Mitigating Global Feature Collapse in Graph ODEs via Hysteretic Topology Evolution

Qinhan Hou
University of Helsinki
hou.qinhan@helsinki.fi

Jing Tang
University of Helsinki
jing.tang@helsinki.fi

Abstract

Graph neural ordinary differential equations provide continuous-time propagation on graphs, but their long-time behavior is largely determined by the asymptotic structure of the mixing operator. We show that diffusion-style Graph ODEs with positive irreducible mixing face a *monostability trap*, where node features converge to a rank-one consensus subspace. This reveals a structural limitation of positive diffusion-style continuous graph propagation: within this consensus class, exact long-time avoidance requires changing the effective support; otherwise, smooth suppression should be interpreted as finite-horizon or metastable mitigation. We address this by introducing a *hysteretic mechanism* for topology evolution, where candidate edges carry latent bistable potentials in a double-well landscape. A node-feature-conditioned force field evolves these potentials, and a smooth sigmoid gate converts them into differentiable propagation weights. We prove global consensus under the stated mixing assumptions and block-wise consensus under idealised support separation. For consensus-form dynamics, we further establish a finite-horizon perturbation bound under small cross-block leakage, and empirically evaluate the resulting nonlinear HGODE on node- and graph-level benchmarks.

1 Introduction

Inspired by neural ordinary differential equations (Chen et al., 2018), Graph ODEs (GODEs) (Poli et al., 2019; Liu et al., 2025) evolve node representations continuously over a graph, supporting adaptive long-range propagation beyond finite-depth message passing. However, continuous-time propagation also changes the nature of feature collapse. In finite-depth Graph Neural Networks (GNNs), over-smoothing is usually described as the effect of increasing aggregation layers. In diffusion-style GODEs, the corresponding risk is more structural: when the mixing operator remains positive and irreducible, consensus is the long-time attractor of the underlying dynamical system.

Though the feature collapse can be described as a classical consensus phenomenon, its implication for Graph ODE design is important: within positive diffusion-style propagation, avoiding exact long-time collapse requires controlling the asymptotic mixing support, rather than merely rescaling positive weights or regularizing feature dynamics. More specifically, the effective propagation operator must be able to move away from a single irreducible support toward separated or weakly interacting components. We therefore introduce a hysteretic mechanism for topology evolution and design an ODE framework, *Hysteresis Graph ODE* (HGODE), to mitigate this global collapse. In our proposed framework, each candidate edge is controlled by a latent bistable potential in a double-well landscape and can occupy either a connected or insulated phase, a process we call *topology polarization*. A learned force field computed from node representations drives these potentials, yielding a coupled system in which topology and features co-evolve. Unlike ACMP/GREAD-style reaction or phase-transition dynamics in the node-feature state (Wang et al., 2023; Choi et al., 2023), HGODE places the bistable state on candidate edges, so the phase transition changes the effective propagation weights

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and, under the analysis projection, the induced support. Empirically, we use soft-attention-based GODE diagnostics to demonstrate the predicted feature collapse by tuning the attention temperature, and we evaluate the resulting framework on real-world benchmarks across graph types and tasks.

Overall, our contributions are:

- We identify the *monostability trap* as a structural limitation of diffusion-style GODEs: positive irreducible mixing drives the long-time dynamics to rank-one consensus. We frame this classical consensus mechanism as a failure mode of dense positive continuous graph propagation. Additionally, we derive the corresponding design principle for this type of propagation: exact asymptotic avoidance requires the effective support to become reducible.
- We instantiate this principle through Hysteresis Graph ODE (HGODE), a coupled feature–topology ODE in which double-well edge potentials, learned forces, and smooth sigmoid gates induce connected/insulated edge phase transitions and persistent topology polarization while preserving differentiability; its strictly positive gates give finite-horizon (metastable) mitigation, which we separate from the idealized block-wise limit.
- We validate the mechanism with theory-driven synthetic diagnostics and real-world node- and graph-level benchmarks, including ablations for hysteresis, topology search, and the force-margin objective.

2 Related Work

Graph neural ODEs and continuous-depth GNNs. Graph neural ODEs extend GCN-style propagation (Kipf and Welling, 2017; Chen et al., 2020a; Luo et al., 2025) by viewing message passing as continuous-time dynamics on graphs. Early continuous-depth models such as GDE (Poli et al., 2019), CGNN (Xhonneux et al., 2020), and GRAND (Chamberlain et al., 2021) instantiate this idea mainly through diffusion-style feature flows, which provide an infinite-depth view of graph propagation. Subsequent work has modified the feature dynamics or the representation geometry to mitigate over-smoothing and improve expressivity: Neural Sheaf Diffusion (Bodnar et al., 2022) uses sheaf-valued diffusion; FLODE (Maskey et al., 2024), FROND (Kang et al., 2024), and DRAGON (Zhao et al., 2024) alter the temporal or spectral law through fractional-order dynamics; GREAD (Choi et al., 2023) and ACMP (Wang et al., 2023) introduce reaction-diffusion or phase-transition dynamics in the node-feature state; and BuNN (Bamberger et al., 2025) performs diffusion over bundle-valued representations. Recent variants further introduce control, second-order, or equivariant dynamics, as in CSG-ODE (Wang et al., 2025b), SEGNO (Liu et al., 2024), and DuSEGO (Wang et al., 2025a). These methods are complementary to HGODE, since they mainly change how features evolve on a given or softly reweighted propagation operator, whereas HGODE makes edge connectivity itself a continuous latent state and studies how its polarization affects the long-time mixing structure.

Over-smoothing and long-time collapse. Over-smoothing in discrete and continuous GNNs is commonly explained through repeated aggregation, Laplacian contraction, or random-walk convergence toward low-frequency or consensus subspaces (Oono and Suzuki, 2020; Rusch et al., 2023; Wu et al., 2023b,c; Hou et al., 2025); Roth and Liebig (2024) identify aggregation-driven rank collapse as a common cause of over-smoothing and over-correlation in discrete-depth GNNs. Residual connections, normalization methods such as PairNorm (Zhao and Akoglu, 2020), and personalized diffusion or scattering methods (Gasteiger et al., 2018; Min et al., 2020; Koishekenov, 2023) are effective finite-depth remedies, and source-term diffusion such as GRAND++ (Thorpe et al., 2022) changes the equilibrium equation so that representations need not become node-constant; none of these directly controls the asymptotic support of the mixing operator. Our analysis isolates a complementary long-time failure mode: an irreducible row-stochastic operator has a single consensus attractor, so exact asymptotic avoidance requires a reducible effective support, while smooth suppression of positive weights gives finite-horizon or metastable mitigation; HGODE introduces persistent edge-level phases that suppress harmful mixing over the learned horizon.

Graph structure learning and adaptive topology. Graph structure learning refines the input graph through adaptive adjacency matrices, metric graphs, denoising objectives, diffusion preprocessing, attention weights, or geometric rewiring criteria (Wang et al., 2019; Zheng et al., 2024; Gasteiger et al., 2019; Jin et al., 2020; Nguyen et al., 2023). GAT-style attention mechanisms (Velickovic et al., 2018; Lee et al., 2019; Ye and Ji, 2021) can also be interpreted as feature-dependent soft structure

learning. However, dense positive attention weights may preserve global irreducible mixing in the long-time limit, and many structure-learning methods learn a static or instantaneous adjacency rather than a persistent topological state. HGODE differs by assigning each candidate edge a bistable latent potential that co-evolves with node features, so that the topology itself carries memory and persistent connected/insulated phases. This positioning makes LDS/IDGL-style (Franceschi et al., 2019; Chen et al., 2020b) structure learning, robust graph denoising, and sparse/top-k attention natural complementary comparisons; our main experiments focus on the continuous-depth diffusion baselines targeted by the theory.

3 Monostability Trap and Hysteretic Topology Dynamics

3.1 Notations

For a given graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, we represent $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ and $\mathcal{E} = \{e_{ij} \mid v_i, v_j \in \mathcal{V}\}$ as the node set and edge set, respectively. For each node v_i, we define an m-dimensional feature vector $\mathbf{h}_i \in \mathbb{R}^m$, forming the global feature matrix $\mathbf{H} \in \mathbb{R}^{N\times m}$. Throughout, $t \geq 0$ is continuous depth (ODE time), T the integration horizon, and $\mathbf{H}(t)$ the feature trajectory from $\mathbf{H}(0)$; an unindexed $\mathbf{H}$ is a generic feature matrix. For $k \in \{1, \ldots, m\}$, $\mathbf{H}_{\cdot k}(t) := (\mathbf{H}_{1k}(t), \ldots, \mathbf{H}_{Nk}(t))^\top$ collects the k-th feature coordinate over all nodes, and $\mathrm{diam}(\mathbf{x}) := \max_i x_i - \min_i x_i$ for $\mathbf{x} \in \mathbb{R}^N$.

Since our framework allows for asymmetric interactions, we define the degree matrix $\mathbf{D}(t) = \mathrm{diag}(d_1(t), \ldots, d_N(t))$ with $d_i(t) = \sum_j \widetilde{\mathbf{A}}_{ij}(t)$, where $\widetilde{\mathbf{A}}(t) = \mathbf{A}(t) + \epsilon\mathbf{I}$ adds a small self-loop for numerical stability. The diffusion is governed by the row-stochastic matrix $\mathbf{P}(t) = \mathbf{D}(t)^{-1}\widetilde{\mathbf{A}}(t)$; when the time dependence is irrelevant we write $\mathbf{P} = \mathbf{D}^{-1}\widetilde{\mathbf{A}}$, and the consensus statements below hold for any row-stochastic $\mathbf{P}$. We denote the topological support of $\mathbf{P}$ as $\mathcal{E}_{supp} := \{(i, j) : \mathbf{P}_{ij} > 0\}$.

For the proposed framework, we introduce a continuous latent potential matrix $\mathbf{U} \in \mathbb{R}^{N\times N}$. To ensure scalability, we restrict computations to a sparse candidate set $\mathcal{E}_{cand} \subset \mathcal{V} \times \mathcal{V}$ that is fixed during each ODE solve (observed edges plus 2-hop, Laplacian-eigenvector, and random proposals; Sec. 4.4). A smooth sigmoid gate maps latent potentials to effective candidate-edge weights, so the forward HGODE dynamics remain differentiable. We denote the topological force function as $\mathcal{F} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ and the force on the directed edge e_{ij} as $\mathcal{F}_{ij}(t) := \mathcal{F}(\mathbf{h}_i(t), \mathbf{h}_j(t))$, which the theory only needs to be bounded and piecewise continuous in t; $\mathcal{F}_\theta$ is its learned version (Sec. 4.2). Finally, τ_{feat} and τ_{topo} denote the timescale constants for feature and topological evolution, respectively. When a hard support is discussed below, it is a thresholded analysis/evaluation projection of the learned smooth weights, not part of the ODE vector field.

All of the proofs can be found in the Appendix A.

3.2 The Monostability of Directed Consensus

We analyze the asymptotic behavior of continuous-time GNNs. Unlike previous works that rely on undirected Dirichlet energy, we consider the general case of directed, non-symmetric interactions characteristic of global soft-attention mechanisms, using *consensus dynamics* on positively weighted graphs.

Definition 3.1 (Directed Diffusion Dynamics)**.** Let $\mathbf{P} = \mathbf{D}^{-1}\widetilde{\mathbf{A}}$ be a (time-invariant) row-stochastic transition matrix of the graph. The evolution of node features $\mathbf{H}(t)$ is governed by the flow

$$\frac{d\mathbf{H}(t)}{dt} = -(\mathbf{I} - \mathbf{P})\mathbf{H}(t), \tag{1}$$

i.e., $\dot{\mathbf{h}}_i(t) = \bar{\mathbf{h}}_i(t) - \mathbf{h}_i(t)$ with $\bar{\mathbf{h}}_i(t) := \sum_j \mathbf{P}_{ij}\mathbf{h}_j(t)$: each node moves toward its current row-stochastic neighbor average.

We begin with the time-invariant case and extend it to time-varying, attention-like operators in Sec. 3.3.

Assumption 3.2 (Strongly connected positive support)**.** The directed support graph $(\mathcal{V}, \mathcal{E}_{supp})$ of $\mathbf{P}$, with $\mathcal{E}_{supp} = \{(i, j) : \mathbf{P}_{ij} > 0\}$, is strongly connected; equivalently, the row-stochastic matrix $\mathbf{P}$ is irreducible.

This assumption is satisfied by dense soft-weight constructions (*e.g.*, global attention or dense similarity kernels) and by any fixed support whose positive weights keep the directed graph irreducible. Explicit masking or sparsification can break irreducibility, so such operators are outside this theorem rather than counterexamples to it. Under Assumption 3.2, we invoke the Perron–Frobenius theorem to characterize the resulting consensus behavior, using the following notion of collapse.

Definition 3.3 (Global and block-wise consensus collapse)**.** Let $\mathcal{C}_{\mathrm{global}} := \{\mathbf{1}\mathbf{y}^\top : \mathbf{y} \in \mathbb{R}^m\}$. A trajectory $\mathbf{H}(t)$ undergoes *global consensus collapse* if $\mathrm{dist}_F(\mathbf{H}(t), \mathcal{C}_{\mathrm{global}}) := \min_{\mathbf{y}\in\mathbb{R}^m} \|\mathbf{H}(t) - \mathbf{1}\mathbf{y}^\top\|_F \to 0$ as $t \to \infty$. For a node partition $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, it reaches *block-wise consensus* if $\mathbf{H}_{\mathcal{C}_k}(t) \to \mathbf{1}\mathbf{y}_k^\top$ for every k.

Block-wise consensus with distinct $\mathbf{y}_k$ is not global collapse, although features still agree within blocks; *mitigating feature collapse* means avoiding or delaying global consensus collapse.

Theorem 3.4 (Consensus Trap under Irreducible Positive Mixing)**.** *Under Assumption 3.2, the Markov matrix $\mathbf{P}$ admits a unique stationary distribution $\boldsymbol{\pi} \succ 0$ such that $\boldsymbol{\pi}^\top\mathbf{P} = \boldsymbol{\pi}^\top$ and $\boldsymbol{\pi}^\top\mathbf{1} = 1$. Moreover, every solution of the directed consensus flow in Eq.* (1) *converges to the rank-one consensus subspace:*

$$\lim_{t\to\infty} \mathbf{H}(t) = \mathbf{1}\big(\boldsymbol{\pi}^\top\mathbf{H}(0)\big). \tag{2}$$

Let $\Delta := \min_{i\geq 2} \mathrm{Re}(1 - \lambda_i(\mathbf{P})) > 0$. Then $\|\mathbf{H}(t) - \mathbf{1}\boldsymbol{\pi}^\top\mathbf{H}(0)\| = O(e^{-(\Delta-\varepsilon)t})$ for every $\varepsilon > 0$, with rate exactly Δ when $\mathbf{P}$ is diagonalizable (Appendix A.3 gives the Jordan-form refinement).

Theorem 3.4 shows that, within this positive irreducible regime, every trajectory undergoes global consensus collapse (Definition 3.3). Escaping this trap therefore requires changing the asymptotic mixing structure, most directly by making the effective propagation reducible (*e.g.*, block-diagonal up to permutation), which yields multiple invariant subspaces corresponding to different clusters.

Remark 3.5 (Invariant consensus subspace)*.* The zero modes of Eq. (1) form $\ker(\mathbf{I} - \mathbf{P})$: $\mathrm{span}\{\mathbf{1}\}$ for irreducible $\mathbf{P}$, and $\mathrm{span}\{\mathbf{1}_{\mathcal{C}_1}, \ldots, \mathbf{1}_{\mathcal{C}_K}\}$ for $\mathbf{P} = \mathrm{diag}(\mathbf{P}_1, \ldots, \mathbf{P}_K)$ with irreducible blocks. Support-preserving reweighting, rate rescaling, lazy mixing, and the Euler layer $\mathbf{I} - h(\mathbf{I} - \mathbf{P})$ leave it unchanged, so discrete-depth propagation shares this subspace (Roth and Liebig, 2024); $K > 1$ separated blocks change it, and $\mathcal{C}_{\mathrm{global}}$ then stays invariant but is no longer attractive. Source terms such as GRAND++ (Thorpe et al., 2022) act differently, by changing the equilibrium equation (Proposition A.1, Appendix A.5).

3.3 Beyond time-invariant mixing

Theorem 3.4 is stated for a time-invariant operator, but the same collapse mechanism also applies to dense time-varying attention. In Appendix A.7, we prove that if a row-stochastic $\mathbf{P}(t)$ is piecewise continuous and uniformly positive, $\mathbf{P}_{ij}(t) \geq \alpha > 0$ for all pairs and times, then each feature coordinate contracts in diameter as

$$\mathrm{diam}(\mathbf{H}_{\cdot k}(t)) \leq e^{-2\alpha t}\,\mathrm{diam}(\mathbf{H}_{\cdot k}(0)), \qquad k = 1, \ldots, m, \tag{3}$$

and the dynamics converge to a fixed rank-one consensus state. The argument is trajectory-wise, so it also covers state-dependent $\mathbf{P}(\mathbf{H}(t), t)$ satisfying these conditions, but not arbitrary nonlinear feature maps, source terms, or normalization layers. Dense softmax attention with temperature τ_{attn} satisfies it with $\alpha \geq N^{-1}\exp(-2B^2/\tau_{\mathrm{attn}})$ for $B := \max_j \|\mathbf{h}_j(0)\|_2$, since features stay in their initial convex hull (Lemma A.10); larger attention temperature thus strengthens rather than removes the consensus pressure. This result targets dense/full support soft mixing; explicit sparsification or masking can avoid uniform positivity, but then the asymptotic behavior depends on whether the induced support becomes reducible.

3.4 Hysteresis via Double-Well Potentials

A natural way to obtain persistent support-level intervention is to make topology dynamics *bistable*. We model each directed edge e_{ij} via a latent order parameter $u_{ij}(t)$ evolving on a non-convex Landau energy landscape with two stable structural phases: *Connected* and *Insulated*, rather than as a directly optimized propagation weight. This is a binary decision at the edge level: an interaction should either support mixing or suppress it, and multi-cluster structure arises from the joint pattern of many binary edge states. The quartic double-well is the lowest-order smooth bistable normal form with an analytic saddle-node threshold; other smooth bistable potentials could be substituted (Appendices C and A.9).

