Contractive Monoids: The Algebra Behind Stable Residual Propagation in Deep Graph Neural Networks

Rohit Sinha [1] **Saroj Kumar** [2]

Abstract

Graph neural networks exhibit a fragile relationship with depth, as adding layers leads to rapid homogenization of node representations (over-smoothing). Recent manifold-constrained approaches using doubly stochastic matrices address this issue, but the fundamental principle underlying their success remains unclear. This study investigates whether doubly stochasticity is essential or exemplifies a broader geometric principle for stable residual propagation in deep GNNs.We formalize three axioms to define residual mixing matrices as contractive monoids. We instantiate this framework through three manifolds (Orthogonal Group, Diagonal Contraction Monoid, Spectral Norm Ball) and evaluate them across nine node-classification benchmarks using four GNN backbones. Results demonstrate that contractive-monoid constraints maintain stable accuracy and low representational collapse at extreme depths, while standard GNNs degrade sharply beyond moderate depth. This work establishes that stable residual propagation arises from contractive, compositional structure anchored at identity, revealing a significantly larger design space for deep GNN architectures while maintaining theoretical guarantees.

1. Introduction

Depth has been a reliable lever for progress in vision and language models (He et al., 2016a; Vaswani et al., 2017; Dosovitskiy et al., 2021), but graph neural networks (GNNs) respond to it differently. Adding layers often causes node representations to collapse toward indistinguishable vectors: a phenomenon known as over-smoothing, even when optimization remains stable and gradients flow normally (Li et al., 2018; Oono & Suzuki, 2020; Zhao & Akoglu, 2020; Chen et al., 2020a). In practice, this has limited most message-passing architectures to shallow configurations.

Residual connections were designed to counteract such degradation by preserving identity mappings across layers (He et al., 2016a), and they are widely used in GNNs (Kipf & Welling, 2017; Xu et al., 2018; Klicpera et al., 2019). However, neighborhood aggregation already contracts node features in graph settings (Cai & Wang, 2020; Rusch et al., 2023), and as depth grows, the residual branch itself participates in this contraction rather than bypassing it.

Recent work on Manifold-Constrained Hyper-Connections (mHC) addresses this by restricting residual mixing matrices to the Birkhoff polytope of doubly stochastic matrices, ensuring mean preservation, bounded propagation, and compositional closure (Xie et al., 2026). This geometric constraint enables stable behavior at extreme depths in language models, and Mishra (2026) adapted it to GNNs, showing that doubly stochastic mixing mitigates over-smoothing and extends expressiveness beyond 1-WL.

These results raise a natural question: is doubly stochasticity with its specific properties of mean preservation, non-negativity, and permutation-matrix decomposition the right structural choice, or does it happen to satisfy a more general principle? The answer has direct design implications. If doubly stochasticity is essential, the Birkhoff polytope is the canonical constraint and future work should refine within it. If it is one instance of a broader class, then a larger design space exists with the same stability guarantees.

We argue for the latter. The operative principle is *contractivity*: stable residual propagation arises whenever mixing operators are non-expansive, closed under composition, and anchored at identity. Together, these three properties define a *contractive monoid*: an algebraic structure that doubly stochastic matrices satisfy, but that is also satisfied by other operator families with different expressiveness–efficiency trade-offs. Individual ingredients of this framework (spectral normalization, orthogonal recurrence, residual identity mappings) each have precedent in deep learning, but the

[1]IIT Hyderabad, India [2]DCG Datacore Systems India, Pvt. Ltd. Correspondence to: Rohit Sinha <rohit123sinha456@gmail.com>.

Accepted to Graph Foundation Models: A New Era for Graph Machine Learning, an ICML 2026 Workshop, Copyright 2026 by the author(s).

compositional residual mixing setting requires all three simultaneously; a requirement absent when constraining individual weight matrices.

We formalize this observation and instantiate it through three manifold families, demonstrating both theoretically and empirically that the contractive monoid structure, not doubly stochasticity, is what governs stable depth scaling.

Contributions

- **Axiomatic framework for stable residual mixing.** We identify three axioms: non-expansiveness, compositional closure, and identity anchoring; and prove that their conjunction is jointly sufficient for stable residual mixing in multi-stream GNNs, with each axiom underpinning a distinct guarantee in our analysis(Sections 3, 4).
- **Generalized architecture family.** We instantiate this framework through three contractive monoids (Orthogonal Group, Diagonal Contraction, Spectral Norm Ball) that offer different expressiveness–structure trade-offs while sharing the same theoretical stability guarantees (Section 4).
- **Birkhoff as a special case.** We show that doubly stochastic mixing is a structured subset of the spectral norm ball, clarifying why it works and revealing a strictly larger design space for deep GNN architectures (Section 3.1).

2. Related Work

Depth Degradation and Local Mitigations Over-smoothing: the convergence of node representations under repeated neighborhood aggregation has been analyzed extensively from spectral and dynamical perspectives (Li et al., 2018; Oono & Suzuki, 2020). These analyses show that graph convolution induces a diffusion process governed by the spectral gap of the normalized adjacency operator, explaining why deep GNNs lose discriminative power.

A range of architectural interventions target this effect. Residual and dense connections (He et al., 2016a; Huang et al., 2018) improve gradient flow; normalization methods such as PairNorm and DiffGroupNorm (Zhao & Akoglu, 2020; Zhou et al., 2020) counteract feature collapse; DropEdge (Rong et al., 2020) stochastically sparsifies the graph during training; and GCNII (Chen et al., 2020b) reintroduces initial features via an identity-anchored residual:

$$H^{(l)} = (1-\alpha)\tilde{A}H^{(l-1)} + \alpha H^{(0)}. \tag{1}$$

GCNII provides partial identity anchoring by preserving initial features, but does not enforce non-expansiveness on the mixing pathway, which helps explain its effectiveness at moderate depths and degradation beyond 16 layers (Table 3).

Over-squashing, which compresses long-range dependencies through bottlenecks (Giraldo et al., 2023), is addressed by graph rewiring (Gutteridge et al., 2023; Qian et al., 2024) but alters the input graph.

These methods address depth-related failure through local corrections. None constrains the algebraic properties of residual operators under composition, leaving open whether stability can be enforced structurally across depth.

Expressiveness Beyond 1-WL Higher-order GNNs (Morris et al., 2019), subgraph methods (Bouritsas et al., 2022), and positional encodings (Dwivedi & Bresson, 2021) surpass the 1-WL barrier through explicit augmentation or higher-order computation. Whether expressiveness can instead emerge from internal architectural structure is less explored.

Multi-Stream and Manifold-Constrained Architectures Expanding residual pathways dates back to Highway Networks (Srivastava et al., 2015) and dense aggregation schemes. Hyper-Connections (Zhu et al., 2024) formalized multi-stream residual mixing with learnable connection matrices, but unconstrained mixing can compromise the identity-mapping property needed for stable deep training.

Manifold-Constrained Hyper-Connections (mHC) (Xie et al., 2026) addressed this by projecting mixing matrices onto the Birkhoff polytope of doubly stochastic matrices, ensuring bounded propagation and mean preservation. This was adapted to GNNs by Mishra (2026), yielding theoretical guarantees and strong depth scaling. Our work builds on this foundation: rather than focusing on doubly stochasticity alone, we view mHC as one instance of a broader principle in which stability arises from contractive, compositional structure anchored at identity.

