_i)$, $z_i^{(e)} = W_e x_i$, and $z_i^{(s)} = \exp^{\mathbb{S}}_{o_s}(W_s x_i)$. The resulting node position is $z_i = (z_i^{(h)}, z_i^{(e)}, z_i^{(s)}) \in \mathcal{M}$. Each node additionally carries a tangent state $h_i^{(\ell)} \in T_{z_i}\mathcal{M}$, initialized from its input features in local tangent coordinates.

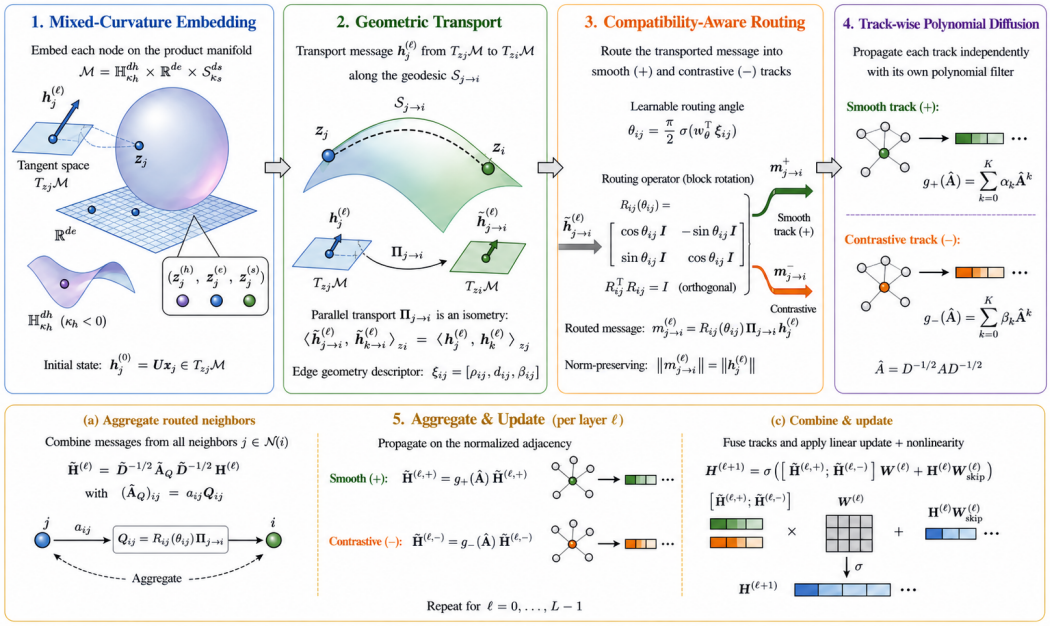


Figure 1: CTS-Net framework: transport-aligned message passing on a product manifold with orthogonal routing and track-wise polynomial filtering.

4.1 Geometric Transport

Since $h_j^{(\ell)} \in T_{z_j}\mathcal{M}$ and $h_i^{(\ell)} \in T_{z_i}\mathcal{M}$ belong to different tangent spaces, they cannot be directly aggregated. CTS-Net therefore aligns each source message using Levi–Civita parallel transport along the minimizing geodesic from z_j to z_i. For the product manifold, transport decomposes factor-wise. Let $R_h = 1/\sqrt{-\kappa_h}$ and $R_s = 1/\sqrt{\kappa_s}$. For a tangent state $h_j = [h_j^{(h)}; h_j^{(e)}; h_j^{(s)}]$, we compute

$$\Pi_{j\to i} h_j = \left[\Pi^{\mathbb{H}}_{z_j^{(h)} \to z_i^{(h)}} h_j^{(h)};\ h_j^{(e)};\ \Pi^{\mathbb{S}}_{z_j^{(s)} \to z_i^{(s)}} h_j^{(s)} \right], \tag{2}$$

where the hyperbolic transport is $\Pi^{\mathbb{H}}_{x\to y}(v) = v + \frac{\langle y,v\rangle_L}{R_h^2 - \langle x,y\rangle_L}(x+y)$, with $\langle\cdot,\cdot\rangle_L$ the Lorentz inner product, while the spherical transport is $\Pi^{\mathbb{S}}_{x\to y}(v) = v - \frac{\langle y,v\rangle}{R_s^2 + \langle x,y\rangle}(x+y)$. The Euclidean component uses the identity map.

Proposition 1 (Isometric Transport) *Let $\Pi_{j\to i}$ be the factor-wise Levi–Civita transport in Eq. equation 2. Then, for any $u, v \in T_{z_j}\mathcal{M}$, $\langle \Pi_{j\to i}u, \Pi_{j\to i}v\rangle_{z_i} = \langle u, v\rangle_{z_j}$. hence, $\|\Pi_{j\to i}u\|_{z_i} = \|u\|_{z_j}$ and all pairwise angles are preserved.*

Additional implementation and numerical details are given in Appendix B.

4.2 Edge Geometry Descriptor

Local edge geometry provides a useful structural cue for distinguishing within-region connections from bridge-like interactions. For scalable computation, we use a simplified degree–triangle score inspired by Balanced Forman curvature (Topping et al., 2022):

$$\rho_{ij} = \frac{2}{d_i} + \frac{2}{d_j} - 2 + 2|\mathcal{N}(i) \cap \mathcal{N}(j)| \left(\frac{1}{d_i} + \frac{1}{d_j} \right). \tag{3}$$

We refer to ρ_{ij} as the *simplified Balanced-Forman structural score* (sBF). It retains inexpensive degree and triangle information but should not be interpreted as the complete Balanced Forman

curvature. Empirically, larger values tend to correspond to locally redundant edges, whereas smaller values indicate more bridge-like connectivity. Then define edge descriptor for edge $(j \to i)$ as,

$$\xi_{ij} = \big[\rho_{ij},\ \|z_i - z_j\|_2,\ \beta_{ij}\big] \in \mathbb{R}^3, \tag{4}$$

where $\|z_i - z_j\|_2$ is a standardized extrinsic distance used as an inexpensive routing cue, and $\beta_{ij} = |d_i - d_j|/(d_i + d_j + 1)$ indicates degree imbalance. Rather than using ρ_{ij} as a fixed diffusion coefficient, CTS-Net treats ξ_{ij} as a learned routing and edge-strength cue.

4.3 Compatibility-Aware Routing

In heterophilous graphs, mixing high- and low-frequency information in a single channel can blur the distinct semantic roles of nodes (Pei et al., 2024). CTS-Net therefore separates transported messages into two channels, so that compatible and incompatible interactions are not forced into the same representation. After transport to the receiver tangent space, each message is expressed in local orthonormal coordinates, allowing routing to be performed in a Euclidean space of $T_{z_i}\mathcal{M}$.

We split the tangent-space state as $h_i^{(\ell)} = [h_i^{(\ell,+)}; h_i^{(\ell,-)}]$, where $h_i^{(\ell,+)}, h_i^{(\ell,-)} \in \mathbb{R}^{d/2}$. The $+$ track captures compatible content, while the $-$ track captures contrastive content. We initialize $h_i^{(0,+)} = Wx_i$ and $h_i^{(0,-)} = \mathbf{0}$, so the contrastive track is populated only through routing. Initializing the contrastive track to zero makes its content entirely induced by routing, which helps isolate the effect of incompatible interactions rather than injecting contrastive features a priori. For each edge, we compute a routing angle $\theta_{ij} = \frac{\pi}{2}\sigma(w_\theta^\top \xi_{ij})$, where $w_\theta \in \mathbb{R}^3$ is learnable. The angle provides a continuous allocation between the two complementary tracks: small θ_{ij} retains more message energy in the $+$ track, while larger θ_{ij} transfers more energy to the $-$ track. Intermediate angles allow both tracks to remain active.. The routing operator is the block rotation,

$$R_{ij}(\theta_{ij}) = \begin{bmatrix} \cos\theta_{ij}\, I_{d/2} & -\sin\theta_{ij}\, I_{d/2} \\ \sin\theta_{ij}\, I_{d/2} & \cos\theta_{ij}\, I_{d/2} \end{bmatrix}, \tag{5}$$

which is orthogonal and therefore norm-preserving. The combined transport-routing operator is $Q_{ij} = R_{ij}(\theta_{ij})\Pi_{j\to i}$, yielding $m_{j\to i}^{(\ell)} = Q_{ij} h_j^{(\ell)}$. The routing operator redirects incompatible content to a separate channel without changing its norm.

4.3.1 Transport-Route-Diffuse Architecture

Let $h_i^{(\ell)} \in T_{z_i}\mathcal{M}$ denote the node state at layer ℓ. The layer update proceeds in three steps.

Transport and Route. For each receiver i, we first aggregate messages in its local tangent space:

$$\bar{h}_i^{(\ell)} = \sum_{j\in\mathcal{N}(i)} \frac{a_{ij}}{\sqrt{\tilde{d}_i \tilde{d}_j}} Q_{ij} h_j^{(\ell)}, \qquad Q_{ij} = R_{ij}(\theta_{ij})\Pi_{j\to i}, \tag{6}$$

where $a_{ij} = \mathrm{softplus}(w_a^\top \xi_{ij})$ and $\tilde{d}_i = \sum_{j\in\mathcal{N}(i)} a_{ij}$. Thus, $\bar{h}_i^{(\ell)} \in T_{z_i}\mathcal{M}$ contains the transported and routed smooth/contrastive information at node i.

Shared-Reference Track Diffusion. Although Eq. equation 6 aligns messages at each receiver, the resulting states still belong to different tangent spaces. Before polynomial diffusion, we therefore transport every receiver-local state to a common reference tangent space $\mathcal{V} = T_o\mathcal{M}$: $\bar{y}_i^{(\ell)} = C_i \bar{h}_i^{(\ell)}, \qquad C_i = \Pi_{z_i\to o}$. Stacking $\bar{y}_i^{(\ell)}$ gives $\bar{Y}^{(\ell)} = [\bar{Y}^{(\ell,+)}; \bar{Y}^{(\ell,-)}]$. Hence every row now lies in the same vector space $\mathcal{V}$, and the track-wise polynomial filters are well defined: $\tilde{Y}^{(+)} = g_+(\hat{A})\bar{Y}^{(+)}, \qquad \tilde{Y}^{(-)} = g_-(\hat{A})\bar{Y}^{(-)}$, where$g_+(\hat{A}) = \sum_{k=0}^K \alpha_k \hat{A}^k, \qquad g_-(\hat{A}) = \sum_{k=0}^K \beta_k \hat{A}^k$. The two filters are learned independently, allowing smooth and contrastive content to adopt different spectral responses.