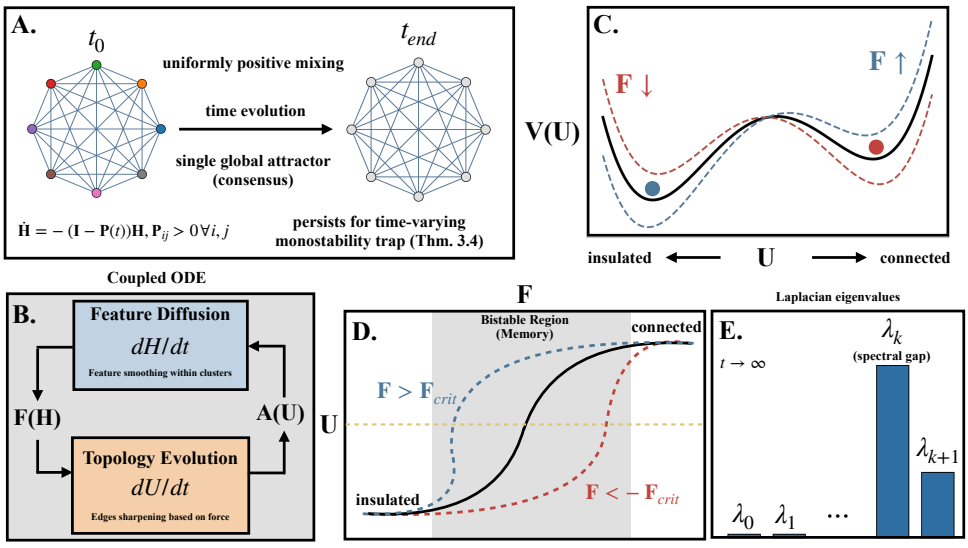


Figure 1: **HGODE mechanism and theory scope.** (A) Positive irreducible or uniformly positive diffusion-style Graph ODEs contract to a rank-one consensus subspace in the long-time regime. (B) HGODE makes topology a dynamical state by coupling feature diffusion with edge-potential evolution. (C–D) A feature-conditioned force tilts a double-well landscape, polarizing candidate edges into connected/insulated phases with hysteretic memory. (E) Edge polarization can move propagation toward block-structured mixing; in the idealized limit this yields cluster-wise invariant subspaces (*e.g.*, multiple near-zero Laplacian eigenvalues) instead of global consensus.

Definition 3.6 (Landau Double-Well Potential for Connectivity)**.** Let $u_{ij} \in \mathbb{R}$ be the latent order parameter associated with the directed edge e_{ij}. Given the node features $\mathbf{h}_i, \mathbf{h}_j$, we define the driving force $\mathcal{F}_{ij}(t) := \mathcal{F}(\mathbf{h}_i(t), \mathbf{h}_j(t))$. We define the local Landau potential (energy) for each edge as

$$V(u_{ij}; \mathcal{F}_{ij}) := \underbrace{\frac{1}{4}u_{ij}^4 - \frac{1}{2}u_{ij}^2}_{\text{intrinsic bistability}} - \underbrace{\mathcal{F}_{ij}\, u_{ij}}_{\text{coupling field}}, \tag{4}$$

where the intrinsic term yields two degenerate minima at $u_{ij} = \pm 1$ in the absence of external drive ($\mathcal{F}_{ij} = 0$), and the topological force $\mathcal{F}_{ij}$ acts as an external bias that breaks the symmetry.

Gradient-flow dynamics. We evolve the edge state by gradient descent on the potential:

$$\frac{du_{ij}}{dt} = -\frac{\partial V(u_{ij}; \mathcal{F}_{ij})}{\partial u_{ij}} = u_{ij} - u_{ij}^3 + \mathcal{F}_{ij}. \tag{5}$$

The cubic nonlinearity in (5) induces **hysteresis**: the state of an edge depends not only on the instantaneous force $\mathcal{F}_{ij}$ but also on its history (*i.e.*, which well it currently occupies), thereby providing a form of structural memory.

Proposition 3.7 (Bistability, saddle-node transitions, and hysteresis)**.** *For each edge (i, j), the equilibria of* (5) *are given by the roots of:*

$$u_{ij}^3 - u_{ij} = \mathcal{F}_{ij} \tag{6}$$

Define the critical force magnitude:

$$\mathcal{F}_{\text{crit}} := \frac{2}{3\sqrt{3}}. \tag{7}$$

Then the system undergoes saddle-node (fold) bifurcations *at $\mathcal{F}_{ij} = \pm\mathcal{F}_{\text{crit}}$, with the following regimes:*

1. ***Bistable regime** ($|\mathcal{F}_{ij}| < \mathcal{F}_{\text{crit}}$)**. Eq. 6 admits two stable equilibria (attractors) and one unstable equilibrium (repeller). Consequently, an edge initialized in the* Connected *well*

($u_{ij} \approx +1$) resists switching to Insulated *even if $\mathcal{F}_{ij}$ decreases moderately, and vice versa, yielding robustness to noise and transient perturbations.*

2. ***Monostable regime*** *($|\mathcal{F}_{ij}| > \mathcal{F}_{\mathrm{crit}}$). Only one real fixed point remains. The barrier disappears via a saddle-node transition, forcing a rapid, deterministic phase change to the single remaining attractor.*

From latent states to effective propagation. A smooth monotone gate (Eq. (10)) maps the Connected well to high and the Insulated well to near-zero propagation weight on $\mathcal{E}_{cand}$, so the forward model stays differentiable; near $|\mathcal{F}_{ij}| \approx \mathcal{F}_{\mathrm{crit}}$ the potential can move rapidly but smoothly, which adaptive solvers (`dopri5`) integrate stably. For the scaled vector field $(1-\lambda)u_{ij} - u_{ij}^3 + \mathcal{F}_{ij}$ used by HGODE, the threshold becomes $\mathcal{F}_{\mathrm{crit}}(\lambda)$ in Eq. (9).

4 The Hysteresis Graph ODE Framework

We now instantiate the coupled feature–topology ODE designed to alter the long-time mixing structure analyzed in Sec. 3. The overview of the proposed HGODE is shown in Figure 1.

4.1 Coupled Evolutionary Dynamics

HGODE governs the co-evolution of node features $\mathbf{H}(t)$ and latent edge potentials $\mathbf{U}(t)$ on a candidate pool $\mathcal{E}_{cand}$. The coupled dynamics are

$$\begin{cases} \tau_{feat}\,\dfrac{d\mathbf{H}(t)}{dt} = \mathcal{G}_\phi\big(\mathbf{H}(t), \mathbf{A}(t)\big) - \gamma\,\mathbf{H}(t), \\ \tau_{topo}\,\dfrac{d\mathbf{U}(t)}{dt} = (1-\lambda)\,\mathbf{U}(t) - \mathbf{U}(t)^{\odot 3} + \mathcal{F}_\theta\big(\mathbf{H}(t)\big), \end{cases} \tag{8}$$

where $\mathcal{G}_\phi$ is a graph neural operator, $\gamma \geq 0$ is a self-decay coefficient used for finite-time stabilization, and $\lambda \in [0, 1)$ controls the double-well depth. Eq. (8) is a framework template: $\mathcal{G}_\phi$ can be any graph operator, whereas our convergence results (Theorems 3.4 and A.9, Proposition 4.1) cover only the consensus form, in which the feature right-hand side reduces to $(\mathbf{P}(\mathbf{A}) - \mathbf{I})\mathbf{H}$, under their stated assumptions. Any additional net decay of $\mathbf{H}$ without a source term drives even block-wise consensus values toward zero, so the decay is treated as an experimental stabilization device rather than part of the collapse theory. Sec. 5 evaluates one nonlinear instantiation of $\mathcal{G}_\phi$, for which we make no convergence claim.

Continuous depth. The time t plays the role of network depth, and $\mathbf{H}(T)$ is passed to the task head. A forward-Euler step of size Δt gives $\mathbf{H}^{(\ell+1)} = \mathbf{H}^{(\ell)} + \frac{\Delta t}{\tau_{feat}}[\mathcal{G}_\phi(\mathbf{H}^{(\ell)}, \mathbf{A}(\mathbf{U}^{(\ell)})) - \gamma\mathbf{H}^{(\ell)}]$, and analogously for $\mathbf{U}^{(\ell)}$, so HGODE is a residual message-passing network whose parameters are shared across depth and whose edge weights on the fixed candidate support change with depth. We integrate Eq. (8) with adaptive `dopri5`, whose function evaluations are numerical evaluations of one vector field rather than separately parameterized layers.

For the scaled topology equation in Eq. 8, the intrinsic wells are at $\pm\sqrt{1-\lambda}$ and the saddle-node force threshold is

$$\mathcal{F}_{\mathrm{crit}}(\lambda) := \frac{2(1-\lambda)^{3/2}}{3\sqrt{3}}. \tag{9}$$

With $a := 1-\lambda$, the rescaling $\bar{u} = u/\sqrt{a}$, $\bar{t} = at/\tau_{topo}$, $\bar{\mathcal{F}} = \mathcal{F}/a^{3/2}$ maps each scalar topology equation to the canonical form $d\bar{u}/d\bar{t} = \bar{u} - \bar{u}^3 + \bar{\mathcal{F}}$ of Eq. (5), so Proposition 3.7 applies with threshold $\mathcal{F}_{\mathrm{crit}}(\lambda)$. The barrier between the wells has height $a^2/4$: increasing λ lowers both the barrier and the switching threshold, making phase changes easier but hysteretic memory weaker; $\lambda = 0$ recovers the canonical case (Appendix A.9).

Force field. $\mathcal{F}_\theta(\mathbf{H}(t)) \in \mathbb{R}^{N\times N}$ is a matrix-valued force field whose entries depend on feature compatibility:

$$\big(\mathcal{F}_\theta(\mathbf{H})\big)_{ij} = \mathcal{F}_\theta(\mathbf{h}_i, \mathbf{h}_j).$$

Its parameterization and margin-inducing objective are specified in Sec. 4.2.

From latent potentials to effective propagation. HGODE uses a sigmoid-only forward gate on the candidate set:

$$\mathbf{A}_{ij}(t) = \sigma\left(\frac{\mathbf{U}_{ij}(t)}{\tau}\right)\mathbf{1}\big[(i,j) \in \mathcal{E}_{cand}\big]. \tag{10}$$

Connected wells map to high and insulated wells to near-zero but nonzero weights. Eq. (10) is the only adjacency in the forward pass and backpropagation, which keeps the coupled ODE differentiable; sigmoid-only HGODE therefore yields finite-horizon or metastable suppression of harmful mixing, whereas exact asymptotic avoidance requires a reducible effective support.

For theory diagnostics and post-hoc evaluation only, we may inspect the thresholded support projection

$$\mathcal{E}_\eta(t) = \{(i,j) \in \mathcal{E}_{cand} : \mathbf{A}_{ij}(t) > \eta\}, \qquad \eta > 0. \tag{11}$$

It is never used in the forward ODE or in training, and represents an idealized saturated regime in which near-zero sigmoid weights are treated as absent.

Proposition 4.1 (Block-wise consensus under an evaluation support projection)**.** *Assume $\gamma = 0$. Consider the idealized projected diffusion flow*

$$\dot{\mathbf{H}}(t) = \big(\mathbf{P}_\eta(t) - \mathbf{I}\big)\mathbf{H}(t)$$

obtained by applying the support projection in Eq. (11) *after time T_0. Suppose $\mathbf{P}_\eta(t)$ is block diagonal up to permutation, with blocks $\mathbf{P}_1(t), \ldots, \mathbf{P}_K(t)$, and each $\mathbf{P}_k(t)$ is row-stochastic, piecewise continuous, and uniformly positive within its block. Then each block converges to its own consensus subspace:*

$$\lim_{t\to\infty} \big\|\mathbf{H}_{\mathcal{C}_k}(t) - \mathbf{1}\mathbf{y}_{k,\infty}^\top\big\|_F = 0, \qquad k = 1, \ldots, K.$$

where $\mathbf{y}_{k,\infty} \in \mathbb{R}^m$ is the block-specific consensus vector. If, in addition, each $\mathbf{P}_k(t) = \mathbf{P}_k$ is time-invariant after T_0 and irreducible, then $\mathbf{y}_{k,\infty} = \boldsymbol{\pi}_k^\top \mathbf{H}_{\mathcal{C}_k}(T_0)$, where $\boldsymbol{\pi}_k$ is the stationary distribution of the k-th block. Unless these block centroids coincide, the idealized projected limit is not global rank-one consensus.

Appendix A.1 gives the gradient-flow (energy) interpretation of both equations.

4.2 Topological Force Field and Scaling

The topological force $\mathcal{F}_\theta$ couples features to topology. It determines whether an edge remains in its current well or crosses the hysteresis threshold $\mathcal{F}_{\text{crit}}(\lambda)$ to switch phase. We parameterize it as a bounded edge-wise score:

$$\mathcal{F}_{ij} := \big(\mathcal{F}_\theta(\mathbf{H})\big)_{ij} = s \cdot \tanh\Big(\mathrm{MLP}_\theta\big([\mathbf{h}_i \| \mathbf{h}_j]\big)\Big), \tag{12}$$

where $(i,j) \in \mathcal{E}_{cand}$. When the margin term is active ($\beta > 0$), we choose $s \geq \mathcal{F}_{\text{crit}}(\lambda) + \delta$ so the force-margin target is feasible under the bounded $\tanh$ force. Positive forces bias an edge toward the connected well, negative forces bias it toward the insulated well, and magnitudes above $\mathcal{F}_{\text{crit}}(\lambda)$ remove one of the wells through a saddle-node transition. The theory needs no MLP, only a bounded, piecewise-continuous force with, for guaranteed switching, a sustained signed margin (Lemmas A.13–A.14); Eq. (12) is one bounded ordered-pair choice allowing $\mathcal{F}_{ij} \neq \mathcal{F}_{ji}$, and $\mathcal{F}_{\text{crit}}(\lambda)$ is fixed by the potential, not by the MLP. Since $\mathcal{F}_{ij}(t)$ follows $\mathbf{H}(t)$, an edge can switch phase in either direction within one forward pass once its force stays beyond the corresponding fold long enough.

4.3 Margin-Inducing Training Objective

Our asymptotic analysis relies on a *force-separation* condition: compatible pairs should receive positive forces exceeding $\mathcal{F}_{\text{crit}}(\lambda)$ up to a margin, while incompatible pairs should receive sufficiently negative forces. Under a sustained force-margin condition, edge potentials enter and remain in the corresponding stable phase even when the learned force is time-varying (Lemmas A.13–A.14, Appendix A.10). We encourage this condition through

$$\mathcal{L}_{margin} = \frac{1}{|\mathcal{P}_s|}\sum_{(i,j)\in\mathcal{P}_s} \mathrm{softplus}\big(\mathcal{F}_{\text{crit}}(\lambda) + \delta - \mathcal{F}_{ij}\big) + \frac{1}{|\mathcal{N}_s|}\sum_{(i,j)\in\mathcal{N}_s} \mathrm{softplus}\big(\mathcal{F}_{\text{crit}}(\lambda) + \delta + \mathcal{F}_{ij}\big), \tag{13}$$

Table 1: Scope of the collapse results, all for consensus-form propagation.

Regime	Assumption	Proven conclusion	Scope of claim
Irreducible diffusion (Thms. 3.4, A.9)	Fixed irreducible $\mathbf{P}$, or uniformly positive $\mathbf{P}(t)$	One global consensus state	Diagnosis of global collapse
Support projection (Prop. 4.1)	Block-diagonal $\mathbf{P}_\eta(t)$, uniformly positive blocks	Block-wise consensus; distinct block centroids persist	K block modes replace one
Sigmoid gates (Prop. A.2)	Cross-block perturbation $\|\mathbf{R}(t)\|_\infty \le \epsilon$	$\|\mathbf{H}_\epsilon - \mathbf{H}_0\|_\infty \le \epsilon t \|\mathbf{H}(0)\|_\infty$ for $t \le T$	Finite-horizon mitigation only

where $\mathcal{P}, \mathcal{N} \subseteq \mathcal{E}_{cand}$ are compatible and incompatible candidate pairs (so the term never adds edges), $\mathcal{P}_s, \mathcal{N}_s$ are random subsets of capped size (empty terms omitted), $\delta > 0$ is the margin, and the forces are evaluated by $\mathcal{F}_\theta$ on the pre-ODE representation $\mathbf{H}(0)$. The total training objective is

$$\mathcal{L} = \mathcal{L}_{task} + \beta\, \mathcal{L}_{margin}, \text{where} \quad \beta \ge 0 \tag{14}$$

Pair construction never uses validation or test labels: single-label node tasks pair training-labeled nodes, multi-label tasks threshold the cosine similarity of training label vectors (Appendix B.3, Table 3), and other nodes are omitted or assigned by a label-free pseudo-partition $\{c_i\}$ obtained by k-means on the ℓ_2-normalized, detached pre-ODE features (Algorithm 3). For graph-level tasks it is computed within each training graph, and graph labels never label node pairs. Margin pairs and pseudo-partitions are not needed at validation or test time, and $\mathcal{L}_{task}$ remains the primary supervision signal.

Theory scope. The edge-level results are mechanism-level (bistability, saddle-node thresholds, and polarization under a sustained margin); Table 1 summarizes the collapse results. We do not prove global convergence of the fully learned coupled system; Appendix A.6 gives finite-horizon leakage control and boundedness of the latent potentials.

4.4 Candidate Pool and Smooth Gating

Large graphs use a sparse candidate pool $\mathcal{E}_{cand}$ to avoid the $\mathcal{O}(N^2)$ cost of all-pairs topology states. The pool combines observed edges, 2-hop common-neighbor proposals, Laplacian-eigenvector (LapPE) cosine proposals, and random exploration pairs, but this is a scalability choice rather than a modeling necessity: small graphs can use $\mathcal{E}_{cand} = \mathcal{V} \times \mathcal{V}$, equivalent to neighbor ratio 1. The pool is built from the input graph alone (no labels or split masks), observed edges start with positive and new proposals with negative potentials, and the pool stays fixed during each ODE solve; for graph-level tasks it is built within each graph, including unseen test graphs. Algorithm 2 gives the exact construction, and Appendix C.1 shows why small random ratios still cover long-distance relation types.