Table 1. Mapping of over-smoothing methods to contractive monoid axioms. Our framework is the only approach that constrains the residual mixer to satisfy all three axioms simultaneously.

Method	Ax. 1	Ax. 2	Ax. 3	Operates On
PairNorm / DGN	×	×	×	Node features (post-hoc)
GCNII	×	×	✓	Feature residual
DropEdge	×	×	×	Graph topology
DRew / Rewiring	×	×	×	Graph topology
mHC-GNN (Birkhoff)	✓	✓	✓	Residual stream mixer
Ours	✓	✓	✓	Residual mixer (generalized)

Operator Constraints for Stability

Several lines of work constrain neural network operators for stability: spectral normalization of individual weights (Miyato et al., 2018), orthogonal recurrence for gradient stability (Arjovsky et al., 2016; Helfrich et al., 2018), and global Lipschitz bounds for robustness (Anil et al., 2019). Each addresses one or two of our axioms; none imposes all three simultaneously on residual stream mixers (Appendix E.4).

Table 1 maps representative methods to our axioms, showing that existing over-smoothing interventions each satisfy at most a subset. The contractive monoid perspective identifies *which* algebraic property each method satisfies, clarifying both why it works and where it falls short.

3. Contractive Residual Mixing: An Axiomatic Framework

We seek to characterize the minimal structural properties that a set of residual mixing matrices must satisfy for stable composition at arbitrary depth. Consider a deep GNN with L layers, where each layer applies a residual mixer H_ℓ to route information across parallel representation streams. The effective transformation after L layers is $\prod_{\ell=1}^{L} H_\ell$, a long chain of matrix multiplications whose behavior must remain controlled as L grows. We identify three properties that jointly guarantee this: each mixer should not amplify signals (*non-expansiveness*), the composition of any two admissible mixers should remain admissible (*closure*), and the identity mapping should be reachable (*identity anchoring*). Together, these define a *contractive monoid*, an algebraic structure that guarantees depth-invariant stability while leaving maximal freedom in the choice of operator family.

Doubly stochastic mixing (Xie et al., 2026; Mishra, 2026) satisfies all three properties, but so do other operator families with different expressiveness–structure trade-offs.

Multi-stream setup and notation. Following (Mishra, 2026), each node i carries a multi-stream representation $\mathbf{x}_i \in \mathbb{R}^{n\times d}$, where $\mathbf{x}_i = [\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(n)}]$ and n is a hyperparameter controlling the number of streams. The layer update for node i at layer l is:

$$\mathbf{x}_i^{l+1} = H_\ell\, \mathbf{x}_i^l + \mathcal{H}_{i,l}^{\text{post}}\, \text{GNN}\Big(\mathcal{H}_{i,l}^{\text{pre}}\, \mathbf{x}_i^l,\ \{\mathbf{x}_j^l : j \in \mathcal{N}_i\};\ \mathcal{W}^l\Big), \tag{2}$$

where $H_\ell \in \mathbb{R}^{n\times n}$ is the residual stream mixing matrix that we constrain to the axiomatic manifolds, and $\mathcal{H}_{i,l}^{\text{pre}}, \mathcal{H}_{i,l}^{\text{post}}$ are projection matrices for stream aggregation. The parametrization of H_ℓ and the projection matrices follows Mishra (2026), adapting the mapping from Xie et al. (2026): raw mixer entries are produced by a lightweight input-dependent network (RMSNorm, sigmoid gating, learnable scalars initialized near zero) and then projected onto the chosen manifold. Full parametrization details are given in Appendix A.

We now specify the three axioms that H_ℓ must satisfy.

Axiom 1: Non-Expansiveness

[Non-Expansiveness] Each residual mixer must be non-expansive under an appropriate operator norm:

$$\|H_\ell\| \le 1. \tag{3}$$

Non-expansiveness prevents composed residual operators $\prod_{\ell=1}^{L} H_\ell$ from exhibiting exponential norm growth, which would otherwise cause gradient explosion or representational blow-up. In the GNN setting, it also controls how the residual pathway interacts with feature diffusion: by bounding the mixer's contribution, we decouple the graph-induced contraction rate (governed by spectral gap) from any amplification the mixer might introduce, allowing deeper networks without premature feature collapse. We formalize this decoupling in Section 4.

Axiom 2: Compositional Closure

[Compositional Closure] The family of admissible mixers $\mathcal{M}$ must be closed under composition:

$$H_a, H_b \in \mathcal{M} \implies H_a H_b \in \mathcal{M}. \tag{4}$$

Closure guarantees that stability holds *globally* across arbitrary depths, not merely locally at each layer. If $\mathcal{M}$ is closed and all $H_\ell \in \mathcal{M}$ satisfy $\|H_\ell\| \le 1$, then $\prod_{\ell=1}^{L} H_\ell \in \mathcal{M}$ and $\|\prod_{\ell=1}^{L} H_\ell\| \le 1$ for any depth L.

Without closure, structural properties can degrade under composition even when norm bounds are preserved. For example, the set of symmetric matrices with $\|H\|_2 \le 1$ is individually non-expansive, yet the product of two symmetric matrices is not generally symmetric. The norm bound survives, but the symmetry a designer may have chosen for interpretability is lost in the effective multi-layer mixer. Closure prevents such pathologies: all structural guarantees survive composition, not just the norm bound.

Axiom 3: Identity Anchoring

[Identity Anchoring] The identity matrix must belong to $\mathcal{M}$:

$$I \in \mathcal{M}. \tag{5}$$

Identity anchoring ensures that networks can initialize near the identity mapping, preserving the residual learning principle (He et al., 2016b): when $H_\ell \approx I$, the update reduces to $\mathbf{x}^{(\ell+1)} \approx \mathbf{x}^{(\ell)} + F_\ell(\mathbf{x}^{(\ell)})$, allowing gradients to flow unimpeded during early training. It also guarantees that adding depth cannot reduce model capacity; deeper networks can always recover shallower behavior by setting additional mixers to identity (Appendix D). During training, mixers learn to deviate from identity where beneficial while retaining the option to preserve information flow, which is particularly important for heterogeneous graphs where different regions may require different mixing strategies.

Formal Definition: Together, the three axioms define a contractive monoid:

Definition 3.1 (Contractive Monoid)**.** A set $\mathcal{M}$ of matrices forms a *contractive monoid* under matrix multiplication if:

1. **(Closure)** $H_a, H_b \in \mathcal{M} \implies H_a H_b \in \mathcal{M}$

2. **(Identity)** $I \in \mathcal{M}$

3. **(Contraction)** $\|H\| \leq 1$ for all $H \in \mathcal{M}$

The choice of manifold constraint is thus a choice of contractive monoid. Different monoids trade off expressiveness (how large is $\mathcal{M}$?) against structure (what additional properties does $\mathcal{M}$ guarantee?).

3.1. Instantiations of Contractive Monoids

We study three instantiations that span this design space.