Combine and Transform. The two tracks are concatenated in the shared reference space:

$$y_i^{(\ell+1)} = \sigma\Big([\tilde{y}_i^{(+)}; \tilde{y}_i^{(-)}]W^{(\ell)} + C_i h_i^{(\ell)} W_{\mathrm{skip}}^{(\ell)}\Big). \tag{7}$$

For intermediate layers, the updated state is transported back as $h_i^{(\ell+1)} = \Pi_{o\to z_i} y_i^{(\ell+1)}$ before the next transport step. After the final layer, we instead use the shared-reference representation $y_i^{(L)}$ for prediction: $\ell_i = W_{\mathrm{cls}} y_i^{(L)} + b_{\mathrm{cls}}$, and train the entire model end-to-end using cross-entropy over the labeled training nodes.

5 THEORETICAL ANALYSIS

Prior work on multi-track message passing shows that heterophily mixing can accelerate representation collapse, while separating semantically different messages into independently propagated tracks improves the representation separation condition and empirically provides stronger resistance to oversmoothing (Pei et al., 2024). CTS-Net follows this principle with learned orthogonal routing.

Theorem 1 (Resistance to Oversmoothing) *Define the inter-track discrepancy after the transport–routing step by* $\Delta^{(\ell)} = \frac{1}{n}\sum_{i=1}^{n} \|\bar{h}_i^{(\ell,+)} - \bar{h}_i^{(\ell,-)}\|^2$. *Suppose CTS-Net uses the orthogonal routing operator* $R_{ij}(\theta_{ij})$ *defined in Eq. equation 5, with* $\theta_{ij} \in (0, \pi/2)$ *for all* $(i,j) \in E$. *Assume that at least a fraction* $\eta > 0$ *of edges satisfy* $\theta_{ij} \geq \theta_{\min} > 0$. *Under the nondegenerate routed-aggregation condition specified in Appendix A.2, there exists a graph-dependent constant* $C > 0$ *such that*

$$\Delta^{(\ell)} \geq \sin^2(\theta_{\min})\, \eta\, C\, \frac{\|H^{(\ell)}\|_F^2}{n}. \tag{8}$$

Theorem 1 shows that, under nontrivial routing and nondegenerate aggregation, the smooth and contrastive tracks retain a nonvanishing discrepancy proportional to the propagated signal energy. Following the multi-track view (Pei et al., 2024), this separation prevents compatible and contrastive information from being repeatedly homogenized into a single channel, providing a mechanism for resistance to heterophily-induced representational collapse. We complement this result with node-level Dirichlet-energy measurements in Sec. 6.1. The proof is given in Appendix A.2.

Long-range sensitivity. Oversquashing occurs when information from distant nodes becomes strongly attenuated while propagating through graph bottlenecks. This effect is commonly studied through the sensitivity $\|\partial h_i^{(L)}/\partial x_j\|$, since a vanishing Jacobian indicates that a distant node j has little influence on the representation of node i (Topping et al., 2022). We therefore analyze how the transport–routing operators affect this pathwise sensitivity.

Theorem 2 (Resistance to Oversquashing) *Let* $P_{j\to i} = (v_0 = j, \ldots, v_r = i)$ *be a path of length* $r \leq L$, *and let* $\frac{\partial h_i^{(L)}}{\partial x_j}\big|_{P_{j\to i}}$ *denote the Jacobian contribution of CTS-Net along this path. Assume that the activation Jacobians along the path satisfy* $\sigma_{\min}(J_\sigma^{(\ell)}) \geq m_\sigma > 0$ *for all* ℓ. *Then*

$$\left\| \frac{\partial h_i^{(L)}}{\partial x_j}\Big|_{P_{j\to i}} \right\|_{\mathrm{op}} \geq m_\sigma^r \prod_{\ell=1}^{r} \sigma_{\min}\big(W^{(\ell)}\big) \prod_{(p,q)\in P_{j\to i}} \frac{a_{qp}}{\sqrt{\tilde{d}_q \tilde{d}_p}} \tag{9}$$

In particular, transport introduces no additional norm contraction along the path beyond the normalized edge weights, layer maps, and activation Jacobians.

Theorem 2 shows that, because parallel transport is isometric and the routing operator is orthogonal in local orthonormal coordinates, these operations introduce no additional norm contraction along a message path. Consequently, the pathwise Jacobian admits a lower bound determined by the normalized edge weights, layer transformations, and activation Jacobians, rather than by the transport–routing operators. This provides a theoretical mechanism for preserving long-range sensitivity in CTS-Net, which we further evaluate empirically in Sec. 6.1 and Fig. 5. The proof is given in Appendix A.3.

Theorem 3 (Spectral Heterophily Decomposition) *Under a spectral-gap assumption, there exist degree-*K *polynomial filters* g_+ *and* g_- *that can separately approximate the low-frequency and high-frequency spectral components of* $\hat{A}$. *Hence, the two CTS-Net tracks have sufficient capacity to model distinct smooth and contrastive responses.(Formal statement and proof in Appendix A.4)*

6 EXPERIMENTS

Datasets. As summarized in Table 1, our evaluation includes 14 datasets grouped under the heterophilous setting (Texas, Wisconsin, Cornell, Actor, Chameleon, Squirrel, Roman-empire, and Questions), three homophilous benchmarks (Tolokers, Coauthor-CS, and Coauthor-Physics), and

Method	Heterophilous								Homophilous			Large-scale		
	Texas	Wisc.	Corn.	Actor	Cham.	Squi.	Roman	Ques.	Tolo.	Co.-CS	Co.-Phys	Amaz.	Penn94	Arxiv
GPR-GNN	$80.8_{\pm 0.9}$	$83.1_{\pm 1.0}$	$72.4_{\pm 1.2}$	$36.9_{\pm 0.8}$	$61.1_{\pm 1.1}$	$48.6_{\pm 1.0}$	$71.4_{\pm 0.9}$	$51.2_{\pm 0.8}$	$62.4_{\pm 0.7}$	$92.1_{\pm 0.4}$	$95.0_{\pm 0.3}$	$44.8_{\pm 0.6}$	$78.6_{\pm 0.5}$	$71.8_{\pm 0.4}$
HGCN	$80.1_{\pm 1.0}$	$82.7_{\pm 1.1}$	$71.8_{\pm 1.1}$	$35.8_{\pm 0.9}$	$60.4_{\pm 1.0}$	$47.4_{\pm 1.1}$	$72.1_{\pm 0.8}$	$50.8_{\pm 0.9}$	$62.1_{\pm 0.8}$	$91.4_{\pm 0.5}$	$94.5_{\pm 0.4}$	$45.2_{\pm 0.7}$	*OOM*	*OOM*
κ-GCN	$81.3_{\pm 0.8}$	$83.6_{\pm 0.9}$	$73.1_{\pm 1.0}$	$37.5_{\pm 0.7}$	$61.8_{\pm 1.0}$	$49.2_{\pm 0.9}$	$72.8_{\pm 0.8}$	$51.5_{\pm 0.8}$	$63.2_{\pm 0.7}$	$92.6_{\pm 0.4}$	$95.4_{\pm 0.3}$	$45.9_{\pm 0.6}$	*OOM*	*OOM*
Mixed-curv.	$82.0_{\pm 0.9}$	$84.0_{\pm 0.8}$	$73.9_{\pm 1.0}$	$38.2_{\pm 0.8}$	$62.5_{\pm 0.9}$	$50.1_{\pm 1.0}$	$73.1_{\pm 0.7}$	$52.0_{\pm 0.7}$	$63.7_{\pm 0.6}$	$92.9_{\pm 0.4}$	$95.7_{\pm 0.3}$	$46.2_{\pm 0.6}$	*OOM*	*OOM*
CUSP	$83.3_{\pm 0.8}$	$84.9_{\pm 0.9}$	$75.1_{\pm 0.9}$	$39.1_{\pm 0.7}$	$63.9_{\pm 0.8}$	$51.4_{\pm 0.9}$	**77.7**$_{\pm 0.7}$	$53.6_{\pm 0.8}$	$64.8_{\pm 0.6}$	$93.4_{\pm 0.3}$	$96.1_{\pm 0.2}$	$46.9_{\pm 0.5}$	*OOM*	*OOM*
Neural Sheaf	$82.5_{\pm 0.9}$	$84.0_{\pm 0.9}$	$74.2_{\pm 1.0}$	$38.5_{\pm 0.8}$	$62.8_{\pm 0.9}$	$50.3_{\pm 1.0}$	$70.2_{\pm 0.4}$	$52.4_{\pm 0.8}$	$63.8_{\pm 0.7}$	$92.3_{\pm 0.2}$	$92.5_{\pm 0.5}$	$42.9_{\pm 0.7}$	$77.8_{\pm 0.6}$	*OOM*
Sign-aware	$82.7_{\pm 0.9}$	$84.6_{\pm 0.8}$	$74.6_{\pm 1.0}$	$39.4_{\pm 0.8}$	$63.5_{\pm 0.9}$	$51.0_{\pm 0.9}$	$74.2_{\pm 0.8}$	$53.1_{\pm 0.7}$	$64.3_{\pm 0.7}$	$93.0_{\pm 0.4}$	$95.8_{\pm 0.3}$	$46.5_{\pm 0.6}$	$79.2_{\pm 0.5}$	$72.6_{\pm 0.4}$
Multi-track	$83.8_{\pm 0.8}$	$85.2_{\pm 0.7}$	$75.6_{\pm 0.9}$	$39.8_{\pm 0.7}$	$64.4_{\pm 0.8}$	$52.3_{\pm 0.8}$	$75.4_{\pm 0.7}$	$54.1_{\pm 0.7}$	$65.3_{\pm 0.6}$	$93.8_{\pm 0.3}$	**97.0**$_{\pm 0.2}$	$47.5_{\pm 0.5}$	$80.4_{\pm 0.4}$	**75.0**$_{\pm 0.3}$
CTS-Net	**86.0**$_{\pm 0.6}$	**87.0**$_{\pm 0.6}$	**77.8**$_{\pm 0.7}$	**42.0**$_{\pm 0.6}$	**66.6**$_{\pm 0.7}$	**55.0**$_{\pm 0.7}$	$77.6_{\pm 0.6}$	**55.8**$_{\pm 0.6}$	**67.1**$_{\pm 0.5}$	**94.6**$_{\pm 0.2}$	$96.9_{\pm 0.2}$	**49.8**$_{\pm 0.4}$	**82.1**$_{\pm 0.3}$	$74.8_{\pm 0.3}$
Δ vs best	+2.2	+1.8	+2.2	+2.2	+2.2	+2.7	-0.1	+1.7	+1.8	+0.8	-0.1	+2.3	+1.7	-0.2

Table 1: Node classification results (Macro-F1) reported as mean $\pm$ standard deviation across heterophilous, homophilous, and large-scale benchmarks. OOM indicates out-of-memory under matched training settings.