5 Experiments

Instantiated operator. For an ordered candidate $i \leftarrow j$ (receiver i, sender j), the implemented gate and transition weights are

$$g_{ij} = \mu\, \sigma\big(u_{ij}/\tau_g\big), \quad \tilde{g}_{ij} = g_{ij} + \epsilon_{\text{loop}} \mathbf{1}[i=j], \quad p_{ij} = \frac{\tilde{g}_{ij}}{\sum_k \tilde{g}_{ik} + \epsilon_{\text{deg}}} \quad \text{(all at time } t\text{)}, \tag{15}$$

i.e., Eq. (10) with a scalar multiplier μ and temperature τ_g (annealable over training steps, fixed within a solve), self-loops $\epsilon_{\text{loop}} = 10^{-3}$, and a guard $\epsilon_{\text{deg}} = 10^{-6}$. With $\mathbf{X}^{(0)} = \mathbf{H}(t)$ and $\ell \in \{0, 1\}$,

$$\mathbf{Z}^{(\ell)} = \tanh\Big(\mathrm{LN}_\ell\big(\mathbf{P}(t)(\mathbf{X}^{(\ell)}\mathbf{W}_\ell + \mathbf{1}\mathbf{b}_\ell^\top)\big)\Big), \qquad \mathbf{X}^{(\ell+1)} = (1-\eta_\ell)\mathbf{X}^{(\ell)} + \eta_\ell \mathbf{Z}^{(\ell)}, \tag{16}$$

with a learned scalar $\eta_\ell = \sigma(r_\ell)$, and $\mathcal{G}_\phi(\mathbf{H}(t), \mathbf{A}(t)) := \mathbf{X}^{(2)}$ in Eq. (8); no dropout is sampled inside the ODE. Learned value maps, LayerNorm, $\tanh$, and residual interpolation make this field non-consensus in general, so the convergence results of Sec. 3 are not claimed for it. An encoder (Linear–LN–GELU–Dropout–Linear–LN–$\tanh$, optional positional encodings and virtual node) gives $\mathbf{H}(0)$; $\mathbf{H}(T)$ passes through Linear–LN–GELU–Dropout to $\mathbf{H}^{post}$ and then a node head fusing

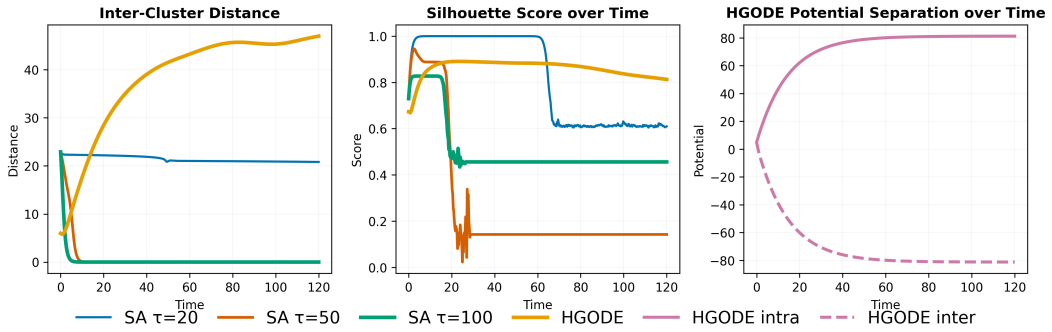


Figure 2: **Left**: Inter-cluster distance under soft attention collapses as τ_{attn} increases, indicating information leakage under increasingly diffuse mixing. **Middle**: The silhouette score decreases under soft attention at large τ_{attn}, whereas HGODE remains stable at long time horizons. **Right**: An implementation-scale HGODE potential statistic polarizes into positive (intra) and negative (inter) regimes; the diagnostic is interpreted by sign separation.

$[\mathbf{H}(0)\|\mathbf{H}^{post}]$ with an $\mathbf{H}^{post}$ residual, or a mean/max-pooling graph head (Appendix B.1). Each full graph (node tasks) or packed mini-batch (graph tasks) is integrated over a validation-selected horizon $[0, T]$.

5.1 Synthetic mechanism checks

On K-block SBM graphs with Gaussian block features, dense soft-attention Graph ODEs use strictly positive row-stochastic mixing, while HGODE follows Eq. 8 with the sigmoid gate in Eq. 10. Figure 2 shows that high attention temperature drives consensus, whereas HGODE preserves cluster separation and polarizes intra/inter-cluster potentials; full protocols are in Appendix B.4.

5.2 Real-world benchmark evaluation

We evaluate HGODE on Cora (McCallum et al., 2000), Chameleon-filtered (Rozemberczki et al., 2021; Platonov et al., 2023), Roman-empire (Platonov et al., 2023), ogbn-proteins (Hu et al., 2020), ZINC (Dwivedi et al., 2023), Peptides-func (Dwivedi et al., 2022), and ogbg-molpcba (Hu et al., 2020). Following Platonov et al., we use the duplicate-removed Chameleon-filtered setting rather than the older unfiltered split with train–test leakage from duplicated nodes. Experiments use NVIDIA A100 GPUs, AdamW, and dataset-specific learning rates; architecture, split details, and selected-configuration fields are cross-referenced in Appendix B.1 and Table 9.

Baselines. Table 2 compares message-passing GNNs (GCN (Kipf and Welling, 2017), GCNII (Chen et al., 2020a), GCN$^+$ (Luo et al., 2024)) with continuous-depth or anti-over-smoothing models (GDE (Poli et al., 2019), CGNN (Xhonneux et al., 2020), GRAND (Chamberlain et al., 2021), Sheaf Diffusion (Bodnar et al., 2022), FLODE (Maskey et al., 2024), GREAD (Choi et al., 2023), BuNN (Bamberger et al., 2025), ACMP (Wang et al., 2023), FROND (Kang et al., 2024), DRAGON (Zhao et al., 2024)). Transformer baselines use different global-attention inductive biases and are reported separately in Appendix B.5.

Ablations. *w/o hysteresis* replaces the topology equation by the single-well relaxation $\tau_{topo}\dot{\mathbf{U}} = \mathcal{F}_\theta(\mathbf{H}) - \mathbf{U}$, the gradient flow of $V_{sw}(u; \mathcal{F}) = \frac{1}{2}u^2 - \mathcal{F}u$, whose unique globally stable equilibrium $u^\star = \mathcal{F}$ keeps soft adaptive weighting but removes bistability and history dependence, with the candidate pool, force network, and gates retained. *w/o topo. search* keeps only the observed input edges in $\mathcal{E}_{cand}$. *w/o force margin* sets $\beta = 0$.

Results, solver, and ablations. We use `dopri5` with 10^{-5} tolerances; Appendix B.7 reports NFEs, memory, and wall-clock time. Table 2 is consistent with the theory: HGODE is strongest on CHAMELEON-filtered, OGBN-PROTEINS, ZINC, and OGBG-MOLPCBA, and second on sparse, chain-like ROMAN-EMPIRE, where heterophily, multi-label relations, or long-range dependencies make a single positive diffusion support prone to harmful mixing. It remains competitive on homophilous CORA and top-three on PEPTIDES-FUNC. The ablations connect the gains to the mechanism:

Table 2: Real-world benchmark results (mean ± standard deviation). All entries, including baselines, are our reruns; run counts, seeds, and splits are given in Appendix B.8. Transformer baselines are in Table 4; "–" denotes unavailable/not applicable. Top three values are green, blue, and gray. Ablation rows are defined in Sec. 5.2.

	Node Classification				Graph-Level Tasks		
Metric	Cora Acc.↑	Chameleon-filt. Acc.↑	Roman-empire Acc.↑	ogbn-proteins ROC-AUC↑	ZINC MAE↓	Peptides-func A.P.↑	ogbg-molpcba A.P.↑
Message-passing GNNs							
GCN	81.42±0.36	40.27±2.48	73.76±0.83	71.18±0.27	0.359±0.015	0.683±0.004	0.204±0.002
GCNII	84.74±0.10	44.47±1.48	83.24±0.88	75.64±0.48	0.178±0.032	0.699±0.003	0.265±0.003
GCN$^+$	85.27±0.55	45.89±2.86	91.20±0.45	77.84±0.56	0.087±0.012	**0.716±0.005**	**0.269±0.002**
Continuous-depth and anti-over-smoothing baselines							
GDE	82.44±0.81	45.63±2.47	82.37±0.61	74.63±0.20	0.239±0.012	0.632±0.014	0.247±0.002
CGNN	83.68±0.46	51.35±1.58	83.42±0.79	77.12±0.61	0.153±0.028	0.651±0.037	0.252±0.003
GRAND	83.61±0.37	57.72±1.86	83.92±0.23	76.45±0.42	0.147±0.018	0.663±0.016	0.260±0.003
Sheaf Diff.	85.53±0.64	68.16±1.62	85.43±0.62	78.96±0.56	0.148±0.021	0.678±0.047	0.263±0.002
FLODE	**86.44±1.17**	67.26±1.37	76.32±0.60	79.23±0.75	0.124±0.034	0.674±0.041	0.267±0.004
GREAD-BS	**87.33±0.84**	**71.42±1.78**	88.24±0.63	79.21±0.74	0.084±0.032	0.704±0.046	0.257±0.004
GREAD-Exp	**87.94±0.42**	69.23±1.14	90.14±0.72	78.61±0.68	0.102±0.047	0.711±0.040	0.262±0.001
BuNN	86.10±0.48	69.13±1.21	90.18±0.68	78.92±0.48	0.087±0.056	0.712±0.067	0.245±0.004
ACMP	84.76±0.74	56.24±2.01	87.12±0.79	65.72±0.72	0.124±0.036	0.679±0.051	0.226±0.005
FROND	84.67±0.86	**71.62±1.61**	**92.19±0.47**	**80.26±0.49**	**0.079±0.028**	0.698±0.034	0.260±0.005
DRAGON	84.30±0.61	70.14±1.33	**93.62±0.54**	**80.46±0.42**	**0.081±0.012**	**0.724±0.045**	**0.266±0.002**
HGODE (ours)	86.26±0.78	**72.56±1.24**	**92.56±0.43**	**81.24±0.63**	**0.078±0.025**	**0.714±0.022**	**0.278±0.003**
w/o hysteresis	83.24±0.32	66.24±1.26	90.24±0.63	75.26±0.15	0.145±0.032	0.671±0.013	0.254±0.005
w/o topo. search	84.14±0.46	70.44±1.41	90.37±0.86	77.19±0.52	0.162±0.017	0.653±0.041	0.262±0.002
w/o force margin	84.36±0.19	61.24±0.73	89.79±0.41	80.24±0.77	0.172±0.080	0.689±0.034	0.260±0.003

removing hysteresis hurts most on CHAMELEON-filtered, ROMAN-EMPIRE, and OGBN-PROTEINS; removing topology search hurts long-range graph tasks such as ZINC and PEPTIDES-FUNC; and removing the force margin hurts CHAMELEON-filtered, ROMAN-EMPIRE, and ZINC, supporting the usefulness of the force-margin objective.

6 Conclusion, Limitations, and Broader Impact

We identify the monostability trap in positive irreducible diffusion-style Graph ODEs and propose HGODE, whose bistable edge potentials learn persistent, task-dependent edge reweighting that can approach near-block mixing when such separation is useful. Exact reducible support remains an idealized analysis limit: the strictly positive sigmoid gates mitigate global consensus collapse over the finite integration horizon rather than avoiding it asymptotically.

Limitations. (i) The consensus theory covers consensus-form propagation, not an arbitrary nonlinear $\mathcal{G}_\phi$ in Eq. (8); we establish neither global convergence of the learned coupled system nor convex-hull preservation beyond Eq. (1). (ii) We evaluate one weighted message-passing $\mathcal{G}_\phi$; attention-based instantiations remain untested, and although theory and implementation allow asymmetric operators, our benchmarks are largely undirected, so the empirical role of edge direction is open. (iii) Each candidate edge carries a latent state, and the pool size must be tuned per dataset. (iv) Structure learning is not our primary comparison axis; the edge-state mechanism may also benefit dynamic structure learning, and directed graphs and attention-based backbones are natural next steps.

Broader Impact. Because HGODE learns latent edge states, inferred high- or low-weight relations should not be read as causal without domain validation; in sensitive applications, learned topology can amplify biases in features, labels, or candidate-pool construction, so deployment should include auditing and task-specific safeguards.

Acknowledgments and Disclosure of Funding

Qinhan Hou is supported by the iCANDOC Precision Cancer Medicine pilot, part of Finland's Ministry of Education and Culture's Doctoral Education Pilot. This work was also supported by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101188432 (DTRIP4H). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources. We thank the anonymous reviewers and the area chair for their constructive feedback. The authors declare no competing interests.

References

Jacob Bamberger, Federico Barbero, Xiaowen Dong, and Michael M. Bronstein. Bundle neural network for message diffusion on graphs. In *International Conference on Learning Representations*, 2025.

Guy Bar-Shalom, Beatrice Bevilacqua, and Haggai Maron. Subgraphormer: Unifying subgraph GNNs and graph transformers via graph products. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=6djDWVTUEq.

Cristian Bodnar, Francesco Di Giovanni, Benjamin P. Chamberlain, Pietro Liò, and Michael Bronstein. Neural sheaf diffusion: a topological perspective on heterophily and oversmoothing in gnns. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Ben Chamberlain, James Rowbottom, Maria I Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. Grand: Graph neural diffusion. In *International conference on machine learning*, pages 1407–1418. PMLR, 2021.

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

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. Iterative deep graph learning for graph neural networks: better and robust node embeddings. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020b. Curran Associates Inc. ISBN 9781713829546.

Jeongwhan Choi, Seoyoung Hong, Noseong Park, and Sung-Bae Cho. Gread: Graph neural reaction-diffusion networks. In *International Conference on Machine Learning*, pages 5722–5747. PMLR, 2023.

Chenhui Deng, Zichao Yue, and Zhiru Zhang. Polynormer: Polynomial-expressive graph transformer in linear time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hmv1LpNfXa.

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

Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24(43):1–48, 2023. URL http://jmlr.org/papers/v24/22-0567.html.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1972–1982. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/franceschi19a.html.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2018.

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

Qinhan Hou, Yilun Zheng, Xichun Zhang, Sitao Luan, and Jing Tang. Exploring heterophily in graph-level tasks. *arXiv preprint arXiv:2509.18893*, 2025.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. Graph structure learning for robust graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 66–74, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403049. URL `https://doi.org/10.1145/3394486.3403049`.

Qiyu Kang, Kai Zhao, Qinxu Ding, Feng Ji, Xuhao Li, Wenfei Liang, Yang Song, and Wee Peng Tay. Unleashing the potential of fractional calculus in graph neural networks with FROND. In *International Conference on Learning Representations*, 2024.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=SJU4ayYgl`.

Yeskendir Koishekenov. Reducing over-smoothing in graph neural networks using relational embeddings. *arXiv preprint arXiv:2301.02924*, 2023.

John Boaz Lee, Ryan A Rossi, Sungchul Kim, Nesreen K Ahmed, and Eunyee Koh. Attention models in graphs: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(6):1–25, 2019.

Yang Liu, Jiashun Cheng, Haihong Zhao, Tingyang Xu, Peilin Zhao, Fugee Tsung, Jia Li, and Yu Rong. SEGNO: Generalizing equivariant graph neural networks with physical inductive biases. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=3oTPsORaDH`.

Zewen Liu, Xiaoda Wang, Bohan Wang, Zijie Huang, Carl Yang, and Wei Jin. Graph odes and beyond: A comprehensive survey on integrating differential equations with graph neural networks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6118–6128, 2025.

Yuankai Luo, Lei Shi, and Xiao-Ming Wu. Classic gnns are strong baselines: Reassessing gnns for node classification. *Advances in Neural Information Processing Systems*, 37:97650–97669, 2024.

Yuankai Luo, Lei Shi, and Xiao-Ming Wu. Can classic GNNs be strong baselines for graph-level tasks? simple architectures meet excellence. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=ZH7YgIZ3DF`.

Sohir Maskey, Raffaele Paolino, Aras Bacho, and Gitta Kutyniok. A fractional graph laplacian approach to oversmoothing. *Advances in Neural Information Processing Systems*, 36, 2024.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

Yimeng Min, Frederik Wenkel, and Guy Wolf. Scattering gcn: Overcoming oversmoothness in graph convolutional networks. *arXiv preprint arXiv:2003.08414*, 2020.

Khang Nguyen, Nong Minh Hieu, Vinh Duc Nguyen, Nhat Ho, Stanley Osher, and Tan Minh Nguyen. Revisiting over-smoothing and over-squashing using ollivier-ricci curvature. In *International Conference on Machine Learning*, pages 25956–25979. PMLR, 2023.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. *In Proceedings of Eighth International Conference on Learning Representations*, 2020.

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

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 *The Eleventh International Conference on Learning Representations*, 2023.

Michael Poli, Stefano Massaroli, Junyoung Park, Atsushi Yamashita, Hajime Asama, and Jinkyoo Park. Graph neural ordinary differential equations. *arXiv preprint arXiv:1911.07532*, 2019.

Ladislav Rampášek, Mikhail 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, 2022.

Andreas Roth and Thomas Liebig. Rank collapse causes over-smoothing and over-correlation in graph neural networks. In *Proceedings of the Second Learning on Graphs Conference*, volume 231 of *Proceedings of Machine Learning Research*, pages 35:1–35:23. PMLR, 2024.

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

T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.

Matthew Thorpe, Hedi Xia, Tan Nguyen, Thomas Strohmer, Andrea L. Bertozzi, Stanley J. Osher, and Bao Wang. GRAND++: Graph neural diffusion with a source term. In *International Conference on Learning Representations*, 2022.