Orthogonal Group. The orthogonal group $\mathcal{O}(n) = \{H \in \mathbb{R}^{n\times n} \mid H^\top H = I\}$ enforces norm-preserving residual mixing. Its elements compose without amplification, yielding maximal stability under depth. However, residual updates are restricted to rotations and reflections, which cannot attenuate or selectively suppress individual streams.

Diagonal Contraction Monoid. Diagonal contraction operators $\mathcal{D} = \{\mathrm{diag}(\lambda_1, \ldots, \lambda_n) \mid |\lambda_i| \leq 1\}$ permit anisotropic attenuation across streams while preserving compositional stability. This structure enables controlled per-stream decay but forbids inter-stream mixing, limiting expressiveness despite its simplicity.

Spectral Norm Ball. The spectral norm ball $\mathcal{B} = \{H \in \mathbb{R}^{n\times n} \mid \|H\|_2 \leq 1\}$ admits arbitrary linear mixing subject only to global non-expansiveness. It imposes no structural bias beyond contraction, allowing the richest class of residual transformations compatible with stable composition: a maximally expressive contractive monoid.

Formal verification that each instantiation satisfies the contractive monoid axioms is deferred to Appendix C.

Birkhoff as a special case of spectral contraction. The doubly stochastic manifold used in mHC can be viewed as a structured subset of the spectral norm ball. Let $\mathcal{B}_n$ denote the Birkhoff polytope, the set of doubly stochastic matrices with $H\mathbf{1} = \mathbf{1}$, $\mathbf{1}^\top H = \mathbf{1}^\top$, and $H \geq 0$ (Xie et al., 2026). By the Birkhoff–von Neumann theorem, every $H \in \mathcal{B}_n$ is a convex combination of permutation matrices $H = \sum_k \alpha_k P_k$. Since each permutation matrix is orthogonal ($\|P_k\|_2 = 1$), convexity of the operator norm gives $\|H\|_2 \leq 1$. Hence $\mathcal{B}_n \subseteq \{H : \|H\|_2 \leq 1\}$: the Birkhoff constraint enforces spectral non-expansiveness while additionally imposing non-negativity and exact row/column-sum conservation. Doubly stochastic matrices are therefore a more structured, strictly smaller instance of the same contraction principle underlying the spectral norm ball (Xie et al., 2026; Mishra, 2026).

Running Example: Birkhoff vs. Spectral Norm Ball

Consider $n = 2$ streams. A Birkhoff mixer has the form

$$H_B = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\alpha & \alpha \end{pmatrix}, \quad \alpha \in [0,1], \tag{6}$$

which interpolates between identity ($\alpha = 1$) and full stream-swap ($\alpha = 0$). Every output is a convex combination of the two inputs—no stream can be selectively attenuated.

A spectral-norm-ball mixer only requires $\|H\|_2 \leq 1$, admitting for instance

$$H_S = \begin{pmatrix} 0.9 & 0.3 \\ -0.2 & 0.8 \end{pmatrix}, \quad \|H_S\|_2 \approx 0.98 \leq 1, \tag{7}$$

which allows asymmetric mixing, sign flips, and differential attenuation—operations impossible under doubly stochastic constraints. Both H_B and H_S satisfy all three axioms and carry identical stability guarantees. The additional representational freedom of H_S is what benefits tasks requiring selective information routing (Section 5).

4. Theoretical Analysis: Stability and Expressiveness

We now analyze how contractive residual mixing addresses the two core limitations of deep GNNs. On the stability side, we provide a manifold-agnostic perturbation bound (Theorem 4.1) and generalize the multi-stream over-smoothing slowdown beyond Birkhoff-specific constraints (Theorem 4.3). On the expressiveness side, we bound stream-level diversity collapse (Theorem 4.5) and show that contractive-monoid mixing on an edge lift strictly exceeds 1-WL (Theorem Theorem 4.6). The unifying technical contribution is that all four results depend only on the contractive-monoid axioms, not on doubly stochastic structure: the Birkhoff-specific analyses of (Mishra, 2026) and (Xie et al., 2026) emerge as one instantiation of bounds and rates that hold across the full design space of Section 3.

4.1. Perturbation Stability: Decoupling the Mixer from the Backbone

Standard analyses bound over-smoothing by tracking how graph convolution contracts node feature distances. We take a different approach: we bound the perturbation amplification introduced by the residual mixer itself, showing that contractive constraints prevent any depth-dependent degradation from that component.

Theorem 4.1 (Perturbation Stability with Contractive Residual Mixing)**.** *Consider the depth-L architecture*

$$h^{(\ell+1)} = H_\ell\, h^{(\ell)} + G_\ell\, F_\ell\big(K_\ell\, h^{(\ell)}\big), \qquad \ell = 0, \ldots, L-1, \tag{8}$$

where H_ℓ is the residual mixer, F_ℓ is an L_ℓ-Lipschitz GNN layer, and K_ℓ, G_ℓ are linear stream projection/aggregation

maps. Assume $\|H_\ell\|_2 \leq 1$ for all ℓ. Then for any initial perturbation $\delta h^{(0)}$, the perturbation at depth L satisfies

$$\|\delta h^{(L)}\|_2 \;\leq\; \|\delta h^{(0)}\|_2 \cdot \prod_{\ell=0}^{L-1} \Big(1 + L_\ell \, \|G_\ell\|_2 \, \|K_\ell\|_2\Big). \quad (9)$$

The key observation is that H_ℓ does not appear in the bound: the residual mixer contributes zero amplification as long as $\|H_\ell\|_2 \leq 1$, regardless of graph structure or message-passing details inside F_ℓ. The only source of perturbation growth is the GNN layer itself, through the factor $(1 + L_\ell \|G_\ell\|_2 \|K_\ell\|_2)$. This decoupling is what makes the framework manifold-agnostic (any contractive monoid yields the same bound) and explains why the stability pattern persists across different GNN backbones in our experiments (Section 5).

This also formalizes the distinction between two rates governing feature diffusion in deep GNNs: the *diffusion rate*, at which graph convolution contracts pairwise node distances (governed by the spectral gap γ of the propagation operator), and the *mixing rate*, the per-layer perturbation from the residual mixer (bounded by $\|H_\ell - I_n\|_2 \leq \epsilon$). Non-expansiveness ensures that the mixer cannot amplify the diffusion-driven contraction, keeping these two sources of signal loss independent. Proof in Appendix D.2.

4.2. Multi-Stream Over-Smoothing Slowdown

The slowdown arises from two mechanisms. First, uniform aggregation over n streams dilutes pairwise differences by a factor of $1/\sqrt{n}$ (Lemma 4.2, proof in Appendix D.3). Second, each layer applies a fractional propagation step $P^{1/n}$, so n layers are needed to match one single-stream contraction step.