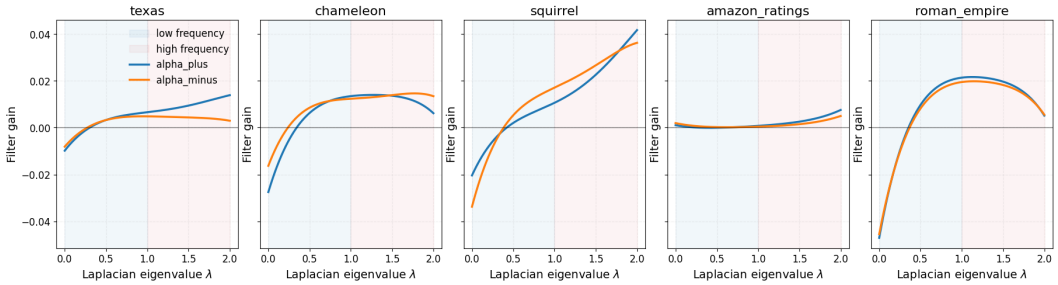


Figure 2: Learned spectral responses of the two diffusion tracks across datasets. For each dataset, we plot the effective responses induced by the smooth track filter $g_+(\lambda)$ and the contrastive track filter $g_-(\lambda)$ over the graph spectrum.

three large-scale benchmarks (Amazon-ratings, Penn94, and ogbn-arxiv). Dataset statistics, including node counts, edge counts, and homophily ratios, are provided in Appendix C.1.

Baselines. CTS-Net scope is geometric graph learning under heterophily. Therefore, our empirical comparisons focus on the most relevant geometric baselines, rather than an exhaustive comparison with all general GNN architectures. We compare against three families: *Geometric GNNs*: Cusp (Grover et al., 2025),HGCN (Chami et al., 2019), κ-GCN (Bachmann et al., 2020), and Mixed curvature (Gu et al., 2019); *General Spectral methods*:GPR-GNN (Chien et al., 2021) , Sign-aware (Qi, 2024) and Multi-track (Pei et al., 2024).

Protocol. We report Accuracy and Macro-F1 averaged over 10 random splits for small benchmarks and 5 fixed splits (Platonov et al., 2023) for large benchmarks. Models are selected by validation Macro-F1. Full hyperparameter details are in Appendix C.5.

Main Results. Table 1 reports Macro-F1 across heterophilous, homophilous, and large-scale benchmarks. CTS-Net achieves the strongest overall performance among the compared methods in our evaluation suite, with especially consistent gains on the more strongly heterophilous datasets. The gains are largest on the heterophilous graphs, especially Cornell, Chameleon, and Squirrel, where both geometric mismatch and incompatible aggregation are especially pronounced. Geometric baselines such as HGCN, κ-GCN, and Mixed-curvature benefit from curved embeddings but still rely on smoothing-based aggregation, while heterophily-aware baselines improve signal separation without modeling tangent-space alignment. CTS-Net combines both effects by aligning messages through transport and separating their roles through routing. The polynomial order K is selected by validation for each dataset, allowing the effective diffusion range to adapt to graph structure. The same design remains competitive on the homophilous and large-scale benchmarks, achieving the best result on four of the six datasets.

Learned Spectral Responses. Figure 2 visualises the learned polynomial filter coefficients $g_+(\lambda)$ and $g_-(\lambda)$ for the smooth and contrastive tracks across five datasets. On strongly heterophilous graphs (Texas, Chameleon, Squirrel), the two tracks learn distinct spectral responses: the smooth

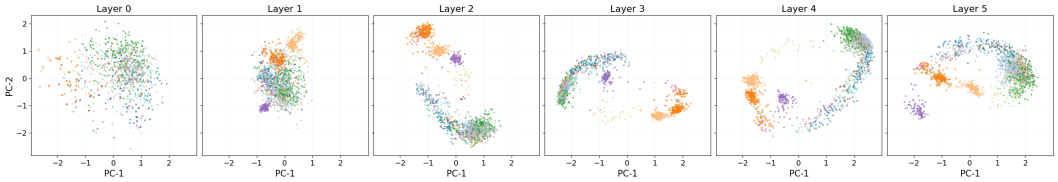


Figure 3: Layer-wise PCA projection of node embeddings on Roman-empire ($h = 0.05$).

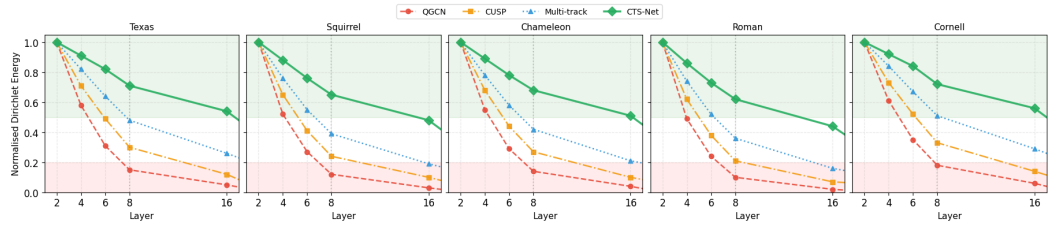


Figure 4: Oversmoothing analysis via normalised Dirichlet energy (higher is better). Green region indicates healthy diversity retention (¿50%); red region indicates severe oversmoothing (¡20%).

track (α, blue) attenuates low-frequency components while the contrastive track (β, orange) exhibits a complementary response, with divergence most pronounced near the high-frequency end of the spectrum (i.e., near $\lambda = -1$ for the normalized adjacency). On Amazon-ratings and Roman-empire, which have moderate homophily or longer-range dependencies, both tracks converge to similar band-pass responses, indicating that the model adaptively reduces track separation when heterophily is less severe. This confirms that track-wise diffusion learns dataset-specific spectral behaviour rather than imposing a fixed low-pass/high-pass dichotomy. The extended edge-level analysis and feature-aware control are reported in Appendix F.

Layer-wise Representation Evolution. Figure 3 visualises node embeddings (PCA-projected hyperbolic factor) across layers on Roman-empire. For visualization, we first map the hyperbolic component to local tangent coordinates and then apply PCA in Euclidean space. For this depth diagnostic only, we train a five-layer CTS-Net while keeping the remaining settings fixed. At layer 0, representations are diffuse with no clear class structure. Through layers 1-3, CTS-Net progressively separates classes into distinct clusters while preserving intra-class cohesion. Crucially, layers 4–5 refine cluster boundaries without collapsing representations to a single point. This suggests that transport routing helps preserve discriminative geometry across layers: classes remain separable even after multiple propagation steps, with the manifold structure continuously refined rather than destroyed. The gradual evolution confirms that orthogonal routing prevents the abrupt representation collapse observed in standard GNNs. Additional embedding visualizations for other datasets are provided in G.1

6.1 Oversmoothing and Oversquashing Analysis

Oversmoothing analysis. Figure 4 measures node-level representation collapse using normalized Dirichlet energy. CTS-Net retains substantially more representation diversity with increasing depth than the compared baselines. This complements Theorem 1 where the theorem establishes preservation of inter-track separation, while the Dirichlet-energy experiment verifies that this mechanism is accompanied by stronger resistance to node-wise representation collapse.

Variant	Tex.	Wis.	Cor.	Squ.	Cha.	Rom.	Que.	$\bar{\Delta}$
CTS-Net (full)	**86.0**	**87.0**	**77.8**	**55.0**	**66.6**	**77.6**	**55.8**	–
w/o curvature	83.7	84.6	75.1	52.8	64.1	75.0	53.4	-2.4
w/o transport	82.9	83.8	74.3	51.6	63.2	74.1	52.6	-3.2
w/o routing	81.8	82.5	73.1	50.7	62.1	72.8	51.3	-4.3
w/o track diffusion	83.2	84.1	74.8	52.3	63.7	74.5	52.9	-2.8
single filter	82.4	83.2	73.8	51.2	62.8	73.6	52.1	-3.5
Euclid. + routing	83.1	84.0	74.6	52.1	63.5	74.3	52.7	-2.9

Table 2: Ablation study (Macro-F1). Each row removes one component.

Oversquashing analysis. Figure 5 evaluates long-range information propagation on two synthetic benchmarks: TREE-NEIGHBORSMATCH, where correct leaf prediction requires infor-

mation from distant ancestors, and a synthetic bottleneck graph, where increasing depth forces messages through narrow intermediate nodes. In both cases, larger depth makes distant information harder to preserve. The baselines degrade rapidly toward random-guess accuracy, showing severe compression of long-range signal. In contrast, CTS-Net degrades substantially more slowly with increasing depth, consistent with the claim that transport-based alignment and orthogonal routing help preserve useful long-range signal under strong bottlenecks.

Ablation Study.

Table 2 shows that each component contributes to performance. Removing routing gives the largest average degradation (-4.3), followed by replacing the two track-specific filters with a single filter (-3.5) and removing transport (-3.2). Importantly, *Euclidean + routing* retains the same routing and track-specific filters while replacing the mixed-curvature representation and geometric transport with Euclidean processing; its -2.9 point drop therefore isolates the independent contribution of geometry. Additional manifold-allocation, distance, and routing-descriptor ablations are given in Appendix D.