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

Yingxu Wang, Nan Yin, Mingyan Xiao, Xinhao Yi, Siwei Liu, and Shangsong Liang. Dusego: Dual second-order equivariant graph ordinary differential equation, 2025a. URL `https://arxiv.org/abs/2411.10000`.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019.

Yuelin Wang, Kai Yi, Xinliang Liu, Yu Guang Wang, and Shi Jin. Acmp: Allen-cahn message passing for graph neural networks with particle phase transition. *International Conference on Learning Representations*, 2023.

Zhiqiang Wang, Xiaoyi Wang, and Jianqing Liang. CSG-ODE: ControlSynth graph ODE for modeling complex evolution of dynamic graphs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 64670–64689. PMLR, 13–19 Jul 2025b. URL `https://proceedings.mlr.press/v267/wang25dd.html`.

Qitian Wu, Wentao Zhao, Chenxiao Yang, Hengrui Zhang, Fan Nie, Haitian Jiang, Yatao Bian, and Junchi Yan. Sgformer: Simplifying and empowering transformers for large-graph representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. *Advances in Neural Information Processing Systems*, 36:35084–35106, 2023b.

Xinyi Wu, Zhengdao Chen, William Wei Wang, and Ali Jadbabaie. A non-asymptotic analysis of oversmoothing in graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023c. URL `https://openreview.net/forum?id=CJd-BtnwtXq`.

Louis-Pascal Xhonneux, Meng Qu, and Jian Tang. Continuous graph neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10432–10441. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/xhonneux20a.html`.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pages 40–48. PMLR, 2016.

Yang Ye and Shihao Ji. Sparse graph attention networks. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):905–916, 2021.

Kai Zhao, Xuhao Li, Qiyu Kang, Feng Ji, Qinxu Ding, Yanan Zhao, Wenfei Liang, and Wee Peng Tay. Distributed-order fractional graph operating network. In *Advances in Neural Information Processing Systems*, 2024.

Lingxiao Zhao and Leman Akoglu. Pairnorm: Tackling oversmoothing in gnns. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rkecl1rtwB`.

Yilun Zheng, Zhuofan Zhang, Ziming Wang, Xiang Li, Sitao Luan, Xiaojiang Peng, and Lihui Chen. Rethinking structure learning for graph neural networks. *arXiv preprint arXiv:2411.07672*, 2024.

A Proofs.

A.1 Energy interpretation details

The topology equation in Eq. 8 is a Euclidean gradient flow conditioned on $\mathbf{H}$: it descends

$$\mathcal{E}_{topo}(\mathbf{U};\mathbf{H}) = \sum_{(i,j)\in\mathcal{E}_{cand}} \left[\frac{1}{4}\mathbf{U}_{ij}^4 - \frac{1-\lambda}{2}\mathbf{U}_{ij}^2 - \big(\mathcal{F}_\theta(\mathbf{h}_i,\mathbf{h}_j)\big)\mathbf{U}_{ij}\right]. \tag{17}$$

Thus $\tau_{topo}\dot{\mathbf{U}}$ follows the negative gradient up to the timescale factor. For symmetric diffusion with $\mathcal{G}_\phi(\mathbf{H},\mathbf{A}) = \mathbf{P}\mathbf{H} - \mathbf{H}$, the feature flow admits the regularized Dirichlet energy

$$\mathcal{E}_{feat}(\mathbf{H};\mathbf{U}) = \frac{1}{2}\mathrm{Tr}\Big(\mathbf{H}^\top(\mathbf{I}-\mathbf{P})\mathbf{H}\Big) + \frac{\gamma}{2}\|\mathbf{H}\|_F^2. \tag{18}$$

For reversible non-symmetric Markov operators the analogous statement uses the stationary-distribution-weighted inner product or a similarity transform. General directed row-stochastic operators need not be Euclidean gradient flows, which is why the main proofs use consensus contraction.

A.2 Proof of Theorem 3.4.

Proof. The evolution of the node features is governed by the linear differential equation:

$$\frac{d\mathbf{H}(t)}{dt} = -\mathbf{L}_{rw}\mathbf{H}(t), \tag{19}$$

where $\mathbf{L}_{rw} = \mathbf{I} - \mathbf{P}$ is the random-walk Laplacian. Since $\mathbf{P}$ is time-invariant in this analysis, the analytical solution is given by the matrix exponential:

$$\mathbf{H}(t) = e^{-t(\mathbf{I}-\mathbf{P})}\mathbf{H}(0). \tag{20}$$

We analyze the asymptotic behavior of the operator $\mathbf{T}(t) = e^{-t(\mathbf{I}-\mathbf{P})}$ via the spectral properties of $\mathbf{P}$. Since $\mathbf{P}$ is row-stochastic and irreducible (by the strong connectivity assumption), the Perron-Frobenius theorem guarantees the following:

1. The spectral radius is $\rho(\mathbf{P}) = 1$, and $\lambda_1 = 1$ is a simple eigenvalue.
2. The right eigenvector associated with λ_1 is $\mathbf{1}$ (since $\mathbf{P}\mathbf{1} = \mathbf{1}$).
3. The left eigenvector associated with λ_1 is the unique stationary distribution $\boldsymbol{\pi}$.
4. All other eigenvalues λ_i (for $i = 2,\dots,N$) satisfy $\mathrm{Re}(\lambda_i) < 1$. Irreducibility makes $\lambda = 1$ simple; periodic chains may have other eigenvalues with $|\lambda_i| = 1$, but their real parts are still strictly below 1, which is sufficient for continuous-time decay.

Let $\mathbf{P} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$ be the eigendecomposition of $\mathbf{P}$ (assuming diagonalizability for exposition; the argument holds generally using Jordan canonical forms, see Appendix A.3). The eigenvalues of the Laplacian $\mathbf{L}_{rw}$ are given by $\mu_i = 1 - \lambda_i$. Using the spectral expansion, we can express the evolution of the features as:

$$\mathbf{H}(t) = \left(\sum_{i=1}^{N} e^{-t(1-\lambda_i)}\mathbf{u}_i\mathbf{v}_i^\top\right)\mathbf{H}(0), \tag{21}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i^\top$ are the corresponding right and left eigenvectors. In the non-diagonalizable case, each Jordan block for λ_i contributes only polynomial factors multiplying $e^{-t(1-\lambda_i)}$.

Separating the dominant mode ($i = 1$) from the non-dominant modes ($i \geq 2$):

$$\mathbf{H}(t) = e^{-t(1-1)}\mathbf{1}\boldsymbol{\pi}^\top\mathbf{H}(0) + \sum_{i=2}^{N} e^{-t(1-\lambda_i)}\mathbf{u}_i\mathbf{v}_i^\top\mathbf{H}(0). \tag{22}$$

For the dominant mode, the exponential term is $e^0 = 1$. For all non-dominant modes $i \geq 2$, since $\mathrm{Re}(\lambda_i) < 1$, we have $\mathrm{Re}(1 - \lambda_i) > 0$. Consequently, the exponential terms decay asymptotically:

$$\lim_{t\to\infty} \left| e^{-t(1-\lambda_i)} \right| = \lim_{t\to\infty} e^{-t\cdot\mathrm{Re}(1-\lambda_i)} = 0. \tag{23}$$

Taking the limit $t \to \infty$ for the full system:

$$\lim_{t\to\infty} \mathbf{H}(t) = \mathbf{1}\boldsymbol{\pi}^\top \mathbf{H}(0) + \mathbf{0} = \mathbf{1}\left(\boldsymbol{\pi}^\top \mathbf{H}(0)\right). \tag{24}$$

Thus, the system converges to a state where every row of $\mathbf{H}$ is identical to the weighted centroid of the initial features, completing the proof. □

A.3 Jordan-form refinement of Theorem 3.4

Let $\mathbf{P} = \mathbf{V}\mathbf{J}\mathbf{V}^{-1}$ be a Jordan decomposition and let $\mathcal{J}_{nc}$ denote the Jordan blocks of the non-consensus eigenvalues λ_i, $i \geq 2$. By the Perron–Frobenius theorem, $\lambda_1 = 1$ is algebraically simple, so its block is 1×1 with spectral projector $\mathbf{1}\boldsymbol{\pi}^\top$. For a block $\mathbf{J}_s(\lambda) = \lambda\mathbf{I}_s + \mathbf{N}_s$ of size s, where $\mathbf{N}_s$ is nilpotent, $\mathbf{I}_s - \mathbf{J}_s(\lambda) = (1-\lambda)\mathbf{I}_s - \mathbf{N}_s$ and the two terms commute, so

$$e^{-t(\mathbf{I}_s - \mathbf{J}_s(\lambda))} = e^{-t(1-\lambda)} \sum_{r=0}^{s-1} \frac{t^r}{r!} \mathbf{N}_s^r.$$

Define $q := \max_{\mathbf{J}\in\mathcal{J}_{nc}} \big(s(\mathbf{J}) - 1\big)$, where $s(\mathbf{J})$ is the size of block $\mathbf{J}$. Since $|e^{-t(1-\lambda_i)}| = e^{-t\,\mathrm{Re}(1-\lambda_i)} \leq e^{-\Delta t}$ for $i \geq 2$, there is a constant C depending only on $\mathbf{P}$ such that

$$\left\| e^{-t(\mathbf{I}-\mathbf{P})} - \mathbf{1}\boldsymbol{\pi}^\top \right\| \leq C\,(1+t)^q e^{-\Delta t}, \qquad t \geq 0,$$

which is $O(e^{-(\Delta-\varepsilon)t})$ for every $\varepsilon > 0$. If $\mathbf{P}$ is diagonalizable, then $q = 0$ and the decay rate is exactly Δ.

A.4 Proof of Proposition 4.1

Proof. After T_0, the idealized projected support has no edges between distinct blocks, so the projected row-stochastic transition matrix is block diagonal up to a permutation:

$$\mathbf{P}_\eta(t) = \mathrm{diag}(\mathbf{P}_1(t), \ldots, \mathbf{P}_K(t)).$$

The projected feature dynamics therefore decouple across blocks. For each block $\mathcal{C}_k$, the assumed row-stochasticity, piecewise continuity, and uniform positivity of $\mathbf{P}_k(t)$ allow Theorem A.9 to be applied to the restricted feature matrix $\mathbf{H}_{\mathcal{C}_k}$, yielding

$$\lim_{t\to\infty} \left\| \mathbf{H}_{\mathcal{C}_k}(t) - \mathbf{1}\mathbf{y}_{k,\infty}^\top \right\|_F = 0.$$

This proves block-wise convergence to separate consensus subspaces. If, in addition, $\mathbf{P}_k(t) = \mathbf{P}_k$ is fixed after T_0 and irreducible, Theorem 3.4 applies within block k and gives

$$\lim_{t\to\infty} \mathbf{H}_{\mathcal{C}_k}(t) = \mathbf{1}\big(\boldsymbol{\pi}_k^\top \mathbf{H}_{\mathcal{C}_k}(T_0)\big).$$

Because no cross-block dynamics remain in the projected system, different blocks retain their own stationary centroids. The projected limit belongs to the global-consensus subspace if and only if all block consensus vectors coincide. □

A.5 Invariant consensus subspaces

Proposition A.1 (Invariant consensus subspace)**.** *Let $\mathbf{P} \in \mathbb{R}^{N\times N}$ be row-stochastic.*

1. *If $\mathbf{P}$ is irreducible, then* $\ker(\mathbf{I} - \mathbf{P}) = \mathrm{span}\{\mathbf{1}\}$.
2. *If, up to a permutation of the nodes, $\mathbf{P} = \mathrm{diag}(\mathbf{P}_1, \ldots, \mathbf{P}_K)$ with irreducible blocks on $\mathcal{C}_1, \ldots, \mathcal{C}_K$, then* $\ker(\mathbf{I} - \mathbf{P}) = \mathrm{span}\{\mathbf{1}_{\mathcal{C}_1}, \ldots, \mathbf{1}_{\mathcal{C}_K}\}$.

3. *For $c \neq 0$, $\rho \neq 1$, and $h \neq 0$, the matrices $c(\mathbf{I}-\mathbf{P})$, $\mathbf{I}-[(1-\rho)\mathbf{P}+\rho\mathbf{I}]$, and $\mathbf{I}-\mathbf{M}_h$ with $\mathbf{M}_h := \mathbf{I}-h(\mathbf{I}-\mathbf{P})$ have the same kernel as $\mathbf{I}-\mathbf{P}$. In cases 1 and 2, so does $\mathbf{I}-\mathbf{P}'$ for every row-stochastic $\mathbf{P}'$ with the same support as $\mathbf{P}$.*

4. *In case 2 with $K \geq 2$, every point of $\mathcal{C}_{\mathrm{global}}$ is an equilibrium of $\dot{\mathbf{H}} = (\mathbf{P}-\mathbf{I})\mathbf{H}$, but $\mathcal{C}_{\mathrm{global}}$ is not attractive. The limit $\lim_{t\to\infty}\mathbf{H}_{\mathcal{C}_k}(t) = \mathbf{1}\boldsymbol{\pi}_k^\top \mathbf{H}_{\mathcal{C}_k}(0)$, with $\boldsymbol{\pi}_k$ the stationary distribution of $\mathbf{P}_k$, lies in $\mathcal{C}_{\mathrm{global}}$ if and only if all block centroids $\boldsymbol{\pi}_k^\top \mathbf{H}_{\mathcal{C}_k}(0)$ coincide, and for every $\varepsilon > 0$ there are initial states within Frobenius distance ε of $\mathcal{C}_{\mathrm{global}}$ whose limits are at positive distance from it.*

Proof. (1) Since $\mathbf{I}-\mathbf{P}$ is real, it suffices to consider real $\mathbf{x}$ with $(\mathbf{I}-\mathbf{P})\mathbf{x} = \mathbf{0}$. Let $M = \max_i x_i$ and $S = \{i : x_i = M\}$. For $i \in S$, $M = x_i = \sum_j \mathbf{P}_{ij}x_j \leq M$, with equality only if $x_j = M$ whenever $\mathbf{P}_{ij} > 0$; hence S is closed under the edges of the support graph, and strong connectivity gives $S = \mathcal{V}$, *i.e.*, $\mathbf{x} = M\mathbf{1}$. Conversely, $\mathbf{P}\mathbf{1} = \mathbf{1}$. (2) The equation $(\mathbf{I}-\mathbf{P})\mathbf{x} = \mathbf{0}$ decouples into $(\mathbf{I}-\mathbf{P}_k)\mathbf{x}_{\mathcal{C}_k} = \mathbf{0}$, $k = 1, \ldots, K$, and part 1 applies to each irreducible block. (3) The three matrices equal $c(\mathbf{I}-\mathbf{P})$, $(1-\rho)(\mathbf{I}-\mathbf{P})$, and $h(\mathbf{I}-\mathbf{P})$. Irreducibility, and block-diagonal structure with irreducible blocks, depend only on the support, so a row-stochastic $\mathbf{P}'$ with the same support falls in the same case and parts 1–2 give the same kernel. (4) Since $\mathbf{P}\mathbf{1} = \mathbf{1}$, $(\mathbf{P}-\mathbf{I})\mathbf{1}\mathbf{y}^\top = \mathbf{0}$ for every $\mathbf{y}$. The flow decouples across blocks, and Theorem 3.4 applied to each block gives the stated limit, which has the form $\mathbf{1}\mathbf{y}^\top$ exactly when all block centroids coincide. For the last claim, fix $\mathbf{y} \in \mathbb{R}^m$, $\delta > 0$, and a unit vector $\mathbf{e} \in \mathbb{R}^m$, and set $\mathbf{H}(0) = \mathbf{1}\mathbf{y}^\top + \delta\,\mathbf{1}_{\mathcal{C}_1}\mathbf{e}^\top$, which lies within distance $\delta\sqrt{|\mathcal{C}_1|}$ of $\mathcal{C}_{\mathrm{global}}$. Because $\boldsymbol{\pi}_k^\top\mathbf{1} = 1$, the limit equals $\mathbf{1}(\mathbf{y}+\delta\mathbf{e})^\top$ on $\mathcal{C}_1$ and $\mathbf{1}\mathbf{y}^\top$ on the other blocks. Minimizing $|\mathcal{C}_1|\,\|\mathbf{y}+\delta\mathbf{e}-\mathbf{z}\|_2^2 + (N-|\mathcal{C}_1|)\,\|\mathbf{y}-\mathbf{z}\|_2^2$ over $\mathbf{z} \in \mathbb{R}^m$ shows that the limit is at distance $\delta\sqrt{|\mathcal{C}_1|(N-|\mathcal{C}_1|)/N} > 0$ from $\mathcal{C}_{\mathrm{global}}$. □

Relation to discrete depth and source terms. Because $\mathbf{I}-\mathbf{M}_h = h(\mathbf{I}-\mathbf{P})$, an explicit Euler layer, and hence any stack of such layers, has the same consensus subspace as the continuous flow; the two differ only in how the non-consensus modes decay, $e^{-t(1-\lambda_i)}$ versus $(1-h(1-\lambda_i))^L$ after L layers. The rank-collapse analysis of Roth and Liebig (2024) studies this aggregation-driven mechanism across discrete depth. Our statements concern the continuous-time flow, including directed, non-reversible, and uniformly positive time-varying operators (Theorems 3.4 and A.9), and couple the support to a bistable edge state. GRAND++ (Thorpe et al., 2022) adds a source term, $\dot{\mathbf{H}} = (\mathbf{P}-\mathbf{I})\mathbf{H} + \mathbf{S}$, so equilibria, when they exist, solve $(\mathbf{I}-\mathbf{P})\mathbf{H} = \mathbf{S}$ and need not be node-constant even though $\ker(\mathbf{I}-\mathbf{P}) = \mathrm{span}\{\mathbf{1}\}$ is unchanged. Support separation instead enlarges $\ker(\mathbf{I}-\mathbf{P})$ itself. Neither mechanism subsumes the other; within the homogeneous class $\dot{\mathbf{H}} = (\mathbf{P}-\mathbf{I})\mathbf{H}$, however, only a change of support, to a reducible operator with at least two closed classes, can remove the single global consensus mode.