Lemma 4.2 (Stream Aggregation Dilution)**.** *Let $H_\ell \approx \frac{1}{n}\mathbf{1}_n^\top$ aggregate n streams uniformly. For any nodes i, j: $\|\widetilde{x}_i^{(\ell)} - \widetilde{x}_j^{(\ell)}\|_F \leq \frac{1}{\sqrt{n}}\|x_i^{(\ell)} - x_j^{(\ell)}\|_F$, where $\widetilde{x}_i^{(\ell)} = H_\ell^{pre} x_i^{(\ell)}$ is the aggregated representation.*

Theorem 4.3 (Multi-stream over-smoothing rate)**.** *Consider a multi-stream GNN with n streams and depth L, whose per-layer propagation operator P has spectral contraction $(1-\gamma)$ and whose residual stream mixers satisfy $\|H_\ell^{\text{res}} - I_n\|_2 \leq \epsilon$. Then for any nodes i, j,*

$$\|x_i^{(L)} - x_j^{(L)}\|_F \;\leq\; C \cdot (1-\gamma)^{L/n} \cdot (1+\epsilon)^L, \quad (10)$$

where C depends only on the initial representation diversity.

Neither mechanism depends on doubly stochastic structure. The Birkhoff polytope is a sufficient instantiation, but any contractive monoid with identity anchoring yields the same asymptotic behavior, generalizing the Birkhoff-specific result of Mishra (2026). Full proof in Appendix D.4.

4.3. Expressiveness via Information Preservation

Beyond stability, contractive monoids preserve *information diversity* across streams, preventing the representational collapse that would reduce a multi-stream architecture to a single-stream MPNN.

Definition 4.4 (Stream Diversity)**.** For a multi-stream representation $x \in \mathbb{R}^{n\times d}$ with streams $x_1, \ldots, x_n \in \mathbb{R}^d$, define the stream mean $\bar{x} := \frac{1}{n}\sum_{s=1}^n x_s$ and diversity $\mathcal{D}(x) := \frac{1}{n}\sum_{s=1}^n \|x_s - \bar{x}\|_2^2$.

Theorem 4.5 (Diversity Preservation Under Near-Identity Mixing)**.** *Let $H \in \mathbb{R}^{n\times n}$ satisfy: (i) $\|H\|_2 \leq 1$, (ii) $H\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top H = \mathbf{1}^\top$ (mean-preserving), and (iii) $\|H - I_n\|_2 \leq \epsilon$ for some $\epsilon \in [0, 1)$. Then for all $x \in \mathbb{R}^{n\times d}$,*

$$\mathcal{D}(Hx) \;\geq\; (1-\epsilon)^2\, \mathcal{D}(x) \;\geq\; (1-2\epsilon)\, \mathcal{D}(x). \quad (11)$$

When ϵ is small uniformly across depth (as guaranteed by identity anchoring), stream diversity cannot collapse rapidly. Since $\mathcal{D}(x) \to 0$ would degenerate multi-stream hyper-connections to a single-stream MPNN, this result links contractive monoids to expressiveness: the algebraic constraints that ensure stability also prevent the loss of representational capacity at depth. This is confirmed empirically by the variance diagnostics in Section 5.3.

4.4. Beyond 1-WL via Multi-Scale Signatures

Theorem 4.6 (Beyond 1-WL via contractive-monoid mixing on an edge lift)**.** *There exist graphs G, H that are indistinguishable by 1-WL, but are distinguished by a two-stream GNN operating on directed edges (an edge-lift) with per-layer stream mixers constrained to any contractive monoid $\mathcal{M}$ that contains I_2 and a non-trivial permutation matrix.*

When the contractive monoid contains a non-trivial permutation, the multi-stream architecture on an edge lift can simulate a 2-WL refinement step, strictly exceeding 1-WL expressiveness. A constructive proof is in Appendix D.6.
Scope of the expressiveness gain: The hypothesis that $\mathcal{M}$ contains a non-trivial permutation matrix is satisfied by the orthogonal group, the spectral norm ball, and the Birkhoff polytope, but not by the diagonal contraction monoid (whose only permutation element is I_n). The strict $> 1 - WL$ guarantee of Theorem 4.6 therefore applies to three of our four instantiations; diagonal contraction trades this expressiveness for its $O(n)$ mixing cost, consistent with the expressiveness–structure trade-off identified in Section 3.1.

5. Experiments

We test whether the contractive monoid framework predicts empirical behavior: (i) do Axioms 1–3 suffice for stable deep training, (ii) does Birkhoff behave as one instance of a

Table 2. Test accuracy (%) of an 8-layer GCN under different residual mixing constraints. Results are reported as mean ± standard deviation over multiple runs. Constrained residual mixers from different contractive monoids (Birkhoff, Diagonal, Orthogonal, Spectral) consistently match or outperform unconstrained hyper-connections across both homophilic and heterophilic datasets, while substantially improving over the standard GCN baseline. Best performance per dataset is shown in **bold.**

Dataset	Unconstrained	Birkhoff	Diagonal	Orthogonal	Spectral	Standard
chameleon	29.92 ± 1.71	30.34 ± 2.82	30.65 ± 1.08	30.92 ± 2.49	**31.51 ± 2.37**	24.27 ± 0.33
texas	64.67 ± 6.24	66.96 ± 1.56	66.16 ± 1.81	66.67 ± 4.13	**67.57 ± 2.70**	62.16 ± 9.36
actor	27.18 ± 1.46	**28.97 ± 1.16**	28.88 ± 0.40	28.90 ± 1.28	28.53 ± 1.11	24.23 ± 1.16
Cora	74.87 ± 1.05	74.67 ± 1.05	75.23 ± 1.47	74.90 ± 0.87	**75.30 ± 0.98**	70.53 ± 0.74
CiteSeer	61.80 ± 4.25	**63.67 ± 1.33**	62.53 ± 1.21	62.97 ± 1.17	63.23 ± 0.67	54.53 ± 2.49
PubMed	76.10 ± 0.30	77.20 ± 1.15	**77.67 ± 1.18**	76.47 ± 0.12	77.20 ± 0.20	74.17 ± 2.87
cornell	43.15 ± 1.56	**46.85 ± 1.56**	45.54 ± 0.00	45.05 ± 1.56	45.34 ± 1.56	40.54 ± 0.00
wisconsin	66.28 ± 4.53	67.32 ± 2.99	66.86 ± 1.96	**69.32 ± 3.00**	67.97 ± 5.66	53.59 ± 2.99