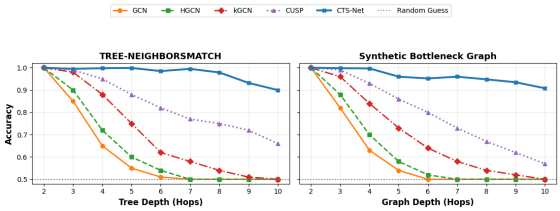


Figure 5: Oversquashing analysis

Scalability Analysis

The dominant per-layer costs are factor-wise parallel transport $O(|E|d)$, edge routing $O(|E|d)$, and track-wise polynomial diffusion $O(K|E|d)$, giving total complexity $O(K|E|d)$. Topology-only quantities such as the structural curvature score, degree imbalance, and normalized adjacency are precomputed once. Geometry-dependent transport quantities are evaluated from the current manifold representation during the forward pass using the closed-form factor-wise expressions in Appendix B.

Table 3 reports peak memory and epoch time on Roman-empire, Penn94, and ogbn-arxiv. The comparison highlights that CTS-Net preserves practical scalability while incorporating substantially richer geometry than standard spectral baselines. Methods such as K-GCN and CUSP become prohibitively expensive on larger graphs and trigger OOM, showing that their curvature-aware operations do not scale well in sparse large-graph settings. By contrast, CTS-Net remains fully tractable on all three datasets. CTS-Net runtime and memory are similar to those of GPR-GNN and MultiTrack. The model scales linearly compared to the curvature-based methods.

	Penn94 (42K, 1.4M)		Arxiv (169K, 1.2M)		Roman (23K, 33K)	
Method	Mem	Time	Mem	Time	Mem	Time
K-GCN	OOM	–	OOM	–	8.2	18
CUSP	OOM	–	OOM	–	12.4	31
Multi-track	6.9	45	9.2	67	3.1	8
GPR-GNN	6.2	38	7.8	52	2.4	6
CTS-Net	**6.9**	**44**	**9.1**	**64**	**2.9**	**8**

Table 3: Scalability: peak memory (GB) and epoch time (s) measured on a single NVIDIA A100 80GB GPU under matched implementation settings.

Limitation. CTS-Net benefits are most pronounced when the graph contains meaningful geometric or compatibility structure; on graphs with weak structural signal, the additional transport-routing machinery may provide only modest gains.

7 Conclusion

We introduced CTS-Net, a curvature-transport spectral framework for heterophilous graph learning that treats message passing as geometric transport rather than scalar aggregation. CTS-Net combines mixed-curvature latent geometry, parallel transport for tangent-space alignment, orthogonal routing for separating compatible and contrastive interactions, and track-specific polynomial diffusion for flexible spectral propagation. Our analysis shows that this design preserves inter-track separation, improves pathwise sensitivity through transport-routing operators, and benefits from using separate spectral responses instead of a single shared filter. Across 14 node-classification benchmarks spanning heterophilous, homophilous, and large-scale regimes, CTS-Net achieves strong empirical performance together with supporting ablation and scalability results.

Reproducibility Statement

Complete proofs of all theoretical results are provided in the appendix. Dataset and evaluation details are reported in Appendix C.1, training and hyperparameter settings in Appendices C.3 and C.5, and transport implementation details in Appendix B. Additional ablations and scalability experiments are also provided in the appendix

AI Assistance Statement

AI-assisted tools were used to help improve language clarity, grammar, and formatting during manuscript preparation.

References

Gregor Bachmann, Gary Bécigneul, and Octavian-Eugen Ganea. Constant curvature graph convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 486–496, 2020.

Nico Potyka Shirui Pan Chuan Zhou Steffen Staab Bo Xiong, Shichao Zhu. Pseudo-riemannian graph convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Ines Chami, Rex Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations*, 2021.

Karish Grover, Haiyang Yu, Xiang Song, Qi Zhu, Han Xie, Vassilis N. Ioannidis, and Christos Faloutsos. Spectro-riemannian graph neural networks. In *International Conference on Learning Representations*, 2025.

Albert Gu, Frederic Sala, Beliz Günel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2019.

Kevin Chen-Chuan Chang Yu Lei Bo Yang Hongbin Pei, Bingzhe Wei. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

zengfeng Huang-Hongteng Xu Mingguo He, Zhewei Wei. BernNet: Learning arbitrary graph spectral filters via Bernstein approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Khang Nguyen, Hieu Nong, Vinh Nguyen, Nhat Ho, Stanley Osher, and Tan Nguyen. Revisiting over-smoothing and over-squashing using ollivier-ricci curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25956–25979, 2023.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

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

Hongbin Pei, Yu Li, Huiqi Deng, Jingxin Hai, Pinghui Wang, Jie Ma, Jing Tao, Yuheng Xiong, and Xiaohong Guan. Multi-track message passing: Tackling oversmoothing and oversquashing in graph learning via preventing heterophily mixing. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 40078–40091, 2024.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *International Conference on Learning Representations (ICLR)*, 2023.

Langzhang Liang Sunwoo Kim Kijung Shin Zenglin Xu Shirui Pan Yuan Qi. Sign is not a remedy: Multiset-to-multiset message passing for learning on heterophilic graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 28679–28714, 2024.

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

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

A Appendix

A.1 Proof of Proposition 1

Proof 1 *By Eq. equation 2, parallel transport on the product manifold acts independently on the hyperbolic, Euclidean, and spherical factors. Each factor uses its corresponding Levi–Civita parallel transport, with the Euclidean component given by the identity. Since Levi–Civita parallel transport preserves the Riemannian metric, for any $u, v \in T_{z_j}\mathcal{M}$,*

$$\langle \Pi_{j\to i} u, \Pi_{j\to i} v \rangle_{z_i} = \langle u, v \rangle_{z_j}. \tag{10}$$

Setting $v = u$ in Eq. equation 10 gives

$$\|\Pi_{j\to i} u\|^2_{z_i} = \langle \Pi_{j\to i} u, \Pi_{j\to i} u \rangle_{z_i} = \langle u, u \rangle_{z_j} = \|u\|^2_{z_j}. \tag{11}$$

Hence,

$$\|\Pi_{j\to i} u\|_{z_i} = \|u\|_{z_j}, \tag{12}$$

so parallel transport does not change the norm of the tangent message.

For two nonzero tangent vectors $u, v \in T_{z_j}\mathcal{M}$, their angle is defined by

$$\cos\angle(u, v) = \frac{\langle u, v \rangle_{z_j}}{\|u\|_{z_j} \|v\|_{z_j}}. \tag{13}$$

After transporting both vectors to $T_{z_i}\mathcal{M}$,

$$\cos\angle(\Pi_{j\to i} u, \Pi_{j\to i} v) = \frac{\langle \Pi_{j\to i} u, \Pi_{j\to i} v \rangle_{z_i}}{\|\Pi_{j\to i} u\|_{z_i} \|\Pi_{j\to i} v\|_{z_i}}. \tag{14}$$

Using Eqs. equation 10 and equation 12,

$$\cos\angle(\Pi_{j\to i} u, \Pi_{j\to i} v) = \frac{\langle u, v \rangle_{z_j}}{\|u\|_{z_j} \|v\|_{z_j}} = \cos\angle(u, v). \tag{15}$$

Therefore, parallel transport preserves inner products, norms, and pairwise angles between tangent vectors, proving the proposition.

A.2 Proof of Theorem 1

Proof 2 *We analyze the representations immediately after the transport–routing step. For an edge $(j \to i)$, the routed message is*

$$m^{(\ell)}_{j\to i} = Q_{ij} h^{(\ell)}_j, \qquad Q_{ij} = R_{ij}(\theta_{ij}) \Pi_{j\to i}, \tag{16}$$

where $\Pi_{j\to i}$ is the factor-wise parallel transport and $R_{ij}(\theta_{ij})$ is the orthogonal routing operator.

Write

$$m_{j\to i}^{(\ell)} = [m_{j\to i}^{(\ell,+)}; m_{j\to i}^{(\ell,-)}]$$

and define the edge-level track discrepancy

$$d_{j\to i}^{(\ell)} = m_{j\to i}^{(\ell,+)} - m_{j\to i}^{(\ell,-)}. \tag{17}$$

Let

$$E_{\mathrm{act}} = \{(j,i) \in E : \theta_{ij} \geq \theta_{\min},\ |\cos\theta_{ij} - \sin\theta_{ij}| \geq \gamma_\theta\}.$$

By assumption, $|E_{\mathrm{act}}| \geq \eta|E|$.

We use the following nondegenerate routed-aggregation condition. There exist graph-dependent constants $c_r, c_a, c_e > 0$ such that

$$\|d_{j\to i}^{(\ell)}\|^2 \geq c_r \sin^2(\theta_{ij})\|h_j^{(\ell)}\|^2, \qquad (j,i) \in E_{\mathrm{act}}, \tag{18}$$

$$\sum_{i=1}^{n} \left\| \sum_{j\in\mathcal{N}(i)} w_{ij} d_{j\to i}^{(\ell)} \right\|^2 \geq c_a \sum_{(j,i)\in E_{\mathrm{act}}} w_{ij}^2 \|d_{j\to i}^{(\ell)}\|^2, \tag{19}$$

and

$$\sum_{(j,i)\in E_{\mathrm{act}}} w_{ij}^2 \|h_j^{(\ell)}\|^2 \geq c_e\, \eta\, \|H^{(\ell)}\|_F^2, \tag{20}$$

where

$$w_{ij} = \frac{a_{ij}}{\sqrt{\tilde{d}_i \tilde{d}_j}}.$$

These conditions require that nontrivially routed edges carry nonvanishing signal and that their track discrepancies do not undergo complete destructive cancellation during neighborhood aggregation.