A.6 Soft-gate leakage and coupled-system boundedness

Proposition A.2 (Finite-horizon leakage under soft gates)**.** *Let $\mathbf{P}_\epsilon(t) = \mathbf{P}_0(t) + \mathbf{R}(t)$ on $t \in [0,T]$, where $\mathbf{P}_0(t)$ is block diagonal and row-stochastic, $\mathbf{P}_\epsilon(t)$ is row-stochastic, and $\|\mathbf{R}(t)\|_\infty \leq \epsilon$ for all t. Let $\mathbf{H}_\epsilon$ and $\mathbf{H}_0$ solve*

$$\dot{\mathbf{H}}_\epsilon = (\mathbf{P}_\epsilon(t)-\mathbf{I})\mathbf{H}_\epsilon, \qquad \dot{\mathbf{H}}_0 = (\mathbf{P}_0(t)-\mathbf{I})\mathbf{H}_0$$

with the same initial condition $\mathbf{H}(0)$. Then

$$\|\mathbf{H}_\epsilon(t) - \mathbf{H}_0(t)\|_\infty \leq \epsilon t\,\|\mathbf{H}(0)\|_\infty, \qquad 0 \leq t \leq T.$$

For the consensus-form specialisation of HGODE, this bound controls the finite-horizon deviation from ideal block-wise propagation when the realised operators satisfy the stated small-leakage condition. Here $\|\cdot\|_\infty$ denotes the induced row-sum norm; for $\mathbf{X} \in \mathbb{R}^{N\times m}$, this means $\|\mathbf{X}\|_\infty = \max_i \sum_{\ell=1}^m |\mathbf{X}_{i\ell}|$.

Proof. Let $\mathbf{D}(t) = \mathbf{H}_\epsilon(t) - \mathbf{H}_0(t)$. Then

$$\dot{\mathbf{D}} = (\mathbf{P}_\epsilon(t)-\mathbf{I})\mathbf{D} + \mathbf{R}(t)\mathbf{H}_0(t), \qquad \mathbf{D}(0) = 0.$$

Let $\boldsymbol{\Phi}_\epsilon(t,s)$ be the state-transition matrix for $\dot{\mathbf{Z}} = (\mathbf{P}_\epsilon(t)-\mathbf{I})\mathbf{Z}$. Since $\mathbf{P}_\epsilon(t)-\mathbf{I}$ is a time-varying Markov generator with nonnegative off-diagonal entries and zero row sums, $\boldsymbol{\Phi}_\epsilon(t,s)$ is

nonnegative and row-stochastic. Hence $\|\mathbf{\Phi}_\epsilon(t,s)\|_\infty \le 1$. The same argument for $\mathbf{P}_0(t) - \mathbf{I}$ gives $\|\mathbf{H}_0(t)\|_\infty \le \|\mathbf{H}(0)\|_\infty$. Variation of constants yields

$$\mathbf{D}(t) = \int_0^t \mathbf{\Phi}_\epsilon(t,s)\mathbf{R}(s)\mathbf{H}_0(s)\,ds,$$

and therefore

$$\|\mathbf{D}(t)\|_\infty \le \int_0^t \|\mathbf{\Phi}_\epsilon(t,s)\|_\infty \|\mathbf{R}(s)\mathbf{H}_0(s)\|_\infty\,ds \le \int_0^t \epsilon\,\|\mathbf{H}(0)\|_\infty\,ds,$$

which proves the bound. □

Corollary A.3 (Time-dependent leakage)**.** *Under the assumptions of Proposition A.2, but with a time-dependent bound $\|\mathbf{R}(t)\|_\infty \le \epsilon(t)$ for an integrable $\epsilon(\cdot)$,*

$$\|\mathbf{H}_\epsilon(t) - \mathbf{H}_0(t)\|_\infty \le \|\mathbf{H}(0)\|_\infty \int_0^t \epsilon(s)\,ds, \qquad 0 \le t \le T.$$

In particular, fix a partition $\{\mathcal{C}_1, \dots, \mathcal{C}_K\}$, write $\mathcal{C}(i)$ for the block of node i, let $\ell_i(t) := \sum_{j \notin \mathcal{C}(i)} \mathbf{P}_{\epsilon,ij}(t) < 1$ be the cross-block mass of row i and $\ell_\infty(t) := \max_i \ell_i(t)$, and let $\mathbf{P}_0(t)$ be obtained from $\mathbf{P}_\epsilon(t)$ by deleting the cross-block entries and renormalizing each row within its block. Then $\|\mathbf{P}_\epsilon(t) - \mathbf{P}_0(t)\|_\infty = 2\ell_\infty(t)$, and hence

$$\|\mathbf{H}_\epsilon(t) - \mathbf{H}_0(t)\|_\infty \le 2\,\|\mathbf{H}(0)\|_\infty \int_0^t \ell_\infty(s)\,ds.$$

Proof. In the proof of Proposition A.2, the integrand is bounded by $\|\mathbf{R}(s)\|_\infty\|\mathbf{H}(0)\|_\infty \le \epsilon(s)\|\mathbf{H}(0)\|_\infty$, which gives the first claim. For the second, $\mathbf{P}_0(t)$ is block diagonal and row-stochastic, because each row keeps its within-block mass $1 - \ell_i(t) > 0$ and divides it by $1 - \ell_i(t)$. Row i of $\mathbf{P}_\epsilon(t) - \mathbf{P}_0(t)$ consists of the deleted cross-block entries, with absolute sum $\ell_i(t)$, and the within-block entries $\mathbf{P}_{\epsilon,ij}(t)\big(1 - (1-\ell_i(t))^{-1}\big)$, with absolute sum $(1-\ell_i(t))\big((1-\ell_i(t))^{-1} - 1\big) = \ell_i(t)$. Taking the maximum over rows gives $2\ell_\infty(t)$. The numerical self-loops of Sec. 3.1 keep every diagonal entry positive, so $\ell_i(t) < 1$ holds automatically. □

Proposition A.4 (Well-posedness and bounded edge potentials)**.** *Assume the candidate set $\mathcal{E}_{cand}$ is finite, $\tau_{feat}, \tau_{topo} > 0$, $\lambda < 1$, the bounded force satisfies $|\mathcal{F}_{ij}(t)| \le F_{\max}$ for all $(i,j) \in \mathcal{E}_{cand}$ and $t \in [0,T]$, and $\mathcal{F}_\theta$ and $\mathcal{G}_\phi$ are locally Lipschitz in their arguments. Assume also that $\mathcal{G}_\phi$ has at most linear growth in $\mathbf{H}$ uniformly over the row-normalized sigmoid gates used by HGODE. Then the coupled HGODE has a unique solution on every finite horizon $[0,T]$ for any initial $(\mathbf{H}(0), \mathbf{U}(0))$. Moreover, if $R > 0$ satisfies*

$$R^3 - (1-\lambda)R - F_{\max} > 0,$$

then every edge potential entering $[-R, R]$ remains in $[-R, R]$, and trajectories initialized outside this interval are driven back toward it.

Proof. The sigmoid gate is smooth and bounded, and the candidate set is finite. Hence the coupled vector field is locally Lipschitz. For $a := 1 - \lambda$, the scalar topology dynamics satisfy

$$\tau_{topo}\dot{u} = au - u^3 + F(t), \qquad |F(t)| \le F_{\max}.$$

At $u = R$, $au - u^3 + F(t) \le aR - R^3 + F_{\max} < 0$, while at $u = -R$, $au - u^3 + F(t) \ge -aR + R^3 - F_{\max} > 0$. Thus the vector field points inward at the boundary of $[-R, R]$. Since $R^3 - aR - F_{\max} > 0$ implies $R^2 > a$, the cubic term dominates outside the interval as well, so trajectories outside are pushed back toward the interval. The latent topology state is therefore bounded on finite horizons. With $\mathbf{U}$ bounded, the gates are bounded, and the assumed linear-growth bound on $\mathcal{G}_\phi$ gives a Gronwall bound for $\|\mathbf{H}(t)\|$ on $[0,T]$. Local existence and uniqueness can therefore be continued to the full finite horizon. □

Remark A.5 (Latent-potential identifiability). The raw potential $\mathbf{U}$ is a latent parameterization of edge phase, not an identifiable physical edge strength. Because propagation depends on $\sigma(\mathbf{U}/\tau)$, only the induced gates, their signs/polarized phases, and thresholded diagnostic supports are interpreted.

A.7 Time-varying consensus statements and proofs

Assumption A.6 (Uniform Positivity)**.** Let $\mathbf{P}(t)$ be a time-varying row-stochastic matrix. We assume that there exists a constant $\alpha > 0$ (necessarily $\alpha \leq 1/N$) such that: (i) $\mathbf{P}(t)$ is piecewise continuous in t; (ii) $\mathbf{P}_{ij}(t) \geq \alpha$ for all $i, j \in \{1, \ldots, N\}$ and all $t \geq 0$.

We first prove a differential contraction bound, from which the window contraction statement follows immediately.

Lemma A.7 (Instantaneous diameter contraction)**.** *Let $\mathbf{x}(t) \in \mathbb{R}^N$ follow the time-varying consensus flow*

$$\dot{\mathbf{x}}(t) = -(\mathbf{I} - \mathbf{P}(t))\mathbf{x}(t) = -\mathbf{x}(t) + \mathbf{P}(t)\mathbf{x}(t),$$

where $\mathbf{P}(t)$ is row-stochastic for all t and satisfies uniform positivity: $\mathbf{P}_{ij}(t) \geq \alpha$ for all i, j, t with some $\alpha > 0$. Define $\operatorname{diam}(\mathbf{x}) := \max_i \mathbf{x}_i - \min_i \mathbf{x}_i$. Then for almost every t,

$$\frac{d}{dt}\operatorname{diam}(\mathbf{x}(t)) \leq -2\alpha \operatorname{diam}(\mathbf{x}(t)).$$

Consequently,

$$\operatorname{diam}(\mathbf{x}(t)) \leq e^{-2\alpha t} \operatorname{diam}(\mathbf{x}(0)), \qquad \forall t \geq 0.$$

Proof. Let $M(t) := \max_i \mathbf{x}_i(t)$ and $m(t) := \min_i \mathbf{x}_i(t)$. Pick indices $i^+(t) \in \arg\max_i \mathbf{x}_i(t)$ and $i^-(t) \in \arg\min_i \mathbf{x}_i(t)$ (the argument holds for almost every t, since $M(t)$ and $m(t)$ are *a.e.* differentiable).

For the maximizer i^+, we have

$$\dot{\mathbf{x}}_{i^+}(t) = -\mathbf{x}_{i^+}(t) + \sum_{j=1}^{N} \mathbf{P}_{i^+ j}(t)\, \mathbf{x}_j(t).$$

Let j^- be any minimizer such that $\mathbf{x}_{j^-}(t) = m(t)$. By uniform positivity, $\mathbf{P}_{i^+ j^-}(t) \geq \alpha$. Since all other $\mathbf{x}_j(t) \leq M(t)$, we obtain

$$\sum_{j=1}^{N} \mathbf{P}_{i^+ j}(t)\, \mathbf{x}_j(t) \leq \alpha m(t) + (1 - \alpha) M(t),$$

and hence $\dot{M}(t) \leq -\alpha(M(t) - m(t))$. Similarly, for the minimizer i^-, using a maximizer j^+ gives $\dot{m}(t) \geq \alpha(M(t) - m(t))$. Combining the two inequalities yields

$$\frac{d}{dt}\operatorname{diam}(\mathbf{x}(t)) \leq -2\alpha \operatorname{diam}(\mathbf{x}(t)),$$

and Gronwall's inequality gives the stated exponential bound. □

Lemma A.8 (Window contraction)**.** *Under Assumption A.6, for any solution $\mathbf{x}(t)$ of the scalar consensus flow in Lemma A.7 and any $T_w > 0$ there exists $\rho \in (0, 1)$ such that*

$$\operatorname{diam}(\mathbf{x}(t + T_w)) \leq (1 - \rho) \operatorname{diam}(\mathbf{x}(t)), \qquad \forall t \geq 0.$$

One can take $\rho := 1 - e^{-2\alpha T_w}$.

Proof. Apply Lemma A.7 over the interval $[t, t + T_w]$. □

Theorem A.9 (Consensus trap under time-varying positive mixing)**.** *Consider the time-varying consensus flow*

$$\frac{d\mathbf{H}(t)}{dt} = -(\mathbf{I} - \mathbf{P}(t))\mathbf{H}(t),$$

where $\mathbf{P}(t)$ satisfies Assumption A.6. Then for each feature dimension $k \in \{1, \ldots, m\}$, there exist constants $C, \kappa > 0$ such that

$$\operatorname{diam}(\mathbf{H}_{\cdot k}(t)) \leq C e^{-\kappa t} \operatorname{diam}(\mathbf{H}_{\cdot k}(0)).$$

Moreover, there exists a fixed vector $\mathbf{y}_\infty \in \mathbb{R}^m$ such that

$$\lim_{t \to \infty} \|\mathbf{H}(t) - \mathbf{1}\mathbf{y}_\infty^\top\|_F = 0.$$

Proof. Fix a feature dimension $k \in \{1, \dots, m\}$ and define $\mathbf{x}(t) := \mathbf{H}_{\cdot k}(t) \in \mathbb{R}^N$. From the matrix ODE, $\mathbf{x}(t)$ satisfies

$$\dot{\mathbf{x}}(t) = -(\mathbf{I} - \mathbf{P}(t))\mathbf{x}(t).$$

Applying Lemma A.7 yields

$$\operatorname{diam}(\mathbf{H}_{\cdot k}(t)) \le e^{-2\alpha t} \operatorname{diam}(\mathbf{H}_{\cdot k}(0)).$$

Thus we can take $C = 1$ and $\kappa = 2\alpha$.

It remains to show that the consensus value is fixed. Let $M(t) := \max_i \mathbf{x}_i(t)$ and $m(t) := \min_i \mathbf{x}_i(t)$. Because each row of $\mathbf{P}(t)$ is stochastic, $(\mathbf{P}(t)\mathbf{x}(t))_i$ lies in $[m(t), M(t)]$ for every i. At an index attaining the maximum, $\dot{M}(t) \le -M(t) + M(t) = 0$ for almost every t; at an index attaining the minimum, $\dot{m}(t) \ge -m(t) + m(t) = 0$. Hence $M(t)$ is non-increasing and bounded below, while $m(t)$ is non-decreasing and bounded above. Both have limits, and the exponential diameter bound implies $M(t) - m(t) \to 0$. Therefore all entries of $\mathbf{x}(t)$ converge to a common scalar $y_{\infty,k}$.

Applying this argument to all feature dimensions gives $\mathbf{y}_\infty = (y_{\infty,1}, \dots, y_{\infty,m})$. Since every coordinate diameter vanishes and every coordinate has the fixed limit above,

$$\lim_{t\to\infty} \|\mathbf{H}(t) - \mathbf{1}\mathbf{y}_\infty^\top\|_F = 0.$$

□

Lemma A.10 (Convex-hull invariance)**.** *Let $\mathbf{P}(t)$ be row-stochastic and piecewise continuous on $[0, \infty)$, and let $\mathbf{\Phi}(t, s)$ be the state-transition matrix of $\dot{\mathbf{X}} = (\mathbf{P}(t) - \mathbf{I})\mathbf{X}$. For all $t \ge s \ge 0$, $\mathbf{\Phi}(t, s)$ is entrywise nonnegative and row-stochastic. Consequently, every solution of $\dot{\mathbf{H}} = (\mathbf{P}(t) - \mathbf{I})\mathbf{H}$ satisfies $\mathbf{h}_i(t) = \sum_j \mathbf{\Phi}_{ij}(t, 0)\, \mathbf{h}_j(0) \in \operatorname{conv}\{\mathbf{h}_1(0), \dots, \mathbf{h}_N(0)\}$ and $\|\mathbf{h}_i(t)\| \le \max_j \|\mathbf{h}_j(0)\|$ for every norm. If $\mathbf{P}$ is fixed and irreducible, the consensus limit $\sum_j \boldsymbol{\pi}_j \mathbf{h}_j(0)$ of Theorem 3.4 is a point of this hull. The same conclusions hold along any solution of a state-dependent flow $\dot{\mathbf{H}} = (\mathbf{P}(\mathbf{H}(t), t) - \mathbf{I})\mathbf{H}$, applied to the realized operator.*

Proof. Write $\mathbf{\Phi}(t, s) = e^{-(t-s)}\mathbf{\Psi}(t, s)$, where $\mathbf{\Psi}$ solves $\partial_t \mathbf{\Psi}(t, s) = \mathbf{P}(t)\mathbf{\Psi}(t, s)$ with $\mathbf{\Psi}(s, s) = \mathbf{I}$. Because $\mathbf{P}(t)$ is bounded, the Peano–Baker series

$$\mathbf{\Psi}(t, s) = \mathbf{I} + \int_s^t \mathbf{P}(t_1)\, dt_1 + \int_s^t \int_s^{t_1} \mathbf{P}(t_1)\mathbf{P}(t_2)\, dt_2\, dt_1 + \cdots$$

converges, and all its terms are entrywise nonnegative, so $\mathbf{\Phi}(t, s) \ge 0$. Since $(\mathbf{P}(t) - \mathbf{I})\mathbf{1} = \mathbf{0}$, the constant function $\mathbf{1}$ solves $\dot{\mathbf{x}} = (\mathbf{P}(t) - \mathbf{I})\mathbf{x}$, hence $\mathbf{\Phi}(t, s)\mathbf{1} = \mathbf{1}$ by uniqueness. Each row of $\mathbf{\Phi}(t, 0)$ therefore contains convex weights, which gives the hull statement; the norm bound follows from convexity of norms. For fixed irreducible $\mathbf{P}$, the limit $\mathbf{1}\boldsymbol{\pi}^\top\mathbf{H}(0)$ has rows $\sum_j \boldsymbol{\pi}_j \mathbf{h}_j(0)$ with $\boldsymbol{\pi} \ge 0$ and $\boldsymbol{\pi}^\top\mathbf{1} = 1$. For a state-dependent flow, a solution defines the piecewise continuous row-stochastic operator $\mathbf{P}(t) := \mathbf{P}(\mathbf{H}(t), t)$, to which the linear argument applies. □

Corollary A.11 (Dense soft-attention induces the consensus trap)**.** *Let $\mathbf{P}(t)$ be defined by fully connected softmax attention with temperature $\tau_{\text{attn}} > 0$,*

$$\mathbf{P}_{ij}(t) := \frac{\exp(\langle \mathbf{h}_i(t), \mathbf{h}_j(t)\rangle/\tau_{\text{attn}})}{\sum_{\ell=1}^N \exp(\langle \mathbf{h}_i(t), \mathbf{h}_\ell(t)\rangle/\tau_{\text{attn}})},$$

and let $\mathbf{H}(t)$ solve $\dot{\mathbf{H}} = (\mathbf{P}(t) - \mathbf{I})\mathbf{H}$. With $B := \max_j \|\mathbf{h}_j(0)\|_2$, we have $\|\mathbf{h}_i(t)\|_2 \le B$ and $\mathbf{P}_{ij}(t) \ge N^{-1}\exp(-2B^2/\tau_{\text{attn}})$ for all i, j and all $t \ge 0$, and the dynamics converge to a fixed rank-one consensus state.