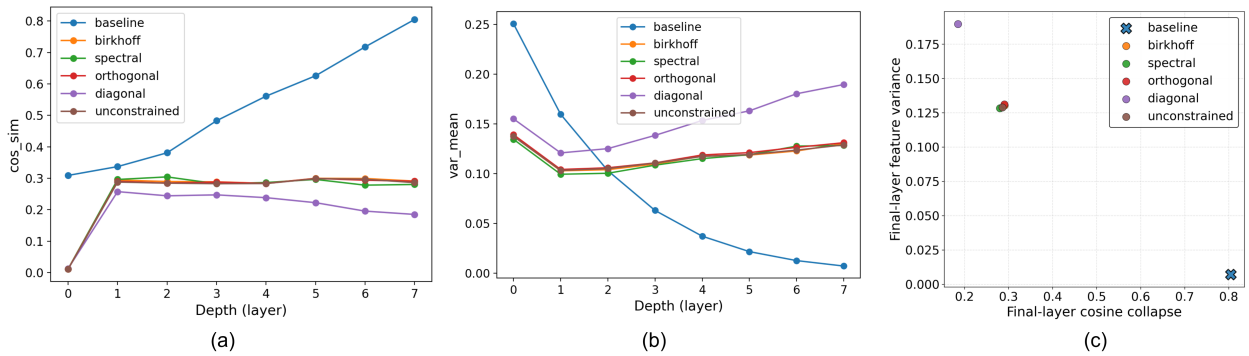

Figure 1. **Embedding stability diagnostics** (a) Cosine similarity vs. depth shows that the baseline GCN exhibits rapid representational collapse as depth increases, a signature of oversmoothing. In contrast, contractive-manifold variants maintain consistently low and depth-stable cosine similarity, indicating controlled non-expansive composition. (b) Mean feature variance vs. depth reveals that the baseline GCN's embeddings lose variance exponentially with depth, whereas contractive models preserve a non-vanishing variance floor, reflecting stable information propagation. (c) Final-layer collapse–variance scatter (cosine similarity vs. variance) isolates the baseline GCN in the high-collapse/low-variance regime, while contractive models cluster in a low-collapse/high-variance region, demonstrating that stability arises from contractive, identity-anchored residual composition rather than architectural depth alone.

broader contractive family, and (iii) do different contractive monoids produce consistent stability patterns?

Datasets. We evaluate on nine node-classification benchmarks: Cora, CiteSeer, PubMed (homophilic); Chameleon, Actor (heterophilic, medium); Texas, Wisconsin, Cornell (heterophilic, small). Additional results on modern heterophilic benchmarks (Platonov et al., 2023), LRGB graph-level tasks (Dwivedi et al., 2022), and OGBN-ARXIV (169K nodes) appear in Appendix B.5 and confirm the same patterns across 18 total benchmarks.

Architectures and Implementation. We integrate contractive residual mixing with four backbones: GCN (Kipf & Welling, 2017), GraphSAGE (William L. Hamilton, 2017), GAT (Veličković et al., 2018), and GIN (Xu et al., 2019). All models use 8 layers, hidden dimension 128, n=4 streams, and train for 500 epochs with Adam; full hyperparameters are in Table **??**. We report mean ± std over 5 seeds. We do not claim state-of-the-art absolute accuracies: the question is whether the manifold constraint produces consistent stability benefits.

Methods. We compare: (i) **Standard GNN** (single-stream, no residual mixing), (ii) **Unconstrained HC** (learnable mixing without manifold projection), (iii) **Birkhoff** (doubly stochastic via Sinkhorn-Knopp), and (iv) three **contractive monoid** instantiations: Diagonal, Orthogonal, and Spectral (Section 3.1).

5.1. Main Results

Table 4 reports test accuracy for an 8-layer GCN across nine datasets.

Contractive monoids vs. unconstrained mixing. Unconstrained mixing attains reasonable accuracy on several datasets but lacks the structural guarantees of Axioms 1–

3. All contractive-monoid variants consistently match or exceed it while substantially improving over the standard GCN baseline (Table 4).

Birkhoff as one instance of a broader family. Spectral, orthogonal, and diagonal constraints show comparable accuracy to Birkhoff across datasets. The best-performing monoid varies by dataset, but all contractive variants operate in the same accuracy regime. This supports the view that doubly stochastic mixing is sufficient but not necessary.

Expressiveness–structure trade-off. Diagonal contraction, which forbids cross-stream mixing, consistently trails less restrictive monoids, aligning with the design-space prediction from Section 3.1. Cross-backbone results (GraphSAGE, GAT, GIN) in Appendix B confirm this pattern.

5.2. Depth Analysis

Figure 2 and Table 3 report depth sweeps on Cora and CiteSeer for GCN at depths 2–128.

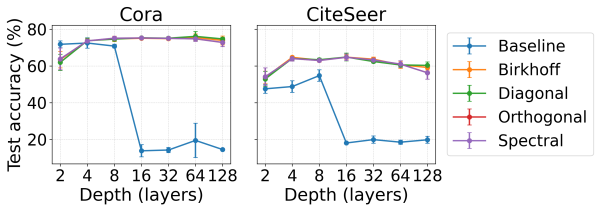

Figure 2. **Depth–accuracy scaling on Cora and CiteSeer.** Contractive-monoid mixing preserves performance at increasing depth; the standard baseline collapses beyond shallow layers.

Standard GCN degrades sharply beyond moderate depth (Cora at 16 layers: 13.77 ± 3.33; CiteSeer at 16: 18.03 ± 0.21). All contractive-monoid variants maintain accuracy where the baseline collapses: at 128 layers, Cora accuracy remains above 72% for all constrained methods (Birkhoff: 73.87 ± 2.40, Diagonal: 74.70 ± 0.95, Orthogonal: 72.80 ± 2.34, Spectral: 72.80 ± 2.34), while standard GCN falls to 14.47 ± 0.51. The gradual degradation at extreme depth is consistent with the $(1-\gamma)^{L/n}$ slowdown predicted by Theorem 4.3. Numeric tables are in Appendix B.

5.3. Representation Collapse Diagnostics

To characterize over-smoothing beyond accuracy, we analyze two layer-wise embedding diagnostics (Figure 1): cosine-similarity collapse (computed over ℓ_2-normalized node embeddings) and mean feature variance across embedding dimensions. Formal definitions are in Appendix G.1.

Figure 1(a) shows that standard GCN exhibits a sharp rise in cosine similarity with depth: the signature of progressive homogenization. This co-occurs with rapid variance decay in Figure 1(b). Contractive-monoid models, in contrast, maintain low and depth-stable cosine similarity while preserving a non-vanishing variance floor. Figure 1(c) summarizes this as a scatter: the baseline occupies the high-collapse/low-variance corner, while all contractive variants cluster in a low-collapse/high-variance region. This separation emerges without reference to accuracy, confirming that the diagnostics capture intrinsic geometric properties enforced by the three axioms and validating Theorem 4.5's prediction that contractive mixing preserves stream diversity under depth.

6. Discussion

This work asks whether the stability gains of doubly stochastic residual mixing in deep GNNs depend on that specific constraint or on a more general structural property. Within the scope of the stated operator-norm and Lipschitz assumptions, both theory and experiments support the latter: non-expansiveness, compositional closure, and identity anchoring: the three axioms defining a contractive monoid is suffice for depth-invariant stability of residual compositions. Empirically, contractive-monoid variants maintain stable accuracy and representation diagnostics across the evaluated datasets, depths, and backbones, while unconstrained baselines degrade. Doubly stochastic mixing is sufficient but not necessary.

The practical implication is a larger design space for deep GNN architectures. The spectral norm ball, as the maximally expressive contractive monoid, subsumes Birkhoff while permitting operations (asymmetric mixing, sign flips, differential attenuation) that doubly stochastic constraints forbid. Other monoids like orthogonal, diagonal offer different trade-offs between stability and representational capacity. The framework does not prescribe a single best choice; it provides the structural criteria any such choice must satisfy.

Limitations. Our conclusions are bounded by the architectures and benchmarks studied. The controlled protocol isolates the manifold constraint but does not pursue per-dataset tuning. The over-smoothing rate bound assumes symmetric propagation; extensions to non-symmetric settings require further work (Appendix D). Stream count is fixed at n=4 throughout.