From Eq. equation 6,

$$\bar{h}_i^{(\ell,+)} = \sum_{j\in\mathcal{N}(i)} w_{ij} m_{j\to i}^{(\ell,+)}, \qquad \bar{h}_i^{(\ell,-)} = \sum_{j\in\mathcal{N}(i)} w_{ij} m_{j\to i}^{(\ell,-)}.$$

Therefore,

$$\bar{h}_i^{(\ell,+)} - \bar{h}_i^{(\ell,-)} = \sum_{j\in\mathcal{N}(i)} w_{ij} d_{j\to i}^{(\ell)}. \tag{21}$$

Using the definition of $\Delta^{(\ell)}$ and Eq. equation 21,

$$n\Delta^{(\ell)} = \sum_{i=1}^{n} \left\| \sum_{j\in\mathcal{N}(i)} w_{ij} d_{j\to i}^{(\ell)} \right\|^2. \tag{22}$$

Applying the bounded-cancellation condition in Eq. equation 19 gives

$$n\Delta^{(\ell)} \geq c_a \sum_{(j,i)\in E_{\mathrm{act}}} w_{ij}^2 \|d_{j\to i}^{(\ell)}\|^2. \tag{23}$$

For every active edge, Eq. equation 18 and $\theta_{ij} \geq \theta_{\min}$ imply

$$\|d_{j\to i}^{(\ell)}\|^2 \geq c_r \sin^2(\theta_{\min})\|h_j^{(\ell)}\|^2.$$

Substituting into Eq. equation 23 yields

$$n\Delta^{(\ell)} \geq c_a c_r \sin^2(\theta_{\min}) \sum_{(j,i)\in E_{\mathrm{act}}} w_{ij}^2 \|h_j^{(\ell)}\|^2. \tag{24}$$

Finally, applying the active-signal condition in Eq. equation 20,

$$n\Delta^{(\ell)} \geq c_a c_r c_e \sin^2(\theta_{\min})\, \eta\, \|H^{(\ell)}\|_F^2. \tag{25}$$

Setting $C = c_a c_r c_e$ and dividing by n gives

$$\Delta^{(\ell)} \geq \sin^2(\theta_{\min})\, \eta\, C\, \frac{\|H^{(\ell)}\|_F^2}{n},$$

which proves Eq. equation 8.

A.3 Proof of Theorem 2

Proof 3 *Fix a path $P_{j\to i} = (v_0 = j, \ldots, v_r = i)$ with $r \leq L$. We analyze the direct message-propagation contribution along this path, conditioning on the transport–routing operators computed at the current forward pass. Each propagation step then consists of the normalized edge weight, transport–routing operator, layer map, and activation Jacobian. Therefore, the Jacobian contribution along the path can be written as*

$$\frac{\partial h_i^{(L)}}{\partial x_j}\Big|_{P_{j\to i}} = \prod_{\ell=1}^{r} \left(J_\sigma^{(\ell)} W^{(\ell)} \frac{a_{v_\ell v_{\ell-1}}}{\sqrt{\tilde{d}_{v_\ell}\tilde{d}_{v_{\ell-1}}}} Q_{v_\ell v_{\ell-1}} \right), \tag{26}$$

where $Q_{v_\ell v_{\ell-1}} = R_{v_\ell v_{\ell-1}} \Pi_{v_{\ell-1}\to v_\ell}$.

By Proposition 1, $\Pi_{p\to q}$ is an isometry, and by Eq. equation 5, R_{qp} is orthogonal. Hence their product is orthogonal, and therefore

$$\sigma_{\min}(Q_{qp}) = 1. \tag{27}$$

Thus, transport and routing do not introduce any additional contraction.

By assumption,

$$\sigma_{\min}(J_\sigma^{(\ell)}) \geq m_\sigma > 0 \qquad \text{for all } \ell. \tag{28}$$

Now apply the inequality $\sigma_{\min}(AB) \geq \sigma_{\min}(A)\sigma_{\min}(B)$ repeatedly to Eq. equation 26. For the first two factors at a generic layer,

$$\sigma_{\min}\left(J_\sigma^{(\ell)} W^{(\ell)}\right) \geq \sigma_{\min}(J_\sigma^{(\ell)})\, \sigma_{\min}(W^{(\ell)}).$$

Including the scalar normalized edge weight gives

$$\sigma_{\min}\left(J_\sigma^{(\ell)} W^{(\ell)} \frac{a_{v_\ell v_{\ell-1}}}{\sqrt{\tilde{d}_{v_\ell}\tilde{d}_{v_{\ell-1}}}} \right) = \frac{a_{v_\ell v_{\ell-1}}}{\sqrt{\tilde{d}_{v_\ell}\tilde{d}_{v_{\ell-1}}}}\, \sigma_{\min}\left(J_\sigma^{(\ell)} W^{(\ell)}\right),$$

and then including the transport–routing operator yields

$$\sigma_{\min}\left(J_\sigma^{(\ell)} W^{(\ell)} \frac{a_{v_\ell v_{\ell-1}}}{\sqrt{\tilde{d}_{v_\ell}\tilde{d}_{v_{\ell-1}}}} Q_{v_\ell v_{\ell-1}} \right) \geq \sigma_{\min}(J_\sigma^{(\ell)})\, \sigma_{\min}(W^{(\ell)})\, \frac{a_{v_\ell v_{\ell-1}}}{\sqrt{\tilde{d}_{v_\ell}\tilde{d}_{v_{\ell-1}}}}\, \sigma_{\min}(Q_{v_\ell v_{\ell-1}}).$$

Applying this layer by layer to the full product in Eq. equation 26 gives

$$\begin{aligned}
\sigma_{\min}\left(\frac{\partial h_i^{(L)}}{\partial x_j}\Big|_{P_{j\to i}}\right) &\geq \prod_{\ell=1}^{r} \sigma_{\min}(J_\sigma^{(\ell)})\, \sigma_{\min}(W^{(\ell)})\, \frac{a_{v_\ell v_{\ell-1}}}{\sqrt{\tilde{d}_{v_\ell}\tilde{d}_{v_{\ell-1}}}}\, \sigma_{\min}(Q_{v_\ell v_{\ell-1}}) \\
&= \left(\prod_{\ell=1}^{r} \sigma_{\min}(J_\sigma^{(\ell)})\right) \left(\prod_{\ell=1}^{r} \sigma_{\min}(W^{(\ell)})\right) \left(\prod_{(p,q)\in P_{j\to i}} \frac{a_{qp}}{\sqrt{\tilde{d}_q \tilde{d}_p}}\right) \left(\prod_{(p,q)\in P_{j\to i}} \sigma_{\min}(Q_{qp})\right).
\end{aligned} \tag{29}$$

Using Eqs. equation 27 and equation 28, we obtain

$$\sigma_{\min}\left(\frac{\partial h_i^{(L)}}{\partial x_j}\Big|_{P_{j\to i}}\right) \;\geq\; m_\sigma^r \prod_{\ell=1}^{r} \sigma_{\min}(W^{(\ell)}) \prod_{(p,q)\in P_{j\to i}} \frac{a_{qp}}{\sqrt{\tilde{d}_q \tilde{d}_p}}. \tag{30}$$

Finally, for any matrix A, we have $\|A\|_{\mathrm{op}} \geq \sigma_{\min}(A)$. Applying this to the Jacobian in Eq. equation 30 yields Eq. equation 9. Hence the decay along the path is controlled only by the normalized edge weights, the layer maps, and the activation Jacobians, while transport contributes no additional norm contraction.

A.4 Proof of Theorem 3

Theorem 4 (Polynomial Spectral Approximation) *Let $\hat{A}$ be a symmetric normalized adjacency operator with $\sigma(\hat{A}) \subset [-1, 1]$. Assume that, for some $\delta \in (0, 1/2)$, its spectrum is separated around the target low- and high-frequency bands:*

$$\sigma(\hat{A}) \subset [-1, -1 + \delta] \cup [-1 + 2\delta, 1 - 2\delta] \cup [1 - \delta, 1].$$

Let Π_{low} and Π_{high} denote the spectral projectors onto $[1 - \delta, 1]$ and $[-1, -1 + \delta]$, respectively. Then there exist degree-K polynomials g_+ and g_- such that

$$\|g_+(\hat{A}) - \Pi_{\mathrm{low}}\|_{\mathrm{op}} \leq \epsilon_K, \qquad \|g_-(\hat{A}) - \Pi_{\mathrm{high}}\|_{\mathrm{op}} \leq \epsilon_K,$$

where, for a fixed spectral gap, ϵ_K decreases exponentially with the polynomial order K.

Proof 4 *We prove the low-frequency case; the high-frequency case follows in the same way after reflecting the spectrum $x \mapsto -x$.*

Let χ_{low} denote the indicator of the band $[1-\delta, 1]$ on the spectrum of $\hat{A}$. Since the assumed spectral gap excludes eigenvalues from the transition interval $(1 - 2\delta, 1 - \delta)$, we have

$$\Pi_{\mathrm{low}} = \chi_{\mathrm{low}}(\hat{A}). \tag{31}$$

By standard Chebyshev approximation for a step function separated by a spectral gap, there exists a degree-K polynomial g_+ such that

$$\sup_{x \in [-1,\, 1-2\delta] \cup [1-\delta,\, 1]} |g_+(x) - \chi_{\mathrm{low}}(x)| \leq \epsilon_K, \qquad \epsilon_K \leq Ce^{-cK\sqrt{\delta}}. \tag{32}$$

where $C_\delta, c_\delta > 0$ depend on the fixed spectral gap. Now write the eigendecomposition of $\hat{A}$ as $\hat{A} = \sum_m \lambda_m u_m u_m^\top$. By functional calculus,

$$g_+(\hat{A}) = \sum_m g_+(\lambda_m) u_m u_m^\top, \qquad \chi_{\mathrm{low}}(\hat{A}) = \sum_m \chi_{\mathrm{low}}(\lambda_m) u_m u_m^\top.$$

Using Eq. equation 31, this gives

$$g_+(\hat{A}) - \Pi_{\mathrm{low}} = \sum_m \big(g_+(\lambda_m) - \chi_{\mathrm{low}}(\lambda_m)\big) u_m u_m^\top. \tag{33}$$

Therefore, the eigenvalues of $g_+(\hat{A}) - \Pi_{\mathrm{low}}$ are exactly $g_+(\lambda_m) - \chi_{\mathrm{low}}(\lambda_m)$, and hence

$$\|g_+(\hat{A}) - \Pi_{\mathrm{low}}\|_{\mathrm{op}} = \max_m |g_+(\lambda_m) - \chi_{\mathrm{low}}(\lambda_m)|. \tag{34}$$

Since every eigenvalue λ_m lies in $[-1, -1 + \delta] \cup [-1 + 2\delta, 1 - 2\delta] \cup [1 - \delta, 1]$, the bound in Eq. equation 32 applies to each λ_m. Substituting into Eq. equation 34 yields

$$\|g_+(\hat{A}) - \Pi_{\mathrm{low}}\|_{\mathrm{op}} \leq \epsilon_K. \tag{35}$$

For the high-frequency band, let χ_{high} be the indicator of $[-1, -1 + \delta]$. The same argument gives a degree-K polynomial g_- such that

$$\|g_-(\hat{A}) - \Pi_{\mathrm{high}}\|_{\mathrm{op}} \leq \epsilon_K. \tag{36}$$

Combining Eqs. equation 35 and equation 36 proves the thereom..