Proof. The vector field is locally Lipschitz, so a unique solution exists locally. Along it, $\mathbf{P}(t)$ is continuous and row-stochastic, so Lemma A.10 gives $\|\mathbf{h}_i(t)\|_2 \le B$; the solution therefore stays bounded and extends to all $t \ge 0$. By Cauchy–Schwarz, $|\langle \mathbf{h}_i(t), \mathbf{h}_j(t)\rangle| \le B^2$. Thus each numerator is at least $\exp(-B^2/\tau_{\text{attn}})$, while the denominator is at most $N\exp(B^2/\tau_{\text{attn}})$. Hence

$$\mathbf{P}_{ij}(t) \ge \frac{1}{N}\exp\left(-\frac{2B^2}{\tau_{\text{attn}}}\right).$$

This satisfies Assumption A.6, so Theorem A.9 applies. □

A.8 Proof of Proposition 3.7

Proof. Equilibria satisfy $u_{ij} - u_{ij}^3 + \mathcal{F}_{ij} = 0$, *i.e.*, (6). At a saddle-node transition, the cubic has a double root, so both $f(u) := u^3 - u - \mathcal{F}_{ij} = 0$ and $f'(u) = 3u^2 - 1 = 0$ hold. Thus $u = \pm 1/\sqrt{3}$ and $\mathcal{F}_{ij} = u^3 - u = \pm 2/(3\sqrt{3})$, giving (7). The stated stability and regime characterization follow from standard one-dimensional bifurcation analysis of the forced double-well potential. □

A.9 Double-well as a minimal bistable normal form

Proposition A.12 (Minimal smooth bistable normal form)**.** *Consider a scalar edge order parameter u whose two phases should be symmetric in the absence of evidence and stable at nonzero values. Among bounded-below even polynomial potentials, the lowest-order family that can represent two stable minima separated by an unstable barrier is*

$$V_0(u) = \frac{\kappa}{4}u^4 - \frac{b}{2}u^2, \qquad \kappa, b > 0.$$

Under an additive force F, the gradient flow $\dot{u} = bu - \kappa u^3 + F$ becomes, after the rescaling $\bar{u} = \sqrt{\kappa/b}\, u$ and $\bar{t} = bt$,

$$\frac{d\bar{u}}{d\bar{t}} = \bar{u} - \bar{u}^3 + \bar{F}, \qquad \bar{F} := \frac{\sqrt{\kappa}}{b^{3/2}}\, F.$$

Thus Eq. (5) *is the canonical low-order smooth model for two persistent edge phases with an analyzable switching threshold.*

Proof. A nonconstant even quadratic potential has at most one stable critical point when bounded below, or is unbounded below otherwise, and therefore cannot encode two stable phases. Adding the quartic term gives the first bounded-below even polynomial with stationary points at $u = 0$ and $u = \pm\sqrt{b/\kappa}$; the outer points are stable minima and the origin is the separating barrier. An additive force contributes a linear tilt $-Fu$, which preserves differentiability and yields the gradient flow $\dot{u} = bu - \kappa u^3 + F$. Substituting $u = \sqrt{b/\kappa}\,\bar{u}$ and $t = \bar{t}/b$ gives $b\sqrt{b/\kappa}\, d\bar{u}/d\bar{t} = b\sqrt{b/\kappa}\,(\bar{u} - \bar{u}^3) + F$, which is the stated normal form. □

The HGODE scaling and the role of λ. Each scalar topology equation in Eq. (8) reads $\tau_{topo}\dot{u} = au - u^3 + \mathcal{F}$ with $a := 1 - \lambda \in (0, 1]$, *i.e.*, $\kappa = 1$ and $b = a$ up to the time constant τ_{topo}. The substitution $\bar{u} = u/\sqrt{a}$, $\bar{t} = at/\tau_{topo}$, $\bar{\mathcal{F}} = \mathcal{F}/a^{3/2}$ gives $d\bar{u}/d\bar{t} = \bar{u} - \bar{u}^3 + \bar{\mathcal{F}}$, so Proposition 3.7 yields saddle-node transitions at $|\mathcal{F}| = a^{3/2} \cdot 2/(3\sqrt{3}) = \mathcal{F}_{\mathrm{crit}}(\lambda)$, as in Eq. (9). For $\mathcal{F} = 0$, the wells $u = \pm\sqrt{a}$ have energy $-a^2/4$ and the barrier at $u = 0$ has energy 0, so the barrier height is $a^2/4$. Increasing λ therefore moves the wells toward zero and lowers both the barrier and the switching threshold: phase changes become easier, while hysteretic memory weakens. The force parameterization (*e.g.*, the MLP in Eq. (12)) does not enter $\mathcal{F}_{\mathrm{crit}}(\lambda)$; it only determines whether and when the realized force crosses it.

A.10 Force margin implies polarization of edge potentials

Lemma A.13 (Force margin implies global polarization)**.** *Consider the scalar edge-potential dynamics*

$$\dot{u} = au - u^3 + F, \tag{25}$$

where $a = 1 - \lambda > 0$ and $F \in \mathbb{R}$ is constant. Define $F_{\mathrm{crit}}(\lambda) := 2(1 - \lambda)^{3/2}/(3\sqrt{3})$. If $|F| > F_{\mathrm{crit}}(\lambda)$, then (25) *admits a* unique *equilibrium $u^\star(F)$, and this equilibrium is* globally asymptotically stable*. Moreover,* $\mathrm{sign}(u^\star(F)) = \mathrm{sign}(F)$*. In particular, if $|F| \geq F_{\mathrm{crit}}(\lambda) + \delta$ for some $\delta > 0$, then every trajectory satisfies $u(t) \to u^\star(F)$ as $t \to \infty$, i.e., the potential polarizes to the connected well for $F > 0$ and to the insulated well for $F < 0$.*

Proof. Define $g(u) := au - u^3 + F$. Equilibria are roots of $g(u) = 0$, i.e., of the cubic $u^3 - au = F$. At a saddle-node transition, both $u^3 - au - F = 0$ and $3u^2 - a = 0$ hold, giving $u = \pm\sqrt{a/3}$ and the critical magnitude $2a^{3/2}/(3\sqrt{3}) = F_{\mathrm{crit}}(\lambda)$. Hence, when $|F| > F_{\mathrm{crit}}(\lambda)$, the cubic has exactly one real root; denote it by $u^\star(F)$.

Local stability follows from $g'(u) = a - 3u^2$. In the monostable regime, the unique equilibrium is stable, so $g'(u^\star) < 0$. Since g is continuous and has a unique zero, $g(u)$ has a fixed sign on each side of $u^\star$: $g(u) < 0$ for all $u > u^\star$ and $g(u) > 0$ for all $u < u^\star$. Thus trajectories are monotone toward $u^\star$ and bounded, so they converge to the unique equilibrium. Finally, $g(0) = F$ and the cubic tails imply $\operatorname{sign}(u^\star(F)) = \operatorname{sign}(F)$. □

Lemma A.14 (Sustained time-varying force margin)**.** *Consider the scalar edge-potential dynamics*

$$\dot{u} = au - u^3 + F(t),$$

where $a = 1 - \lambda > 0$ and $F(t)$ is bounded and piecewise continuous. Let $c = \sqrt{a/3}$, and call $\{u \geq c\}$ the positive phase and $\{u \leq -c\}$ the negative phase. If there exist $\delta > 0$ and $T_0 \geq 0$ such that $F(t) \geq F_{\mathrm{crit}}(\lambda) + \delta$ for all $t \geq T_0$, then every trajectory enters the positive phase in finite time and remains there. If $F(t) \leq -F_{\mathrm{crit}}(\lambda) - \delta$ for all $t \geq T_0$, then every trajectory enters the negative phase in finite time and remains there. If additionally $F(t) \to F_\infty$, then $u(t) \to u^\star(F_\infty)$, the unique equilibrium of $au - u^3 + F_\infty = 0$.

Proof. Let $g_t(u) := au - u^3 + F(t)$. The polynomial $au - u^3$ has local minimum $-F_{\mathrm{crit}}(\lambda)$ at $u = -c$ and local maximum $F_{\mathrm{crit}}(\lambda)$ at $u = c$. If $F(t) \geq F_{\mathrm{crit}}(\lambda) + \delta$, then $g_t(u) \geq \delta$ for all $u \leq c$. Hence any trajectory below c increases at speed at least δ until it reaches c, and at $u = c$ the vector field is positive, so the trajectory cannot cross back into $u < c$. This proves finite-time entry and persistence of the positive phase. The negative-force case is symmetric: if $F(t) \leq -F_{\mathrm{crit}}(\lambda) - \delta$, then $g_t(u) \leq -\delta$ for all $u \geq -c$, giving finite-time entry into and persistence in the negative phase.

The cubic term $-u^3$ makes all trajectories bounded. When $F(t) \to F_\infty$, the limiting autonomous scalar equation is in the monostable regime and has a unique globally stable equilibrium $u^\star(F_\infty)$ by Lemma A.13. Standard one-dimensional comparison with the autonomous equations using forces $F_\infty \pm \varepsilon$ traps the limsup and liminf of $u(t)$ between roots that converge to $u^\star(F_\infty)$ as $\varepsilon \downarrow 0$. Thus $u(t) \to u^\star(F_\infty)$. □

B Supplementary Experiments and Implementation Details

This appendix provides the full synthetic protocol, graph-transformer comparisons, efficiency profiling, and reproducibility details referenced from Sec. 5.2. The main text focuses on the core mechanism and the closest continuous-depth baselines.

B.1 Implementation architecture and algorithm

This subsection specifies the released implementation. Let d be the hidden width, p the dropout rate, $\mathbf{x}_i$ the input features of node i, and Drop_p dropout with rate p. Message passing follows the convention of Eq. (15): an ordered candidate (i, j) carries a message from sender j to receiver i.

Encoder and initial state. The encoder is

$$\mathbf{h}_i^{enc} = \tanh\Big(\mathrm{LN}\big(\mathbf{W}_2\,\mathrm{Drop}_p\big(\mathrm{GELU}(\mathrm{LN}(\mathbf{W}_1\mathbf{x}_i + \mathbf{b}_1))\big) + \mathbf{b}_2\big)\Big).$$

An optional positional encoding is added after the bounded activation, giving $\mathbf{h}_i^{enc} + \mathrm{PE}_i$. The default degree encoding is $\mathrm{PE}_i = \mathrm{GELU}\big(\mathrm{LN}(\mathbf{W}_{pe}[\mathbf{E}_{\min(\deg_i,100)} \| (\mathbf{w}\log(1+\deg_i)+\mathbf{b})] + \mathbf{b}_{pe})\big)$, where $\mathbf{E}$ is a learned embedding table and $\deg_i$ counts the arcs received by node i; Laplacian and random-walk encodings are available alternatives. With the optional virtual node, a learned vector $\mathbf{v}^{(0)}$, shared by all graphs, is broadcast to the nodes of each graph g,

$$\mathbf{h}_i \leftarrow \mathbf{h}_i + \sigma(a)\,\mathrm{MLP}_b(\mathbf{v}_{g(i)}), \qquad \mathbf{v}_g \leftarrow \mathbf{v}_g + \mathrm{MLP}_u\Big(\tfrac{1}{|g|}\textstyle\sum_{i\in g}\mathbf{h}_i\Big),$$

with a learned scalar a (initialized at -2) and $\mathrm{MLP}_b = $ Linear–LN–GELU; MLP_u is Linear($d \to 2d$)–LN–GELU–$\mathrm{Drop}_{0.1}$–Linear($2d \to d$)–LN, and the mean uses the already broadcast node states. The resulting node states form $\mathbf{H}(0)$. The virtual node is not part of the ODE state.

Algorithm 1 One HGODE training step

Require: graph(s) with arcs $\mathcal{E}$ and features $\mathbf{X}$; training labels; hyperparameters (Tables 8–11)

1: Build $\mathcal{E}_{cand}$ and $\mathbf{U}(0)$ from $\mathcal{E}$ alone (Algorithm 2)
2: $\mathbf{H}(0) \leftarrow$ encoder, positional encoding, and virtual-node broadcast applied to $\mathbf{X}$
3: $(\mathbf{H}(T), \mathbf{U}(T)) \leftarrow$ integrate Eq. (8) on $[0, T]$, evaluating Eqs. (15)–(16) at every solver call
4: Compute predictions with the node or graph head from $\mathbf{H}(0)$ and $\mathbf{H}(T)$
5: Form $\mathcal{P}, \mathcal{N} \subseteq \mathcal{E}_{cand}$ from training labels, or from Algorithm 3 when no label partition is supplied (Appendix B.3)
6: Evaluate $\mathcal{L}_{margin}$ (Eq. (13)) with $\mathcal{F}_\theta$ applied to $\mathbf{H}(0)$ on sampled pairs
7: Update all parameters with $\nabla(\mathcal{L}_{task} + \beta\mathcal{L}_{margin})$ and gradient-norm clipping at 1.0

ODE, force network, and solver. Eq. (8) is integrated from $(\mathbf{H}(0), \mathbf{U}(0))$, with $\mathcal{G}_\phi$ given by Eqs. (15)–(16) and $\mathbf{U}(0)$ by Algorithm 2. The force network maps the receiver-then-sender concatenation $[\mathbf{h}_i \| \mathbf{h}_j]$ through Linear($2d \to 64$)–ReLU–Linear($64 \to 32$)–ReLU–Linear($32 \to 1$) before the bounded output $s\tanh(\cdot)$ of Eq. (12), so $\mathcal{F}_{ij} \neq \mathcal{F}_{ji}$ in general. We use `torchdiffeq` with `dopri5`, `rtol` $=$ `atol` $= 10^{-5}$, and at most 1000 solver steps; if adaptive integration fails, the implementation falls back to a fixed-step RK4 solve with 8 steps on $[0, T]$. Non-finite entries of the terminal state are replaced by finite values before the readout.

Readout. With a virtual node, the updated $\mathbf{v}_g$ is broadcast once more into $\mathbf{H}(T)$ as above. The post-ODE map is $\mathbf{h}_i^{post} = \mathrm{Drop}_{p/2}\big(\mathrm{GELU}(\mathrm{LN}(\mathbf{W}_{post}\mathbf{h}_i(T) + \mathbf{b}_{post}))\big)$. The node head fuses pre- and post-ODE states and adds a residual,

$$\mathbf{f}_i = \mathrm{Drop}_{p/2}\Big(\mathrm{GELU}\big(\mathrm{LN}(\mathbf{W}_f[\mathbf{h}_i(0) \| \mathbf{h}_i^{post}] + \mathbf{b}_f)\big)\Big), \qquad \hat{\mathbf{y}}_i = \mathrm{MLP}_c\big(\mathbf{f}_i + \mathbf{h}_i^{post}\big),$$

and the graph head computes $\hat{\mathbf{y}}_g = \mathrm{MLP}_c\big(\mathrm{pool}_{i \in g}\, \mathbf{h}_i^{post}\big)$ with mean, max, sum, or concatenated mean/max pooling. In both heads, MLP_c is Linear–LN–GELU–Drop_p–Linear. The graph head therefore has no pre-/post-ODE fusion or residual.

Edge attributes. When edge attributes $\mathbf{a}_{ij}$ are available, an encoder $e_{ij} = \sigma\big(\mathbf{w}_e^\top \mathrm{GELU}(\mathbf{W}_e\mathbf{a}_{ij} + \mathbf{b}_e) + c_e\big) \in (0, 1)$ rescales the initial potential of each observed edge to $u_{ij}(0) = u_{obs}\, e_{ij}$. Edge attributes enter no other component.

Training and evaluation. Algorithm 1 summarizes one training step. Validation and test passes run lines 1–4 with the same horizon T and use no labels, margin pairs, or pseudo-partitions.

B.2 Candidate-pool construction

Algorithm 2 builds the candidate pool of one graph; batched graphs are processed one graph at a time, so graph-level tasks build a pool within each graph, including unseen test graphs. The construction uses only the input graph, never labels or split masks, and it is deterministic (seeded random proposals, canonicalized eigenbases, and quantized similarities), so it is rebuilt identically at every forward pass and stays fixed during each ODE solve. Observed arcs are kept as supplied, so directed inputs remain directed, whereas proposals are inserted in both directions. The LapPE ranking does not exclude observed neighbors; such selections coincide with an observed arc and keep its potential after deduplication. A proposal type is skipped for graphs larger than its size cap, so the realized pool can be smaller than the nominal budget $(k_{2hop}, k_{lap}, r_{rand})$.

B.3 Force-margin pair construction

Sampling. At every training step, the compatible and incompatible candidate pairs $\mathcal{P}, \mathcal{N} \subseteq \mathcal{E}_{cand}$ are formed as in Table 3, and at most n_{pair} pairs of each type are drawn uniformly without replacement to give $\mathcal{P}_s$ and $\mathcal{N}_s$ in Eq. (13). For single-label node tasks, the partition is given by the training labels, and every pair with an endpoint outside the training set is excluded.