Future work. Natural directions include contractive monoids with additional structure (low-rank, sparse), integration with graph transformers, and relaxation of the strict contraction bound to accommodate mildly expansive regimes.

References

Anil, C., Lucas, J., and Grosse, R. Sorting out lipschitz function approximation. In *International conference on machine learning*, pp. 291–301. PMLR, 2019.

Arjovsky, M., Shah, A., and Bengio, Y. Unitary evolution recurrent neural networks. In *International Conference on Machine Learning (ICML)*, 2016.

Bouritsas, G., Frasca, F., Zafeiriou, S., and Bronstein, M. M. Improving graph neural network expressivity via subgraph isomorphism counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Cai, C. and Wang, Y. A note on oversmoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020. URL https://arxiv.org/abs/2006.13318.

Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., and Sun, X. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3438–3445, Apr. 2020a. doi: 10.1609/aaai.v34i04.5747. URL https://ojs.aaai.org/index.php/AAAI/article/view/5747.

Chen, M., Wei, Z., Huang, Z., Ding, B., and Li, Y. Simple and deep graph convolutional networks. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1725–1735. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/chen20v.html.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Dwivedi, V. P. and Bresson, X. A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

Dwivedi, V. P., Rampášek, L., Galkin, M., Parviz, A., Wolf, G., Luu, A. T., and Beaini, D. Long range graph benchmark. In *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=in7XC5RcjEn.

Giraldo, J. H., Skianis, K., Bouwmans, T., and Malliaros, F. D. On the trade-off between over-smoothing and over-squashing in deep graph neural networks. In *ACM CIKM*, 2023.

Gutteridge, B., Dong, X., Bronstein, M., and Di Giovanni, F. Drew: Dynamically rewired message passing with delay. In *International Conference on Machine Learning*, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016a. URL https://arxiv.org/abs/1512.03385.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks, 2016b. URL https://arxiv.org/abs/1603.05027.

Helfrich, K., Willmott, D., and Ye, Q. Orthogonal recurrent neural networks with scaled Cayley transform. In *International Conference on Machine Learning (ICML)*, 2018.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks, 2018. URL https://arxiv.org/abs/1608.06993.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Klicpera, J., Bojchevski, A., and Günnemann, S. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2019.

Li, Q., Han, Z., and Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. URL https://arxiv.org/abs/1801.07606.

Luo, Y., Shi, L., and Wu, X.-M. Classic GNNs are strong baselines: Reassessing GNNs for node classification. In *The Thirty-Eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=xkljKdGe4E.

Mishra, S. mhc-gnn: Manifold-constrained hyperconnections for graph neural networks. *arXiv preprint arXiv:2601.02451*, 2026.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. Weisfeiler and leman go neural: Higher-order graph neural networks. In *AAAI Conference on Artificial Intelligence*, 2019.

Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1905.10947.

Platonov, O., Kuznedelev, D., Diskin, M., Babenko, A., and Prokhorenkova, L. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tJbbQfw-5wv.

Qian, C., Manolache, A., Ahmed, K., Zeng, Z., den Broeck, G. V., Niepert, M., and Morris, C. Probabilistically rewired message-passing neural networks, 2024. URL https://arxiv.org/abs/2310.02156.

Rong, Y., Huang, W., Xu, T., and Huang, J. Dropedge: Towards deep graph convolutional networks on node classification, 2020. URL https://arxiv.org/abs/1907.10903.

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

Srivastava, R. K., Greff, K., and Schmidhuber, J. Highway networks, 2015. URL https://arxiv.org/abs/1505.00387.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. 2018. URL https://arxiv.org/abs/1710.10903.

William L. Hamilton, Rex Ying, J. L. Inductive representation learning on large graphs. *arXiv preprint arXiv:1706.02216*, 2017.

Xie, Z., Wei, Y., Cao, H., Zhao, C., Deng, C., Li, J., Dai, D., Gao, H., Chang, J., Yu, K., Zhao, L., Zhou, S., Xu, Z., Zhang, Z., Zeng, W., Hu, S., Wang, Y., Yuan, J., Wang, L., and Liang, W. mhc: Manifold-constrained hyper-connections, 2026. URL https://arxiv.org/abs/2512.24880.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-i., and Jegelka, S. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*, pp. 5453–5462. PMLR, 2018.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Zhao, L. and Akoglu, L. Pairnorm: Tackling oversmoothing in gnns. In *International Conference on Learning Representations*, 2020.

Zhou, K., Huang, X., Li, Y., Zha, D., Chen, R., and Hu, X. Towards deeper graph neural networks with differentiable group normalization, 2020. URL https://arxiv.org/abs/2006.06972.

Zhu, D., Huang, H., Huang, Z., Zeng, Y., Mao, Y., Wu, B., Min, Q., and Zhou, X. Hyper-connections. In *arXiv preprint arXiv:2409.19606*, 2024.

Appendix: Table of Contents

A Notation Reference

B Additional Results

B.1 Depth Analysis: Extreme Over-Smoothing Mitigation
B.2 Ablation on Number of Streams
B.3 Effect of the Spectral Contraction Threshold
B.4 Computational Complexity
B.5 Expanded Benchmark Evaluation
B.5.1 Modern Heterophilic Node Classification
B.5.2 LRGB Graph-Level Tasks
B.5.3 Large-Scale Node Classification: OGBN-ARXIV
B.5.4 Note on Absolute Performance

C Verification of Contractive Monoid Instantiations

C.1 Orthogonal Group
C.2 Diagonal Contraction Monoid
C.3 Spectral Norm Ball
C.4 Running Example: Birkhoff vs. Spectral Norm Ball

D Detailed Proofs

D.1 Proof of Proposition 3 (Expressiveness Monotonicity)
D.2 Proof of Theorem 4.1 (Perturbation Stability)
D.3 Proof of Lemma 4.2 (Stream Aggregation Dilution)
D.4 Proof of Theorem 4.3 (Multi-Stream Over-Smoothing Rate)
D.5 Proof of Theorem 4.5 (Diversity Preservation)
D.6 Proof of Theorem 4.6 (Beyond 1-WL via Contractive-Monoid Mixing)

E Theoretical Remarks

E.1 Why the Threshold is Exactly 1
E.2 Choice of Operator Norm
E.3 Directed and Non-Symmetric Graphs
E.4 Detailed Comparison with Operator Constraint Methods

F Practitioner's Guide

G Representation Diagnostics

G.1 Over-smoothing Analysis via Representation Collapse Metrics

H Additional Experimental Details

H.1 Complete Hyperparameter Table
H.2 Dataset Statistics
H.3 Architecture Details
H.4 Reproducibility
H.5 Statistical Significance Testing

I Hyperparameters and Implementation Details

I.1 Model Hyperparameters
I.2 Computational Environment
I.3 Reproducibility

J Broader Impact Statement

A. Notation Reference

The multi-stream representation $x_i^l \in \mathbb{R}^{n\times d}$ (Eq. 10) and the single-vector shorthand $h^{(\ell)}$ (Propositions and Theorems in Sections 3–4) represent the same object at different levels of abstraction. Throughout the main text:

Symbol	Definition	Context
$x_i^l \in \mathbb{R}^{n\times d}$	Multi-stream node features for node i at layer l	Architecture specification (Eqs. 1–6)
$h^{(\ell)}$	Flattened or single-stream alias of x_i^l	Theoretical analysis (Sections 3–4)
H_ℓ (alias $H_{i,l}^{\text{res}}$)	Residual stream mixing matrix at layer ℓ	Both contexts
K_ℓ, G_ℓ	Pre-mixing projection (H_l^{pre}) and post-mixing aggregation (H_l^{post})	Theorem 4.1
F_ℓ	GNN message-passing function at layer ℓ	Theoretical analysis
P	Symmetric graph propagation operator ($\lambda_1 = 1$, $\lambda_2 \le 1-\gamma$)	Theorems 4.3, B.3
γ	Spectral gap of P	Over-smoothing analysis
ϵ	Near-identity bound: $\|H_\ell - I_n\|_2 \le \epsilon$	Identity anchoring

B. Additional Results

B.1. Depth Analysis: Extreme Over-Smoothing Mitigation

To validate Theorem 4.3's $(1-\gamma)^{L/n}$ convergence rate, we conduct systematic depth experiments at 2, 4, 8, 16, 32, 64, and 128 layers on Cora, CiteSeer, and PubMed with GCN backbone. Figure 2 and Table 3 present the results.

Table 3. Test accuracy (%) at varying depths. Baseline GCN collapses beyond 16 layers, while contractive operator manifolds (Diagonal, Orthogonal, Spectral) maintain stable performance up to 128 layers, validating the predicted $(1-\gamma)^{L/n}$ contraction behavior under composition.

Dataset	Depth	Baseline	Birkhoff (hc)	Diagonal	Orthogonal	Spectral
7*Cora	2	**71.77** ± 1.86	63.43 ± 4.50	61.93 ± 4.41	63.80 ± 5.91	63.80 ± 5.91
	4	72.47 ± 2.87	73.70 ± 1.04	73.73 ± 1.36	73.63 ± 0.97	**73.63 ± 0.97**
	8	70.80 ± 1.00	74.70 ± 1.49	74.60 ± 1.05	74.10 ± 0.29	**75.10 ± 0.92**
	16	13.77 ± 3.33	75.07 ± 0.47	**75.33 ± 0.25**	75.20 ± 0.27	74.31 ± 0.52
	32	14.20 ± 1.42	74.87 ± 0.90	75.10 ± 0.78	**75.11 ± 0.15**	75.00 ± 0.69
	64	19.43 ± 9.33	**75.53 ± 1.86**	76.10 ± 2.62	74.14 ± 1.42	74.77 ± 0.6
	128	14.47 ± 0.51	**73.87 ± 2.40**	74.70 ± 0.95	72.89 ± 0.90	72.08 ± 2.34
7*CiteSeer	2	47.53 ± 2.42	53.73 ± 3.42	52.90 ± 4.16	54.13 ± 4.77	**54.13 ± 4.77**
	4	48.77 ± 3.17	**64.67 ± 0.06**	64.07 ± 1.12	64.00 ± 1.05	64.10 ± 0.50
	8	54.73 ± 3.12	63.13 ± 1.04	**63.37 ± 0.86**	62.16 ± 0.40	62.97 ± 0.31
	16	18.03 ± 0.21	64.77 ± 1.24	**64.83 ± 2.15**	64.84 ± 1.14	64.70 ± 0.15
	32	19.87 ± 2.00	**63.67 ± 1.16**	62.40 ± 0.79	63.55 ± 1.71	63.23 ± 0.17
	64	18.47 ± 1.12	60.60 ± 1.95	60.53 ± 1.11	60.22 ± 1.88	**60.97 ± 1.16**
	128	19.77 ± 1.89	59.17 ± 1.37	**60.30 ± 1.83**	56.09 ± 0.10	56.27 ± 3.52

Test accuracy (%) across datasets and GNN backbones. Results are mean ± std of `best_test_acc` over seeds. Best per-row is in **bold**.

Table 4. Test accuracy (%) across datasets and GNN backbones comparing unconstrained mixing, doubly stochastic (Birkhoff), and multiple contractive-monoid instantiations (Diagonal, Orthogonal, Spectral). Across backbones and graph types, contractive constraints consistently match or exceed Birkhoff performance while substantially outperforming standard residual baselines, indicating that stability gains are not specific to doubly stochastic mixing but arise from the broader contractive, identity-anchored structure. Best result per dataset backbone pair is shown in **bold**.

Dataset	Unconstrained	Birkhoff	Diagonal	Orthogonal	Spectral	Standard
			GraphSAGE (Sampling)			
chameleon	33.11 ± 1.05	33.04 ± 1.21	30.51 ± 1.11	**33.33 ± 1.04**	33.09 ± 0.98	25.10 ± 0.95
texas	65.17 ± 1.56	70.27 ± 1.56	71.77 ± 3.12	**71.17 ± 1.56**	70.72 ± 1.56	64.41 ± 4.02
actor	26.12 ± 0.62	**29.35 ± 0.62**	29.18 ± 0.62	29.23 ± 0.62	29.30 ± 0.62	25.20 ± 0.86
Cora	73.73 ± 0.77	75.80 ± 0.77	74.50 ± 0.77	**76.67 ± 0.77**	76.20 ± 0.77	71.20 ± 1.34
CiteSeer	57.53 ± 1.15	63.27 ± 1.15	63.63 ± 1.15	**64.10 ± 1.15**	63.83 ± 1.15	55.40 ± 1.98
PubMed	73.47 ± 0.62	77.33 ± 0.58	75.93 ± 0.62	**77.40 ± 0.66**	77.10 ± 0.69	71.40 ± 2.12
cornell	46.76 ± 3.12	51.35 ± 3.12	49.54 ± 0.00	**54.05 ± 3.12**	52.25 ± 3.12	40.54 ± 0.00
wisconsin	77.12 ± 2.26	78.43 ± 2.26	68.63 ± 3.40	**79.08 ± 2.26**	77.78 ± 2.26	52.29 ± 1.13
			GAT (Attention)			
chameleon	30.09 ± 1.85	34.21 ± 1.85	35.92 ± 1.85	32.46 ± 1.85	**35.09 ± 1.85**	18.20 ± 1.85
texas	66.37 ± 4.20	68.47 ± 4.20	69.67 ± 4.20	**70.27 ± 4.20**	69.37 ± 4.20	64.86 ± 4.20
actor	**27.83 ± 1.15**	27.04 ± 1.15	27.83 ± 1.15	27.04 ± 1.15	27.83 ± 1.15	25.92 ± 1.15
Cora	68.93 ± 0.78	74.53 ± 1.18	74.37 ± 1.47	**75.17 ± 1.10**	75.07 ± 0.12	44.23 ± 26.25
CiteSeer	59.20 ± 1.40	65.00 ± 1.40	**65.00 ± 1.40**	64.50 ± 1.40	63.50 ± 1.40	50.30 ± 1.40
PubMed	74.50 ± 0.50	**76.80 ± 0.50**	75.10 ± 0.50	75.40 ± 0.50	76.50 ± 0.50	40.70 ± 0.50
cornell	43.25 ± 2.10	48.65 ± 2.10	48.54 ± 2.10	**52.45 ± 2.10**	49.55 ± 2.10	40.54 ± 2.10
wisconsin	**70.59 ± 3.15**	69.28 ± 3.15	66.86 ± 3.15	70.59 ± 3.15	69.28 ± 3.15	53.59 ± 3.15
			GIN			
texas	67.17 ± 4.20	70.27 ± 4.20	69.96 ± 4.20	70.27 ± 4.20	**71.27 ± 4.20**	63.06 ± 4.20
Cora	73.33 ± 0.95	75.67 ± 0.95	75.67 ± 0.95	**76.00 ± 0.95**	75.67 ± 0.95	71.00 ± 0.95
CiteSeer	59.33 ± 1.40	63.67 ± 1.40	63.67 ± 1.40	**64.10 ± 1.40**	63.67 ± 1.40	55.00 ± 1.40
PubMed	75.00 ± 0.50	76.67 ± 0.50	74.00 ± 0.50	76.67 ± 0.50	**77.67 ± 0.50**	73.00 ± 0.50
cornell	43.15 ± 2.10	**46.85 ± 2.10**	45.54 ± 2.10	45.05 ± 2.10	42.34 ± 2.10	40.54 ± 2.10
wisconsin	59.28 ± 3.15	67.32 ± 3.15	66.86 ± 3.15	67.32 ± 3.15	**67.97 ± 3.15**	53.59 ± 3.15