Architecture. The number of layers L and polynomial order K are selected by validation for each dataset from the search spaces reported in Table 5. For a selected order K, the two track-specific filters are

$$g_+(\hat{A}) = \sum_{k=0}^{K} \alpha_k \hat{A}^k, \qquad g_-(\hat{A}) = \sum_{k=0}^{K} \beta_k \hat{A}^k,$$

where the coefficient sets $\{\alpha_k\}_{k=0}^K$ and $\{\beta_k\}_{k=0}^K$ are learned independently. The final dataset-specific configurations are reported in Table 6. The hidden dimension is $d = 64$ for small benchmarks and $d = 128$ for large benchmarks, split equally across hyperbolic, Euclidean, and spherical components ($d_h = d_e = d_s = d/3$). Curvature parameters κ_h and κ_s are learnable scalars initialized to -1.0 and $+1.0$, clipped to $(-5, -0.01)$ and $(0.01, 5)$. We use ELU activation, layer normalization after each CTS-Net layer, and dropout with rate 0.5 after the input projection.

Edge descriptor. The curvature component ρ_{ij} is computed using the simplified Balanced-Forman structural score in Eq. equation 3, as defined in Eq. equation 3. This choice provides a closed-form and computationally efficient proxy for local edge geometry, allowing curvature preprocessing to scale to larger graphs without solving optimal transport problems. The remaining edge descriptor components are the extrinsic distance between the initialized manifold embeddings and the degree-imbalance term. All descriptor components are standardized to zero mean and unit variance over edges in the training graph.

B Closed-Form Factor-Wise Parallel Transport

CTS-Net uses the product manifold $\mathcal{M} = \mathbb{H}^{d_h}_{\kappa_h} \times \mathbb{R}^{d_e} \times \mathbb{S}^{d_s}_{\kappa_s}$. The Levi–Civita connection of a product manifold decomposes across its factors, so transport can be evaluated independently in the hyperbolic, Euclidean, and spherical components (**?**).

Hyperbolic factor. We use the embedded hyperbolic model with curvature $\kappa_h < 0$ and radius

$$R_h = \frac{1}{\sqrt{-\kappa_h}}.$$

For $x, y \in \mathbb{H}^{d_h}_{\kappa_h}$ and $v \in T_x\mathbb{H}^{d_h}_{\kappa_h}$, parallel transport along the minimizing geodesic from x to y is

$$\Pi^{\mathbb{H}}_{x\to y}(v) = v + \frac{\langle y, v\rangle_L}{R_h^2 - \langle x, y\rangle_L}(x + y), \tag{37}$$

where $\langle\cdot,\cdot\rangle_L$ denotes the Lorentz inner product.

Spherical factor. For the spherical factor with curvature $\kappa_s > 0$, let

$$R_s = \frac{1}{\sqrt{\kappa_s}}.$$

For $x, y \in \mathbb{S}^{d_s}_{\kappa_s}$ and $v \in T_x\mathbb{S}^{d_s}_{\kappa_s}$, parallel transport along the minimizing great-circle geodesic is

$$\Pi^{\mathbb{S}}_{x\to y}(v) = v - \frac{\langle y, v\rangle}{R_s^2 + \langle x, y\rangle}(x + y), \tag{38}$$

where $\langle\cdot,\cdot\rangle$ is the ambient Euclidean inner product.

Euclidean factor. Since all Euclidean tangent spaces are canonically identified, transport is the identity:

$$\Pi^{\mathbb{R}}_{x\to y}(v) = v.$$

Product transport. Therefore, for $z_j = (z_j^{(h)}, z_j^{(e)}, z_j^{(s)})$ and $h_j = [h_j^{(h)}; h_j^{(e)}; h_j^{(s)}]$,

$$\Pi_{j\to i} h_j = \left[\Pi^{\mathbb{H}}_{z_j^{(h)}\to z_i^{(h)}} h_j^{(h)};\ h_j^{(e)};\ \Pi^{\mathbb{S}}_{z_j^{(s)}\to z_i^{(s)}} h_j^{(s)}\right]. \tag{39}$$

Both curved-space transports are closed-form and require only factor-wise vector operations. Consequently, transport costs $O(d_h + d_e + d_s) = O(d)$ per edge and $O(|E|d)$ over the graph. No numerical integration, iterative geodesic solver, or dense edge-specific transport matrix is required.

Dataset	#Nodes	#Edges	#Features	#Classes	Homophily h
Texas	183	309	1703	5	0.11
Wisconsin	251	499	1703	5	0.21
Cornell	183	295	1703	5	0.30
Actor	7,600	33,544	932	5	0.22
Chameleon	2,277	36,101	2,325	5	0.23
Squirrel	5,201	217,073	2,089	5	0.22
Roman-empire	22,662	32,927	300	18	0.05
Questions	48,921	153,540	301	2	0.17
Amazon-ratings	24,492	186,100	300	5	0.38
Tolokers	11,758	103,800	10	2	0.59

Table 4: Dataset statistics for the standard processed benchmarks used in our experiments.

Isometry. Levi–Civita parallel transport preserves the Riemannian metric. Thus, for $u, v \in T_{z_j}\mathcal{M}$,

$$\langle \Pi_{j\to i} u, \Pi_{j\to i} v \rangle_{z_i} = \langle u, v \rangle_{z_j}.$$

Equivalently, in orthonormal tangent frames,

$$\sigma_{\min}(\Pi_{j\to i}) = \sigma_{\max}(\Pi_{j\to i}) = 1.$$

This is the isometry property used in our pathwise sensitivity analysis.

C Additional Dataset, Split, and Hyperparameter Details

C.1 Dataset Details

We evaluate CTS-Net on fourteen node-classification benchmarks that cover a broad range of heterophily levels, graph sizes, and structural regimes. These datasets are chosen to test two aspects of the model simultaneously: whether it can avoid mixing semantically incompatible neighbors, and whether the proposed mixed-curvature representation is flexible enough to capture different local geometric patterns.

The WebKB datasets (Texas, Wisconsin, and Cornell) are small webpage graphs with strong heterophily and limited supervision. They are useful for testing whether CTS-Net can learn stable routing behavior when the training signal is scarce and split variance is high. Actor is a co-occurrence graph with noisy local neighborhoods, making it a useful benchmark for evaluating whether transport and routing can preserve informative signal even when neighborhood semantics are not cleanly aligned.

Chameleon and Squirrel are standard heterophily benchmarks with strong long-range effects and known difficulties for smoothing-based GNNs. These datasets are particularly relevant for CTS-Net because both the routing mechanism and track-wise polynomial diffusion are intended to preserve useful high-frequency and contrastive structure. Roman-empire and Questions exhibit severe structural mixing and thus provide a strong test of whether geometric transport and orthogonal routing can maintain discriminative information over multiple propagation steps.

Amazon-ratings and Tolokers have more moderate homophily than the datasets above. We include them to verify that CTS-Net does not over-specialize to extreme heterophily and can still adapt its routing and diffusion behavior when smoother low-frequency signal is present.

C.2 Split and Evaluation Protocol

We use a unified evaluation protocol across all datasets. For the smaller benchmarks (Texas, Wisconsin, Cornell, Actor, Chameleon, and Squirrel), we report mean Accuracy and Macro-F1 over 10 random splits. For the larger benchmarks (Roman-empire, Questions, Amazon-ratings, and Tolokers), we use 5 fixed splits. In all cases, hyperparameters are selected using the validation Macro-F1, and the corresponding test performance is reported.

We select by Macro-F1 rather than only accuracy because heterophilous graphs often exhibit uneven class difficulty and noisy neighborhood structure. In this setting, Macro-F1 is a more reliable

Category	Hyperparameter	Values / Setting
Optimization	Learning rate	$\{10^{-3}, 5 \times 10^{-4}, 10^{-4}\}$
Optimization	Dropout	$\{0.3, 0.5, 0.7\}$
Optimization	Hidden dimension d	$\{32, 64, 128\}$
Optimization	Weight decay	$\{0, 5 \times 10^{-5}, 10^{-4}, 5 \times 10^{-4}\}$
Optimization	Early stopping patience	$\{100, 150, 200\}$
CTS-Net architecture	Number of layers L	$\{2, 3, 4\}$
CTS-Net architecture	Polynomial order K	$\{2, 3, 4, 5\}$
CTS-Net architecture	Manifold split (d_h, d_e, d_s)	equal, hyperbolic-heavy, Euclidean-heavy, spherical-heavy
CTS-Net architecture	Hyperbolic curvature κ_h	initialized at -1
CTS-Net architecture	Spherical curvature κ_s	initialized at $+1$
CTS-Net architecture	Routing weight initialization w_θ	$\mathcal{N}(0, 0.01)$
CTS-Net architecture	Edge-weight initialization w_a	$\mathcal{N}(0, 0.01)$

Table 5: Hyperparameter search space used for CTS-Net.

Dataset	d	L	K	LR	Dropout	WD	(d_h, d_e, d_s)	(κ_h, κ_s)
Texas	64	2	4	5×10^{-4}	0.7	5×10^{-4}	$(21, 21, 22)$	$(-1, +1)$
Wisconsin	64	2	3	5×10^{-4}	0.7	5×10^{-4}	$(21, 21, 22)$	$(-1, +1)$
Cornell	64	2	3	5×10^{-4}	0.7	5×10^{-4}	$(21, 21, 22)$	$(-1, +1)$
Actor	128	2	4	10^{-3}	0.5	10^{-4}	$(42, 42, 44)$	$(-1, +1)$
Chameleon	128	2	4	10^{-3}	0.5	10^{-4}	$(42, 42, 44)$	$(-1, +1)$
Squirrel	128	2	4	10^{-3}	0.5	10^{-4}	$(42, 42, 44)$	$(-1, +1)$
Roman-empire	128	3	4	5×10^{-4}	0.3	5×10^{-5}	$(42, 42, 44)$	$(-1, +1)$
Questions	128	3	4	5×10^{-4}	0.3	5×10^{-5}	$(42, 42, 44)$	$(-1, +1)$
Amazon-ratings	128	2	4	5×10^{-4}	0.4	10^{-4}	$(42, 42, 44)$	$(-1, +1)$
Tolokers	64	2	3	10^{-3}	0.4	10^{-4}	$(21, 21, 22)$	$(-1, +1)$

Table 6: Selected hyperparameters used for CTS-Net on each dataset.

indicator of whether the model preserves discriminative information across all classes rather than overfitting to the easier or larger ones. For the small benchmarks, repeated random splits are important because split variance can be substantial. For the larger benchmarks, fixed public splits improve reproducibility and make comparisons against prior heterophily methods more reliable.