Multi-label tasks. For multi-label node datasets such as ogbn-proteins, each training-labeled node has a binary label vector $\mathbf{y}_i \in \{0, 1\}^C$. For candidate pairs whose endpoints both have training labels,

Algorithm 2 Static candidate pool for one graph with n nodes

Require: observed arcs $\mathcal{E}$; budgets $k_{2hop}, k_{lap}, r_{rand}$; cap c_{rand}; LapPE dimension d_{PE}; size caps $n_{2hop}^{\max}, n_{lap}^{\max}$; potentials $u_{obs} > 0 > u_{new}$

$\mathbf{A}^{sym} \leftarrow$ 0/1 adjacency of the symmetrization of $\mathcal{E}$, without self-loops
if $k_{2hop} > 0$ and $n \le n_{2hop}^{\max}$ **then**
    For each i, propose the k_{2hop} non-neighbors $j \neq i$ with the most common neighbors, keeping only $(\mathbf{A}^{sym}\mathbf{A}^{sym})_{ij} > 0$
if $k_{lap} > 0$ and $n \le n_{lap}^{\max}$ **then**
    Eigendecompose $\mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}^{sym}\mathbf{D}^{-1/2}$ in float64 and canonicalize degenerate eigenspaces
    $\mathbf{z}_i \leftarrow$ row i of eigenvectors $2, \dots, d_{PE}+1$, ℓ_2-normalized
    For each i, propose the k_{lap} nodes $j \neq i$ with the largest $\mathbf{z}_i^\top \mathbf{z}_j$, rounded to 10^{-6} with ties broken by node index
if $r_{rand} > 0$ **then**
    For each i, propose $\min(\lceil r_{rand} n \rceil, c_{rand})$ nodes $j \neq i$ drawn uniformly, with a generator seeded by a hash of $\mathcal{E}$
Insert every proposal in both directions; $\mathcal{E}_{cand} \leftarrow \mathcal{E} \cup \{\text{proposals}\}$ without duplicates, observed arcs taking precedence
$u_{ij}(0) \leftarrow u_{obs}$ (times e_{ij} with edge attributes) for observed arcs and u_{new} for the remaining candidates
return $\mathcal{E}_{cand}$, $\mathbf{U}(0)$

Algorithm 3 Label-free pseudo-partition of one training graph

Require: detached pre-ODE states $\mathbf{h}_1(0), \dots, \mathbf{h}_n(0)$ of the graph; cluster count K_{cfg}

$K \leftarrow \min(K_{cfg}, n)$ (a single cluster if $K = 1$) and $\mathbf{z}_i \leftarrow \mathbf{h}_i(0)/\|\mathbf{h}_i(0)\|_2$
$\mathbf{m}_k \leftarrow \mathbf{z}_{\iota_k}$ with $\iota_k = \operatorname{round}\big((k-1)(n-1)/(K-1)\big)$, $k = 1, \dots, K$
for 6 iterations **do**
    $c_i \leftarrow \arg\min_k \|\mathbf{z}_i - \mathbf{m}_k\|_2$ for every node i
    $\mathbf{m}_k \leftarrow \ell_2$-normalized mean of $\{\mathbf{z}_i : c_i = k\}$, unchanged if the cluster is empty
return the partition $\{c_i\}$

we compute

$$c_{ij} = \frac{\langle \mathbf{y}_i, \mathbf{y}_j \rangle}{\|\mathbf{y}_i\|_2 \|\mathbf{y}_j\|_2 + 10^{-12}}.$$

Given a threshold ρ_{ml}, pairs with $c_{ij} \ge \rho_{ml}$ are compatible and enter $\mathcal{P}$, while pairs with $c_{ij} < \rho_{ml}$ enter $\mathcal{N}$; pairs touching validation/test nodes are omitted from the supervised margin term or assigned only by the label-free pseudo-partition. The threshold is selected using the training/validation protocol only.

Label-free pseudo-partition. When no label partition is supplied, which is the case for graph-level tasks, Algorithm 3 partitions each training graph by k-means on its detached pre-ODE states. It runs without gradients and is never used at validation or test time. In the released implementation it is recomputed whenever the margin loss is evaluated, from the pre-ODE states of the current training forward pass, which include encoder dropout.

Table 3: Centralized force-margin pair construction. Pair labels are used only for the auxiliary margin loss; the task loss remains the primary supervision signal.

Regime	Positive pairs $\mathcal{P}$	Negative pairs $\mathcal{N}$	Labels used
Single-label node classification	Same training label among candidate pairs	Different training labels among candidate pairs	Training labels only
Multi-label node classification	Training-vector cosine $c_{ij} \ge \rho_{ml}$	Training-vector cosine $c_{ij} < \rho_{ml}$	Training label vectors only
Unlabeled, validation, or test nodes	Label-free pseudo-partition pairs, or omitted from supervised margin	Label-free pseudo-partition pairs, or omitted from supervised margin	No validation/test labels
Graph-level tasks	Same within-graph pseudo-cluster on training graphs	Different within-graph pseudo-clusters on training graphs	No graph labels for node pairs

B.4 Synthetic diagnostic protocol and perturbation tests

SBM generation. We generate graphs from a K-block SBM with node partition $\{C_1, \dots, C_K\}$ and block sizes $\{n_k\}_{k=1}^{K}$ ($\sum_k n_k = N$). For $i \in C_k$ and $j \in C_\ell$, edges are sampled independently as

$$\mathbf{A}_{ij}^{(0)} \sim \text{Bernoulli}(p_{k\ell}), \qquad p_{k\ell} = \begin{cases} p_{\text{in}}, & k = \ell, \\ p_{\text{out}}, & k \neq \ell, \end{cases}$$

with $p_{\text{in}} > p_{\text{out}}$ unless stated otherwise. Node features are initialized as

$$\mathbf{h}_i(0) = \mu_{c(i)} + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}),$$

where $c(i) \in \{1, \dots, K\}$ denotes the SBM block label.

Synthetic baselines and diagnostics. The soft-attention Graph ODE baseline constructs a strictly positive, row-stochastic transition matrix

$$\mathbf{P}_{ij}(t) = \frac{\exp(s_{ij}(t)/\tau_{\text{attn}})}{\sum_{k=1}^{N} \exp(s_{ik}(t)/\tau_{\text{attn}})}, \qquad s_{ij}(t) = \langle \mathbf{h}_i(t), \mathbf{h}_j(t) \rangle,$$

and evolves features via

$$\frac{d\mathbf{H}(t)}{dt} = -\big(\mathbf{I} - \mathbf{P}(t)\big)\mathbf{H}(t).$$

HGODE follows Eq. 8, with effective adjacency recovered by $\mathbf{A}_{ij}(t) = \sigma(\mathbf{U}_{ij}(t)/\tau)$. We validate Theorem 3.4 by sweeping τ_{attn} and tracking (i) representation mixing via intra- and inter-cluster distances, (ii) clustering quality via the silhouette score, and (iii) HGODE's mean edge potential $\mathbf{U}_{ij}(t)$ for intra- and inter-cluster pairs.

Synthetic perturbation tests. To evaluate robustness under feature corruption, we generate SBM graphs with increased cross-cluster connectivity by setting $p_{\text{out}} = 0.3$ while keeping other SBM parameters fixed. We sweep the noise standard deviation σ in the Gaussian feature initialization above, use 2000 graphs in total with an 80/20 train/validation split, and compare a DeepSets-style MLP baseline, the soft-attention Graph ODE baseline, and HGODE. For each model and noise level, we run 10 random seeds and train for 10 epochs, reporting validation accuracy as mean with shaded uncertainty.

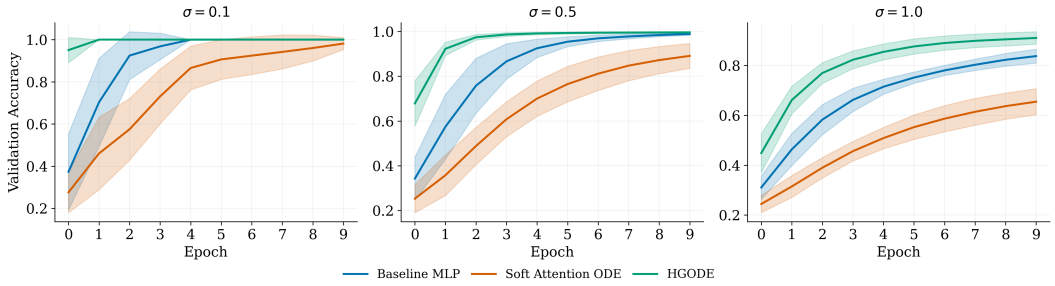


Figure 3: Validation accuracy under feature perturbations on SBM graphs ($\mu = 0.5$, $p_{\text{out}} = 0.3$). Subplots correspond to noise levels **Left**: $\sigma = 0.1$, **Middle**: $\sigma = 0.5$, **Right**: $\sigma = 1.0$. Solid lines denote mean over 10 seeds and shaded regions indicate ± 1 standard deviation.

Figure 3 shows that the soft-attention Graph ODE becomes less robust as σ increases, consistent with increased representation mixing under diffuse global attention. HGODE remains comparatively stable, indicating that dynamic topology can suppress spurious feature diffusion and recover informative signal even in low signal-to-noise regimes.

B.5 Graph transformer baselines

Graph transformers use high-capacity attention or subgraph-tokenization mechanisms that are complementary to the continuous-time dynamics studied in the main text. We choose four recent strong graph transformers, *i.e.*, GraphGPS (Rampášek et al., 2022), Polynormer (Deng et al., 2024), SGFormer (Wu et al., 2023a), and Subgraphormer (Bar-Shalom et al., 2024), and compare them with HGODE. We report them separately to keep Table 2 focused on message-passing and continuous-depth baselines.

Table 4: Comparison with graph-transformer baselines. Entries are mean ± standard deviation under matched splits and comparable parameter budgets when applicable. Roman-empire uses the ten fixed splits of Platonov et al. (2023); no protocol-matched node-classification implementation of Subgraphormer was available. The top three values are marked using the same convention as Table 2.

Model	Cora Acc.↑	Chameleon Acc.↑	Roman-empire Acc.↑	ogbn-proteins ROC-AUC↑	ZINC MAE↓	Peptides-func A.P.↑	ogbg-molpcba A.P.↑
GraphGPS	82.95±1.23	41.04±3.87	**83.38±0.76**	77.25±0.42	**0.069±0.008**	**0.655±0.004**	**0.291±0.001**
Polynormer	**83.24±0.72**	**41.69±3.22**	**92.06±0.62**	**79.41±0.56**	–	–	–
SGFormer	**84.59±0.43**	**45.77±3.48**	81.47±0.59	**79.84±0.61**	–	–	–
Subgraphormer	–	–	–	–	**0.082±0.04**	**0.661±0.005**	**0.293±0.004**
HGODE	**86.26±0.78**	**72.56±1.24**	**92.56±0.43**	**81.24±0.63**	**0.078±0.025**	**0.714±0.022**	**0.278±0.003**

B.6 Learned-topology diagnostics and rewiring comparisons

To inspect the operators that HGODE actually learns, we compare them with two static topology interventions on four node-classification datasets: curvature-based rewiring with BORF (Nguyen et al., 2023) and diffusion-based graph construction with GDC/PPR (Gasteiger et al., 2019). CiteSeer uses its public Planetoid split (Yang et al., 2016); Chameleon (Rozemberczki et al., 2021), Texas, and Wisconsin use Geom-GCN split columns 0–2 (Pei et al., 2020), paired with seeds 42, 1042, and 2042. This Chameleon protocol differs from the duplicate-filtered version used in Table 2. Node features are normalized and input graphs are symmetrized.

Protocol. Input, BORF, and GDC/PPR share the same two-layer GCN (16 hidden units, dropout 0.5, learning rate 0.01, weight decay 5×10^{-4}, 200 epochs), so their task differences isolate the static topology change. HGODE uses its own backbone (two 256-dimensional message blocks, dropout 0.5, `dopri5` on $[0, 1]$ with $\texttt{rtol} = \texttt{atol} = 10^{-5}$, at most 120 epochs with patience 20, learning rate 5×10^{-4} on CiteSeer and 10^{-4} elsewhere, weight decay 5×10^{-4}), so its task values are descriptive rather than a rewiring-only comparison. All three interventions are selected with validation accuracy and label-free topology guards only, and the selections are frozen before any test or label-dependent quantity is computed. Table 6 lists the selected settings and the resulting support changes.

Spectral diagnostics. For each operator we remove self-loops, symmetrize the weights as $\mathbf{W} = (\mathbf{A} + \mathbf{A}^\top)/2$, and form $\mathbf{L}_{sym} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$. We report the algebraic connectivity $\lambda_2(\mathbf{L}_{sym})$ and the number $N_{\lambda<10^{-2}}$ of nontrivial eigenvalues below 10^{-2}, both on the largest connected component, and the off-block mass, *i.e.*, the fraction of symmetrized weight that crosses a deterministic, graph-only K-way spectral partition with K equal to the number of classes. Smaller λ_2, more near-zero modes, and lower off-block mass indicate stronger near-decomposability; we use them as mechanism diagnostics, not as objectives. For HGODE, the spectra use the final soft operator $\mathbf{A}(T)$ of Eq. (10), one per split, while the support counts in Table 6 use the thresholded projection of Eq. (11).

Findings. The matched-GCN controls show no monotone relation between block structure and accuracy. On Wisconsin, BORF and GDC/PPR both improve accuracy over the input graph (from 59.48% to 63.40% and 65.36%), although BORF increases λ_2 and the off-block mass while GDC/PPR decreases both. On Chameleon, GDC/PPR produces the strongest near-decomposability (λ_2 from 0.00640 to 0.00165, near-zero modes from 2 to 9), yet accuracy slightly decreases (from 56.29% to 55.77%). On Texas, GDC/PPR lowers the off-block mass from 0.2509 to 0.0643 while accuracy is unchanged, although macro-F1 improves. HGODE does not maximize block structure either: it lowers the off-block mass on CiteSeer, Texas, and Wisconsin but adds no near-zero modes, and its λ_2 exceeds that of the input graph on Texas and Wisconsin. The learned weighted operators thus remain globally well connected while redistributing propagation weight, a pattern of selective reweighting rather than decomposition into blocks. This is consistent with Proposition 4.1, which characterizes the invariant subspace under exact support separation and makes no claim about downstream optimality. On the task side, HGODE's accuracy is below that of the matched GCN on CiteSeer and Chameleon and above it on Texas and Wisconsin, where its macro-F1 is also roughly twice as high; because the backbones differ, these values are not a rewiring-only comparison. With three runs and small, imbalanced Texas and Wisconsin test sets (37 and 51 nodes), all differences are descriptive.

Table 5: Learned-topology diagnostics. Accuracy and macro-F1 are mean ± standard deviation over three runs. Spectral quantities are computed on the frozen operators; for HGODE they are mean ± standard deviation over the three split-specific operators. Input, BORF, and GDC/PPR share the same two-layer GCN, whereas HGODE uses its own backbone. Arrows indicate stronger near-decomposability, not task desirability.

Dataset	Method	Acc. (%)↑	Macro-F1 (%)↑	$\lambda_2(\mathbf{L}_{sym})\downarrow$	$N_{\lambda<10^{-2}}\uparrow$	Off-block mass↓
CiteSeer	Input+GCN	70.43±0.86	66.22±0.46	0.00155	9	0.0212
	BORF+GCN	70.60±0.61	66.38±0.53	0.00169	7	0.0233
	GDC/PPR+GCN	70.57±0.81	66.97±1.32	0.00249	6	0.0274
	HGODE	73.87±1.10	68.22±1.27	0.00169±0.00023	7.7±1.2	0.0160±0.0045
Chameleon (Geom-GCN)	Input+GCN	56.29±4.82	55.29±4.42	0.00640	2	0.0089
	BORF+GCN	55.70±4.96	54.84±4.88	0.01362	0	0.0326
	GDC/PPR+GCN	55.77±1.46	55.22±1.77	0.00165	9	0.0130
	HGODE	55.19±0.33	54.97±0.39	0.00632±0.00023	2.0±0.0	0.0102±0.0010
Texas	Input+GCN	58.56±6.80	18.70±0.95	0.06323	0	0.2509
	BORF+GCN	58.56±8.69	19.33±6.07	0.06320	0	0.2500
	GDC/PPR+GCN	58.56±4.13	22.68±2.20	0.02429	0	0.0643
	HGODE	76.58±8.26	53.16±10.55	0.10671±0.01245	0.0±0.0	0.2035±0.0608
Wisconsin	Input+GCN	59.48±9.67	25.47±3.29	0.07652	0	0.2000
	BORF+GCN	63.40±6.89	27.79±1.12	0.11401	0	0.2589
	GDC/PPR+GCN	65.36±6.30	28.96±1.49	0.03750	0	0.0891
	HGODE	79.08±6.89	53.94±9.84	0.10289±0.01113	0.0±0.0	0.1773±0.0238

Table 6: Selected settings and support changes of the topology interventions in Table 5. The input graphs have 9,104 (CiteSeer), 62,742 (Chameleon), 558 (Texas), and 900 (Wisconsin) arcs. Change is $100(1-J)$, where J is the Jaccard index between the input arcs and the resulting arcs, and Added/Removed count arcs. For HGODE the resulting arcs are the thresholded projection of Eq. (11) at $\eta = 0.5\mu$ (mean ± standard deviation over three runs), while the forward operator remains a soft gate on $\mathcal{E}_{cand}$. BORF budgets are arcs added/removed per loop; HGODE anneals its gate temperature in all four settings.