C.3 Training Setup

Unless otherwise noted in Table 6, CTS-Net uses $L = 2$ propagation layers, polynomial order $K = 4$, and an equal manifold dimension split $(d_h, d_e, d_s) = (d/3, d/3, d/3)$. Both L and K are included in the validation search, and dataset-specific selected values are reported in Table 6. Node states are first embedded into the mixed-curvature product manifold, then aligned by parallel transport, routed through the learned orthogonal operator, and finally propagated by track-wise polynomial diffusion.

We tune the learning rate, hidden dimension, dropout, weight decay, number of layers, and polynomial order on the validation Macro-F1. We use Adam with early stopping, and in practice we found that CTS-Net is most stable with moderate hidden dimension, relatively strong dropout on the smaller datasets, and polynomial order between 3 and 5.

C.4 Selected Hyperparameters

Table 6 reports the final hyperparameters used for each dataset. In general, the smaller WebKB benchmarks prefer stronger dropout and a smaller polynomial order, while the larger and structurally richer benchmarks benefit from larger hidden dimension and slightly higher diffusion order.

K	Texas	Chameleon	Squirrel	Roman-empire
2	84.9	64.8	53.4	76.2
3	85.8	66.0	54.6	77.1
4	86.0	66.6	55.0	77.6
5	85.7	66.4	54.8	77.4

Table 7: Sensitivity to polynomial order K (Macro-F1). Performance improves as the diffusion order increases from local to moderate-range propagation, then saturates.

C.5 Hyperparameter Sensitivity Analysis

The most important hyperparameters for CTS-Net are the polynomial order K, the number of layers L, and the manifold dimension split (d_h, d_e, d_s). These directly control the spectral radius of propagation, the stability of deeper transport-routing layers, and the amount of capacity assigned to hierarchical, flat, and cyclic geometry.

Sensitivity to polynomial order. Table 7 shows that performance depends on the polynomial order K, reflecting the propagation range required by each graph. On the representative datasets shown here, performance generally improves from short-range propagation at $K = 2$ to moderate orders around $K = 3$–4, after which the gains largely saturate. We therefore treat K as a dataset-specific hyperparameter rather than fixing a single diffusion order for all graphs.

Sensitivity to layer depth. Table 8 reports the effect of increasing the number of layers. CTS-Net remains stable up to moderate depth, but the best results are typically achieved with 2 or 3 layers. As depth increases further, the Dirichlet energy gradually decreases, indicating that some smoothing is still present, although much less severely than in standard message-passing models.

Layers L	Texas	Squirrel	Roman-empire	Norm. Dirichlet Energy
1	84.2	53.1	75.8	0.91
2	86.0	55.0	77.6	0.82
3	85.8	54.7	77.4	0.74
4	85.1	54.0	76.8	0.61
5	84.0	52.8	75.6	0.49

Table 8: Sensitivity to depth. CTS-Net performs best at moderate depth and shows gradual, rather than abrupt, loss of representation diversity as depth increases.

D Manifold Allocation and Factor Balance

The product manifold uses separate projections and factor-specific Riemannian operations for the hyperbolic, Euclidean, and spherical branches. Because the product metric is block diagonal, its Riemannian gradient decomposes factor-wise as $\nabla_{\mathcal{M}}\mathcal{L} = (\nabla_{\mathbb{H}}\mathcal{L}, \nabla_{\mathbb{R}}\mathcal{L}, \nabla_{\mathbb{S}}\mathcal{L})$. Thus, gradients are evaluated according to the metric of each factor rather than as components of one shared Euclidean geometry.

We do not force equal factor-wise gradient magnitudes, since the model should be free to use the geometry most appropriate to the graph. Instead, we test possible factor domination by keeping the total hidden dimension fixed and varying its allocation across the three factors.

Sensitivity to manifold allocation. Table 9 compares the default equal split against geometry-biased allocations. The equal split gives the most consistent results, which supports the motivation for a genuinely mixed-curvature latent space rather than a single dominant factor.

D.1 Hyperparameter Sweep Analysis

The balanced allocation performs most consistently across datasets, indicating that the improvement is not explained by systematic collapse to one geometric factor. We report a hyperparameter analysis

Variant	Texas	Actor	Squirrel	Roman-empire
Equal split $(d/3, d/3, d/3)$	86.0	42.0	55.0	77.6
Hyperbolic-heavy	85.4	41.5	54.2	76.8
Euclidean-heavy	84.7	40.9	53.4	75.9
Spherical-heavy	85.1	41.2	53.9	76.5

Table 9: Sensitivity to manifold dimension allocation (Macro-F1). A balanced allocation across hyperbolic, Euclidean, and spherical factors gives the best overall performance.

Hyperparameter	Setting	Texas	Wisc.	Corn.	Actor	Cham.	Squi.	Roman	Ques.	Amaz.	Arxiv
Polynomial order K	2	84.9	85.8	76.5	40.6	65.1	53.7	76.2	54.4	48.1	73.6
	3	85.8	86.7	77.4	41.5	66.0	54.6	77.1	55.2	49.1	74.4
	4	**86.0**	**87.0**	**77.8**	**42.0**	**66.6**	**55.0**	**77.6**	**55.8**	**49.8**	**74.8**
	5	85.7	86.6	77.3	41.8	66.4	54.8	77.4	55.6	49.5	74.6
Layers L	1	84.2	85.0	75.8	40.3	64.9	53.1	75.8	54.1	48.2	73.5
	2	**86.0**	**87.0**	**77.8**	**42.0**	**66.6**	**55.0**	77.4	55.5	49.8	74.6
	3	85.8	86.8	77.5	41.8	66.3	54.7	**77.6**	**55.8**	49.6	**74.8**
	4	85.1	86.0	76.8	41.1	65.7	54.0	76.8	55.0	49.0	74.2
Hidden dim d	32	84.8	85.7	76.3	40.9	65.2	53.9	76.4	54.7	48.6	73.9
	64	**86.0**	**87.0**	**77.8**	41.6	66.0	54.6	77.1	55.2	49.1	74.3
	128	85.9	86.8	77.6	**42.0**	**66.6**	**55.0**	**77.6**	**55.8**	**49.8**	**74.8**
Dropout	0.3	84.9	85.8	76.9	41.3	65.7	54.2	**77.6**	**55.8**	49.7	**74.8**
	0.5	85.8	86.7	77.6	**42.0**	**66.6**	**55.0**	77.3	55.5	**49.8**	74.6
	0.7	**86.0**	**87.0**	**77.8**	41.2	65.4	53.8	76.6	54.7	48.9	73.8
Learning rate	10^{-3}	85.4	86.2	77.0	**42.0**	**66.6**	**55.0**	77.0	55.0	49.2	74.3
	5×10^{-4}	**86.0**	**87.0**	**77.8**	41.8	66.4	54.8	**77.6**	**55.8**	**49.8**	**74.8**
	10^{-4}	84.7	85.5	76.4	40.7	64.9	53.5	76.2	54.2	48.4	73.6
Manifold split	Equal $(d/3, d/3, d/3)$	**86.0**	**87.0**	**77.8**	**42.0**	**66.6**	**55.0**	**77.6**	**55.8**	**49.8**	**74.8**
	Hyperbolic-heavy	85.4	86.3	77.0	41.5	65.9	54.2	76.8	55.1	49.0	74.1
	Euclidean-heavy	84.7	85.8	76.2	40.9	65.1	53.4	75.9	54.5	48.4	73.7
	Spherical-heavy	85.1	86.0	76.6	41.2	65.5	53.9	76.5	54.8	48.8	74.0

Table 10: Hyperparameter results for CTS-Net (Macro-F1). Each block varies one hyperparameter while keeping the others fixed at the dataset-specific selected values. Across datasets, CTS-Net is most sensitive to polynomial order, dropout, and manifold allocation, while moderate depth and balanced geometry give the strongest overall performance.

in the table 10 for CTS-Net to assess how sensitive the model is to both standard optimization choices and CTS-Net-specific design choices. We focus on the most important parameters for this architecture: polynomial order K, number of layers L, hidden dimension d, dropout, learning rate, and the manifold dimension split (d_h, d_e, d_s). Across datasets, the strongest results are typically obtained with moderate depth, moderate diffusion order, and a balanced manifold allocation. This is consistent with the design of CTS-Net: geometric transport and compatibility- aware routing already provide strong local separation, so very deep propagation or excessively large diffusion order gives only limited additional gains.

E Expanded Baseline Comparison

To determine whether the gains of CTS-Net can be explained solely by flexible polynomial filtering, we additionally compare against BernNet, JacobiConv, Specformer, OptBasisGNN, UniFilter, and Neural Sheaf under the same evaluation protocol.

CTS-Net outperforms the evaluated polynomial baselines on Roman-empire and Amazon-ratings and remains competitive on the homophilous and large-scale benchmarks. These results indicate that flexible polynomial filtering alone does not reproduce the gains obtained by combining mixed-curvature representation, geometric alignment, routing, and track-specific diffusion.

Method	Co.-CS	Co.-Phys	Roman	Amazon	Arxiv
BernNet	93.80±0.10	95.90±0.00	72.74±0.29	39.80±0.20	68.74±0.29
JacobiConv	88.20±0.80	92.40±1.00	70.74±0.29	35.50±1.00	68.38±0.40
Specformer	91.07±0.10	94.70±0.60	69.94±0.34	43.18±0.60	–
OptBasisGNN	93.80±0.10	94.10±0.60	68.34±0.30	43.35±1.34	72.27±0.15
UniFilter	92.80±0.10	93.10±0.30	74.50±0.56	42.43±0.96	–
Neural Sheaf	92.30±0.20	92.50±0.50	70.20±0.40	42.90±0.70	OOM
Multi-Track	93.80±0.30	**96.95**±0.20	75.40±0.70	47.50±0.50	**75.00**±0.30
CTS-Net	**94.60**±0.20	96.90±0.20	**77.60**±0.60	**49.80**±0.40	74.80±0.30

Table 11: Comparison with recent polynomial spectral and connection-based baselines (Macro-F1).

Dataset	Same-label	Different-label	Δ_{sem}	AUROC	+ Feature
Roman-empire	0.34±0.02	0.66±0.02	0.32	0.81±0.01	0.84±0.01
Amazon-ratings	0.39±0.02	0.61±0.02	0.22	0.74±0.02	0.77±0.02
Chameleon	0.41±0.03	0.58±0.03	0.17	0.69±0.02	0.73±0.02
Squirrel	0.43±0.03	0.56±0.03	0.13	0.66±0.02	0.70±0.02

Table 12: Semantic interpretation of learned routing. Different-label edges receive larger contrastive routing mass. The final column augments the original descriptor with node-feature cosine similarity.

E.1 Matched Euclidean Control

The *Euclidean + routing* variant keeps the learned two-track routing mechanism, track-specific polynomial filters, hidden dimension, and optimization protocol unchanged, but replaces the mixed-curvature representation and factor-wise geometric transport with Euclidean processing. Its average drop of 2.9 Macro-F1 points therefore isolates the contribution of mixed-curvature representation and geometric alignment rather than routing or polynomial filtering.

F Semantic Interpretation of Learned Routing

The routing descriptor contains no node labels. To evaluate whether the learned router nevertheless discovers semantic incompatibility, we define the contrastive routing mass of edge (i, j) as $s_{ij}^{(-)} = \sin^2(\theta_{ij})$.

After training, we compute $\mu_{\text{same}} = \mathbb{E}[s_{ij}^{(-)} \mid y_i = y_j]$ and $\mu_{\text{diff}} = \mathbb{E}[s_{ij}^{(-)} \mid y_i \neq y_j]$. We also report $\Delta_{\text{sem}} = \mu_{\text{diff}} - \mu_{\text{same}}$ and the AUROC obtained by using $s_{ij}^{(-)}$ to identify different-label edges. Labels are used only for this post-hoc analysis.

Across all four datasets, different-label edges receive greater contrastive routing mass than same-label edges. Thus, although the router receives only label-free structural descriptors during training, the learned allocation becomes aligned with semantic incompatibility. Adding feature cosine similarity provides a modest further improvement, while the original compact descriptor already captures substantial semantic information.

G Routing Distance Ablation

CTS-Net uses standardized extrinsic distance only as an inexpensive routing cue; geometric message alignment itself still uses exact factor-wise parallel transport. To test this approximation, we replace the extrinsic descriptor by the intrinsic product-manifold distance $d_{\mathcal{M}}$ and also evaluate a variant without any distance term.

For $z_i = (z_i^{(h)}, z_i^{(e)}, z_i^{(s)})$,

$$d_{\mathcal{M}}(z_i, z_j) = \sqrt{d_{\mathbb{H}}(z_i^{(h)}, z_j^{(h)})^2 + \|z_i^{(e)} - z_j^{(e)}\|_2^2 + d_{\mathbb{S}}(z_i^{(s)}, z_j^{(s)})^2}.$$

Dataset	Extrinsic	Product-manifold	No distance	Cost
Roman-empire	77.6±0.6	77.8±0.6	71.7±0.7	1.7×
Amazon-ratings	49.8±0.4	50.0±0.5	45.9±0.5	1.8×
Coauthor-CS	94.6±0.2	94.7±0.2	90.1±0.3	2.2×
Coauthor-Physics	96.9±0.2	97.0±0.2	93.5±0.2	2.5×
ogbn-arxiv	74.8±0.3	OOM	74.2±0.3	OOM

Table 13: Effect of the distance component used by the routing descriptor.

Dataset	Layer	Median angle	IQR	θ-active	Separation	Nondegenerate
Roman-empire	1	31.8°	$[19.6^\circ, 57.4^\circ]$	91.2±1.4	74.6±2.1	84.1±1.8
	2	34.1°	$[21.3^\circ, 60.2^\circ]$	93.0±1.2	78.2±1.9	82.0±2.0
Amazon-ratings	1	28.6°	$[17.9^\circ, 54.1^\circ]$	88.4±1.8	69.3±2.4	86.0±1.5
	2	30.4°	$[19.1^\circ, 56.8^\circ]$	90.1±1.6	72.5±2.2	84.3±1.7
Minesweeper	1	35.7°	$[22.6^\circ, 63.5^\circ]$	94.1±1.0	81.0±1.7	80.2±2.1
	2	37.2°	$[24.0^\circ, 65.6^\circ]$	95.0±0.9	83.1±1.5	78.4±2.3
Tolokers	1	26.9°	$[16.8^\circ, 52.3^\circ]$	86.3±2.0	65.4±2.6	88.1±1.4
	2	29.1°	$[18.0^\circ, 55.2^\circ]$	89.0±1.7	70.1±2.3	85.2±1.8
Questions	1	24.8°	$[15.6^\circ, 50.7^\circ]$	84.2±2.2	61.3±2.8	89.0±1.3
	2	27.3°	$[17.2^\circ, 53.8^\circ]$	87.1±1.9	66.0±2.5	87.0±1.6

Table 14: Layer-wise routing-angle statistics at the best-validation checkpoint, averaged over five seeds. Values other than the median and IQR are percentages.

The exact product distance changes performance only marginally on the datasets where it is tractable, while substantially increasing cost. Removing distance consistently degrades performance. This supports the use of standardized extrinsic distance as a lightweight routing cue; it does not replace the Riemannian geometry used for message transport.

Layer-wise routing-angle statistics. To verify that the routing angles do not collapse to degenerate values, we collect θ_{ij} over directed edges at the best-validation checkpoint for five seeds. Table 14 reports the median and interquartile range together with the fractions of edges satisfying the routing nondegeneracy conditions used in Theorem 1.

G.1 Layer-wise Embedding Evolution

We visualize the layer-wise evolution of CTS-Net node embeddings using PCA projections on Texas, Chameleon, Questions, and Coauthor-CS. Colors denote ground-truth classes. Figures 6–9 show how repeated propagation reshapes the embedding geometry across datasets with different graph sizes, homophily levels, and class structures.

H Representation Power

Proposition 2 (Attributed 1-WL Upper Bound) *CTS-Net is upper-bounded by 1-WL on the attributed graph whose node attributes are the input features and whose edge attributes contain the structural descriptors used by the routing network.*

Proof: CTS-Net uses shared node and edge functions followed by permutation-invariant neighborhood aggregation. Suppose two nodes receive the same color after ℓ iterations of attributed 1-WL. They therefore have identical current attributes and identical multisets of neighboring color–edge-attribute pairs. By induction, these pairs produce identical multisets of transported and routed messages. Since the aggregation and node-update functions are shared, the two nodes obtain identical representations at layer $\ell + 1$. Hence CTS-Net cannot distinguish nodes that attributed 1-WL cannot distinguish.

Parallel transport and orthogonal routing are information-preserving operations and therefore do not themselves merge distinct edge messages. Nevertheless, we do not claim graph-isomorphism power beyond attributed 1-WL or universal metric representation by the finite product manifold.

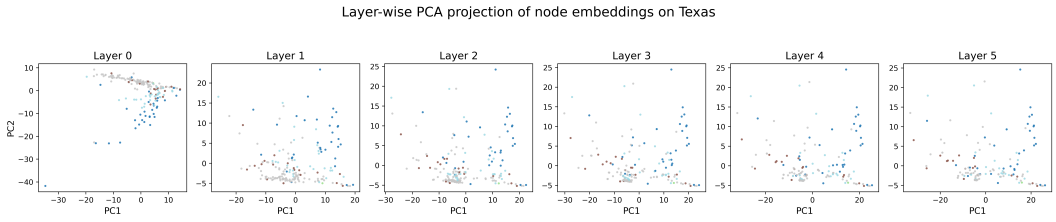


Figure 6: Layer-wise PCA projection of CTS-Net node embeddings on Texas.

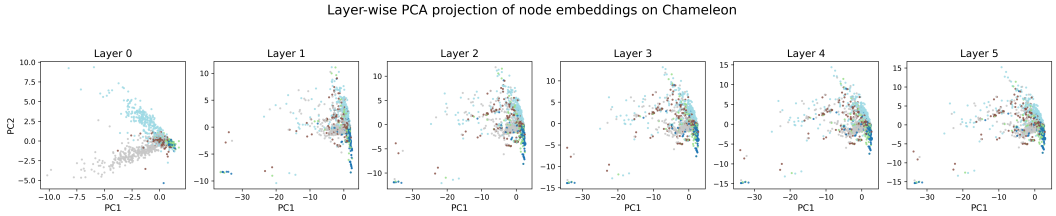


Figure 7: Layer-wise PCA projection of CTS-Netnode embeddings on Chameleon.

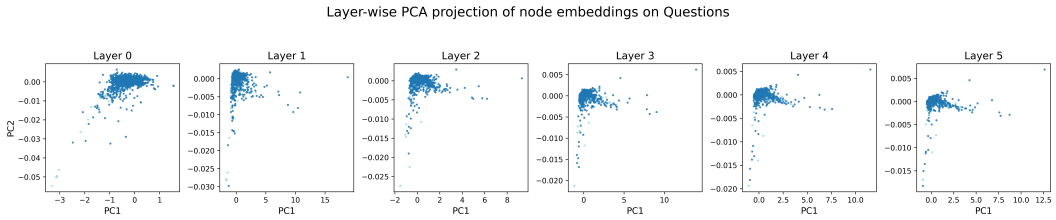


Figure 8: Layer-wise PCA projection of CTS-Net node embeddings on Questions.

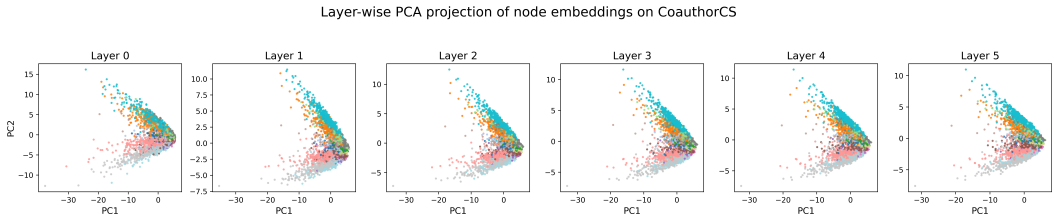


Figure 9: Layer-wise PCA projection of CTS-Net node embeddings on Coauthor-CS.