Dataset	Method	Selected setting	Change (%)	Added	Removed
CiteSeer	BORF	3 loops; budget 31/15	2.97	186	90
	GDC/PPR	$\alpha = 0.15$, top-$k = 16$	82.53	41,130	329
	HGODE	$u_{obs}/u_{new} = 3/-0.72$; $\lambda = 0.25$, $s = 1.1$, $\beta = 0.15$, $\delta = 0.12$	13.89±12.78	1,621±1,540	0
Chameleon	BORF	3 loops; budget 157/157	2.58	696	942
	GDC/PPR	$\alpha = 0.15$, top-$k = 8$	79.90	2,770	49,573
	HGODE	$u_{obs}/u_{new} = 0.05/-0.3$; $\lambda = 0.4$, $s = 1.5$, $\beta = 0.3$, $\delta = 0.2$	18.21±6.88	2,140±2,214	9,715±4,155
Texas	BORF	3 loops; budget 3/0 (no removal)	3.12	18	0
	GDC/PPR	$\alpha = 0.3$, top-$k = 8$	69.30	849	126
	HGODE	$u_{obs}/u_{new} = 3/-0.3$; $\lambda = 0.25$, $s = 1.1$, $\beta = 0.15$, $\delta = 0.12$	23.72±29.45	280±414	0
Wisconsin	BORF	2 loops; budget 25/25	19.84	98	100
	GDC/PPR	$\alpha = 0.15$, top-$k = 16$	79.77	2,980	115
	HGODE	$u_{obs}/u_{new} = 3/-0.12$; $\lambda = 0.25$, $s = 1.1$, $\beta = 0.15$, $\delta = 0.12$	47.93±30.31	1,195±960	0

B.7 Efficiency profiling

Times in Table 7 are measured on one NVIDIA A100 40GB GPU and include the complete model-side training or inference process: candidate lookup/gating, ODE integration, task-head computation, loss/metric computation, and backward/optimizer steps for training. For node-classification datasets, the reported training time is milliseconds per full-graph epoch and inference time is milliseconds per full-graph evaluation pass. For graph-level datasets, times are milliseconds per mini-batch; each mini-batch is packed with as many graphs as fit into the 40GB memory budget under the corresponding model. Forward/backward NFEs are averaged over test batches for graph-level datasets and over the full-graph evaluation pass for node-level datasets.

Table 7: Efficiency profiling on one NVIDIA A100 40GB GPU in FP32 with adjoint training and hidden dimension 256. "–" denotes an unavailable NFE count for the corresponding fixed-flow implementation.

Dataset	Model	NFE		Time [ms]		Memory [MB]	
		Fwd.	Bwd.	Train	Infer.	Train	Infer.
Cora	HGODE	86	128	356.79	111.11	548.49	237.39
Cora	GRAND	82	117	324.64	109.45	509.69	179.18
Cora	FLODE	–	–	317.25	106.17	522.16	234.64
Cora	GREAD	76	124	327.16	98.87	549.34	201.68
Chameleon	HGODE	74	110	303.52	85.87	387.47	163.85
Chameleon	GRAND	77	102	312.48	89.44	331.34	83.16
Chameleon	FLODE	–	–	298.14	64.38	378.61	160.82
Chameleon	GREAD	81	106	316.37	68.19	364.34	146.37
ZINC	HGODE	88	127	2630.62	649.51	8078.65	4546.13
ZINC	GRAND	84	118	2232.48	526.18	6465.37	3589.16
ZINC	FLODE	–	–	2429.95	622.81	7801.24	4678.24
ZINC	GREAD	87	104	2559.34	619.13	7762.36	4518.39
Peptides-func	HGODE	92	134	2715.87	740.47	8051.21	4665.97
Peptides-func	GRAND	87	124	2348.34	529.16	6466.17	3674.13
Peptides-func	FLODE	–	–	2431.64	654.34	7762.33	4526.18
Peptides-func	GREAD	78	116	2498.36	668.36	7893.47	4329.16

Table 8: Roles of the HGODE hyperparameters; values are given in Tables 9–11.

Role	Hyperparameters	Rule
Implementation constants	two message blocks; force-MLP widths $(64, 32)$; $\epsilon_{\text{loop}} = 10^{-3}$, $\epsilon_{\text{deg}} = 10^{-6}$; `rtol=atol=` 10^{-5}; gradient clipping 1.0	Not tuned
Shared by all selected configurations	hidden width d, dropout p, τ_{feat}, τ_{topo}	Searched (Table 10); identical in Table 9
Constrained by the theory	$0 \le \lambda < 1$; $\tau_{feat}, \tau_{topo}, T > 0$; $s \ge \mathcal{F}_{\text{crit}}(\lambda) + \delta$ whenever $\beta > 0$	Required for a well-posed, feasible margin target
Validation-selected	λ, gate temperature τ, γ, s, δ, β, ρ_{ml}, T, learning rate	Tables 10 and 11
Candidate budget	k_{2hop}, k_{lap}, r_{rand}, together with the size caps of Algorithm 2	Selected on validation under a resource budget

B.8 Hyperparameter search and reproducibility

We use adaptive `dopri5` with `rtol=atol=`10^{-5} in all reported HGODE experiments. The optimizer is AdamW. Unless otherwise noted, non-GCN baselines are matched to roughly comparable parameter budgets by adjusting hidden dimensions. Candidate sets are constructed before integration by Algorithm 2. training labels, as described in Sec. 4.3. For Chameleon, we follow Platonov et al. (Platonov et al., 2023): the filtered variant removes duplicate nodes from the standard splits to avoid train–test leakage. For Roman-empire, we follow their ten fixed random 50/25/25 train/validation/test splits and report accuracy.

Table 8 groups the HGODE hyperparameters by role.

Table 9 is the compact reporting table for final selected configurations. All experiments are rerun 5 times and the seeds are all chosen as 42, 1042, 2042, 3042 and 4042. In addition, we attach the general hyperparameter tuning range in Table 10 and the recommended hyperparameter tuning range for each dataset in Table 11. Dataset provenance, benchmark splits, and available license or terms information are summarized in Table 12.

In the hyperparameters, k_{2hop} and k_{lap} are the number of 2-hop and Laplacian-based structural candidates per node, respectively. r_{rand} is the random candidate ratio per node. These three hyperparameters determine the candidate-pool size (Algorithm 2). In our chosen hyperparameter combinations, those values are chosen to reach good-enough performance without any additional

Table 9: Hyperparameter chosen for each reported benchmarks.

Dataset	Split	λ	τ	τ_{feat}	τ_{topo}	γ	s	δ	β	T	hid.	drop.	lr	k_{2hop}	k_{lap}	r_{rand}
Cora	Planetoid	0.3	0.2	1.0	1.0	0.05	1.0	0.1	0.3	0.3	256	0.5	5e-4	4	6	0.1
Chameleon	Chameleon-filtered	0.4	0.05	1.0	1.0	0.05	1.5	0.2	0.3	0.8	256	0.5	5e-4	4	6	0.1
Roman-empire	Platonov 50/25/25	0.5	0.3	1.0	1.0	0.05	1.0	0.2	0.3	1.0	256	0.5	5e-4	4	6	0.1
ogbn-proteins	OGB official	0.5	0.1	1.0	1.0	0.1	1.0	0.1	0.1	1.0	256	0.5	1e-4	6	12	1e-3
ZINC	PyG	0.5	0.1	1.0	1.0	0.05	1.0	0.1	0.2	1.0	256	0.5	5e-4	4	6	0.01
Peptides-func	LRGB	0.6	0.2	1.0	1.0	0.1	1.0	0.1	0.1	1.5	256	0.5	1e-4	6	12	0.05
ogbg-molpcba	OGB official	0.5	0.1	1.0	1.0	0.05	1.0	0.2	0.1	1.0	256	0.5	1e-4	4	6	0.1

Table 10: HGODE hyperparameter search space used for the reported experiments.

Group	Hyperparameter	Range
Structural / hysteresis	λ	$\{0.1, 0.3, 0.4, 0.5, 0.6, 0.8\}$
Structural / hysteresis	τ	$\{0.05, 0.1, 0.2, 0.3, 0.5\}$
Structural / hysteresis	τ_{feat}, τ_{topo}	$\{0.3, 0.5, 1.0\}$
Structural / hysteresis	γ	$\{0, 0.05, 0.1, 0.5\}$
Force / margin	s	$\{1, 1.5\}$
Force / margin	δ	$\{0.1, 0.2, 0.3, 0.5\}$
Force / margin	β	$\{0, 0.1, 0.2, 0.3, 0.5\}$
Force / margin	ρ_{ml}	$\{0.4, 0.5\}$
Candidate pool	r_{rand}	$\{0, 10^{-3}, 10^{-2}, 5\times10^{-2}, 10^{-1}\}$
Candidate pool	k_{2hop}	$\{0, 4, 6, 8, 12\}$
Candidate pool	k_{lap}	$\{0, 2, 4, 6, 8, 10, 12\}$
Backbone / optimization	`hidden_dim`	$\{128, 256, 512\}$
Backbone / optimization	`dropout`	$\{0.2, 0.5\}$
Backbone / optimization	`lr`	$\{10^{-4}, 5\times10^{-4}, 10^{-3}\}$
Backbone / optimization	T	$\{0.3, 0.6, 0.8, 1.0, 1.5, 2.0\}$

computational cost, *i.e.*, without need to increase the candidate-pool size. We run all the models with the same epochs: 80 epochs on Cora, Chameleon; 200 for the other datasets. The early-stopping patience is 10 epochs for all the training. We select the checkpoint and hyperparameters by validation performance and report the corresponding test performance. On the single-graph benchmarks, we run the models for a single graph per epoch, and for the multi-graph benchmarks, we load each batch to fulfill our experiment GPU memory budget (40 GB on single A100 GPU). Gradient clips are set to 1.0 for all the training. All the benchmarks results are reported with our rerun results.

C Bistability and Candidate-Pool Interpretation

Per-edge bistability is sufficient for multi-class structure because the local decision is binary: should this pair support feature mixing, or should it be insulated? A graph with many candidate edges then has many possible stable sign patterns, so global multistability arises compositionally from many locally bistable edge units. When intra-cluster edges polarize to the connected phase and cross-

Table 11: Representative HGODE starting subgrids by dataset regime. These are not additional search values; each entry is a subset of Table 10.

Dataset regime	Examples	Hysteresis subgrid	Force subgrid
Homophilous local	Cora, ZINC	$\lambda \in \{0.1, 0.3\}$; $\tau \in \{0.2, 0.3\}$; $\tau_{feat} = \tau_{topo} = 1.0$	$s = 1.0$; $\delta = 0.1$; $\beta \in \{0, 0.1\}$
Homophilous global	ogbg-molpcba	$\lambda \in \{0.3, 0.5\}$; $\tau \in \{0.1, 0.2\}$; $\tau_{feat} = \tau_{topo} = 1.0$	$s \in \{1.0, 1.5\}$; $\delta \in \{0.1, 0.2\}$; $\beta \in \{0.1, 0.3\}$
Heterophilous sparse	Chameleon, Roman-empire	$\lambda \in \{0.4, 0.5, 0.6\}$; $\tau \in \{0.05, 0.1\}$; $\tau_{feat} = \tau_{topo} = 1.0$	$s \in \{1.0, 1.5\}$; $\delta \in \{0.2, 0.3\}$; $\beta \in \{0.3, 0.5\}$
Heterophilous global	ogbn-proteins, Peptides-func	$\lambda \in \{0.5, 0.6, 0.8\}$; $\tau \in \{0.05, 0.1\}$; $\tau_{feat} = 1.0$; $\tau_{topo} \in \{0.5, 1.0\}$	$s \in \{1.0, 1.5\}$; $\delta \in \{0.2, 0.3\}$; $\beta \in \{0.1, 0.3\}$

Table 12: Dataset provenance, splits, and license or terms information. When an upstream benchmark wrapper or repository does not state a separate dataset license, we report that explicitly rather than inferring one.

Asset	Source / access path	Split used	License / terms reported by source
Cora	Planetoid / PyG citation-network wrapper; original Planetoid benchmark (Yang et al., 2016)	Planetoid public split	No separate dataset license stated in the PyG wrapper; used as a public research benchmark with citation to the original source.
Chameleon	WikipediaNetwork / PyG wrapper with duplicate filtering from Platonov et al. (Platonov et al., 2023); original Wikipedia networks (Rozemberczki et al., 2021)	Chameleon-filtered splits	Filtered benchmark and code are released in the Platonov et al. repository under MIT; original Wikipedia network source is cited.
Roman-empire	Platonov et al. heterophilous-graphs benchmark / PyG wrapper (Platonov et al., 2023)	10 random 50/25/25 splits	Platonov et al. repository is MIT-licensed; dataset is based on the English Wikipedia Roman Empire article with citation to the source benchmark.
ogbn-proteins	Open Graph Benchmark (Hu et al., 2020)	OGB official species split	CC-0 according to the OGB node-property benchmark page.
ZINC	PyG ZINC wrapper for the Benchmarking GNNs subset (Dwivedi et al., 2023)	Standard split	ZINC is distributed as a free molecular database; no separate PyG benchmark-subset license is stated.
Peptides-func	Long Range Graph Benchmark (Dwivedi et al., 2022)	LRGB official split	CC-BY-NC-4.0 according to the LRGB Hugging Face dataset card.
ogbg-molpcba	Open Graph Benchmark (Hu et al., 2020)	OGB official scaffold split	MIT according to the OGB graph-property benchmark page.
CiteSeer (App. B.6)	Planetoid / PyG citation-network wrapper; original Planetoid benchmark (Yang et al., 2016)	Planetoid public split	No separate dataset license stated in the PyG wrapper; used as a public research benchmark with citation to the original source.
Chameleon, Texas, Wisconsin (App. B.6)	WikipediaNetwork and WebKB / PyG wrappers with Geom-GCN splits (Pei et al., 2020); original Wikipedia networks (Rozemberczki et al., 2021)	Geom-GCN split columns 0–2	No separate dataset license stated in the PyG wrappers; used as public research benchmarks with citation to the original sources.

cluster edges polarize to the insulated phase, the effective propagation matrix becomes approximately block-structured even though each individual edge has only two phases.

The candidate pool is a scalability design rather than a separate modeling assumption. The full all-pairs version evolves $\mathbf{U}(t) \in \mathbb{R}^{N\times N}$ and is conceptually valid when resources permit. Sparse pools choose a tractable subset of this full matrix: observed edges preserve the input graph support, local proposals add multi-hop completion, Laplacian-eigenvector proposals add global relations, and random pairs provide exploration. Increasing the pool size therefore interpolates between graph-support polarization and dense topology search. In the notation of an implementation-level neighbor ratio r, the dense case is $r = 1$, where every ordered pair is included and HGODE reduces to dense-matrix topology evolution.

C.1 Coverage from small random proposal ratios

The purpose of a small random-pair ratio is not to include every possible long-distance edge, but to ensure that each relation group with many possible node pairs is likely to be represented at least once. The following elementary bound makes this statement explicit.

Proposition C.1 (Hitting relation groups with sparse proposals)**.** *Let $\mathcal{R}_1, \ldots, \mathcal{R}_K \subseteq \mathcal{V} \times \mathcal{V}$ be relation groups of interest, such as long-distance cross-cluster pairs, and let $m_k = |\mathcal{R}_k|$. Suppose each ordered pair is independently added to the random proposal pool with probability $r \in [0, 1]$. Then*

$$\Pr\left[\forall k,\ \mathcal{E}_{cand} \cap \mathcal{R}_k \neq \varnothing\right] \geq 1 - \sum_{k=1}^{K} (1-r)^{m_k}.$$

In particular, if $m_{\min} = \min_k m_k > 0$ *and* $\eta \in (0,1)$*, then* $r \geq \frac{1}{m_{\min}} \log \frac{K}{\eta}$ *is sufficient to cover every group with probability at least* $1 - \eta$*.*

Proof. A relation group $\mathcal{R}_k$ is missed only if none of its m_k ordered pairs is sampled, which has probability $(1-r)^{m_k}$. A union bound over K groups gives the first inequality. Since $(1-r)^{m_k} \leq e^{-rm_k} \leq e^{-rm_{\min}}$, choosing $r \geq m_{\min}^{-1} \log(K/\eta)$ makes the failure probability at most η. □

This bound explains why small ratios can cover long-distance relation *types* when each type contains many possible node pairs, as is common in large graphs. 2-hop and Laplacian-eigenvector proposals provide short- and mid-range coverage, while the random component gives nonzero probability to far-apart pairs that would otherwise be absent from a local candidate pool.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The title, abstract, and introduction claim mitigation of global consensus collapse and state the positivity assumptions; Table 1 separates asymptotic global consensus, idealized block-wise consensus, and finite-horizon mitigation by the implemented soft gates, matching Secs. 3–5.2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 6 discusses the scope of the consensus theory (consensus-form propagation and no global convergence guarantee for the learned coupled system), the evaluated operator family, directed-graph coverage, candidate-pool cost, and graph-structure-learning risks; Table 1 states the finite-horizon scope of the soft-gate result.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Theoretical assumptions are stated with each theorem/lemma in Sec. 3, and the corresponding proofs appear in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The appendix now gives architecture, algorithm, solver, search ranges, efficiency protocol, split/provenance fields, and the selected-configuration reporting table in Table 9.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We use public benchmark datasets and provide algorithmic, solver, hyperparameter, split, and reporting details in Appendix B.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper reports optimizer, solver tolerances, benchmark splits, architecture, search ranges, selected hyperparameters, seeds, candidate-pool settings, and efficiency protocol in Appendix B.8.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Main benchmark tables report mean $\pm$ standard deviation, and the synthetic perturbation study states that shaded regions are ± 1 standard deviation over 10 seeds.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B.7 reports the A100 40GB setting, NFE, time, memory, and the measurement unit for node- and graph-level experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses standard public graph benchmarks and method development; Sec. 6 discusses responsible use of learned graph relations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Sec. 6 discusses potential benefits for graph learning and risks from learned or inferred graph relations, including bias amplification and causal misinterpretation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-risk foundation model, scraped dataset, or other asset requiring special misuse safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Table 12 lists the benchmark sources, splits, and available license or terms information; when an upstream wrapper does not specify a separate dataset license, the table states this explicitly and the paper uses the public benchmark split with citation.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper introduces a method and experimental implementation, but the current manuscript does not release a new dataset or standalone asset requiring separate documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The experiments do not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The experiments do not involve crowdsourcing or human subjects, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the core HGODE methodology, experiments, or scientific contribution described in the paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.