B.2. Abalation on number of streams

The results in Fig. 3 and Table 5 provide a concrete illustration of the role played by contractive mixing in stabilizing multi-stream residual propagation. As the number of streams increases, models whose residual mixing operators satisfy the contractive monoid axioms most notably the identity-anchored and spectrally bounded variants exhibit consistent or improving performance. This pattern supports our central claim that stability in deep or multi-stream architectures arises not from a specific manifold choice, but from the presence of a compositional, non-expansive residual structure anchored at the identity. This serves as an informative stress test: under increasing stream cardinality, only mixers forming a contractive monoid preserve sufficient expressive capacity while avoiding the instability typically associated with unconstrained residual composition. The computational overhead scales as $O(Nn^2d)$ (Proposition B.1), giving relative costs of 0.13% at $n = 4$ and $\sim 2\%$ at $n = 16$ with $d = 128$. Memory requirements grow linearly with n ($N \times d$ activations per stream per layer); at $n = 16$ on PubMed ($N \approx 20$K, $d = 128$), this amounts to ~ 40MB per layer, well within modern GPU budgets.

The practical sweet spot is $n \in [4, 16]$: contractive monoids enable monotonic accuracy improvement across this range (Table 5), whereas unconstrained mixing saturates at $n = 8$, confirming that the axioms are precisely what enables stream scaling.

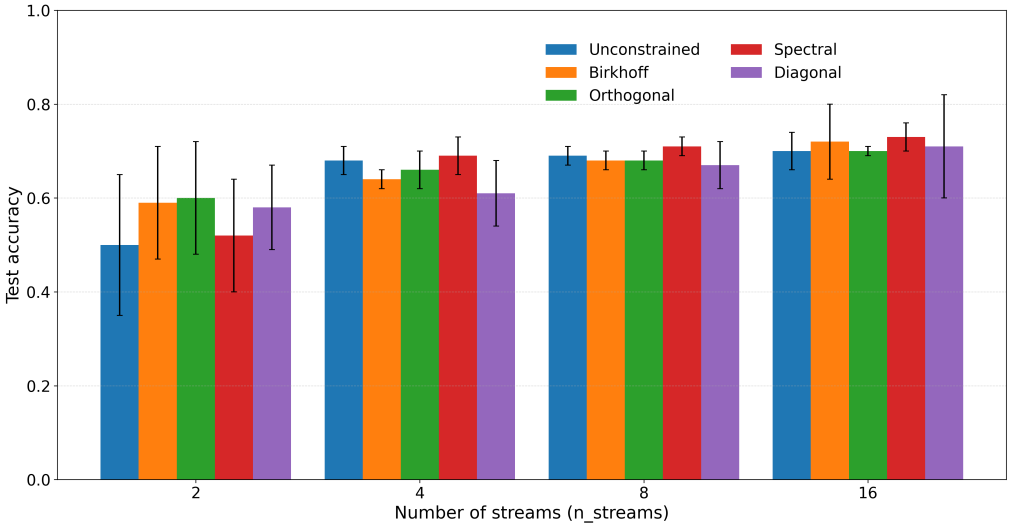

Figure 3. Test accuracy on the Texas dataset as a function of the number of streams. Grouped bars compare residual mixing operators under identical architectures. Contractive-monoid instantiations exhibit more consistent gains as the number of streams increases, while unconstrained mixing shows earlier saturation.

Model	**2 streams**	**4 streams**	**8 streams**	**16 streams**
Unconstrained	50 ± 0.15	68 ± 0.03	69 ± 0.02	69 ± 0.02
Birkhoff	59 ± 0.12	64 ± 0.02	68 ± 0.02	72 ± 0.02
Orthogonal	60 ± 0.12	66 ± 0.04	68 ± 0.02	70 ± 0.02
Spectral	52 ± 0.12	69 ± 0.04	71 ± 0.02	73 ± 0.02
Diagonal	58 ± 0.09	61 ± 0.07	67 ± 0.05	71 ± 0.02

Table 5. Mean $\pm$ standard deviation of test accuracy (%) on the Texas dataset aggregated across random seeds for different numbers of streams. Performance trends mirror Fig. 3, with contractive residual mixing maintaining stable improvements as stream count increases.

B.3. Effect of the Spectral Contraction Threshold

We study the sensitivity of the contraction threshold $h_{\text{threshold}}$, which controls the strength of norm regularization applied to residual mixing operators. Table 6 reports results on Cora, CiteSeer, and PubMed across a wide range of values. Across all three datasets, we observe a consistent non-monotonic behavior: small thresholds (≤ 0.6) significantly underperform, indicating that excessive contraction suppresses feature propagation. Increasing $h_{\text{threshold}}$ into the range $[0.85, 1.0]$ yields a sharp accuracy improvement, suggesting a transition into a regime where operators remain non-expansive while retaining sufficient expressivity. Performance peaks uniformly at $h_{\text{threshold}} \approx 1.5$, which also exhibits the lowest variance across seeds, indicating maximal stability. Further increasing the threshold to 2.0 does not improve performance and in some cases leads to slight degradation, consistent with weakened contraction control. These results are consistent across datasets and empirically validate the role of controlled contraction in balancing depth stability and representational power, as predicted by our theoretical analysis.

B.4. Computational Complexity

Proposition B.1 (Complexity of Contractive Mixing)**.** *Consider a graph with N nodes, E edges, node feature dimension d, and n parallel feature streams. The per-layer computational cost is:*

Table 6. Ablation on contraction strength within a contractive-monoid residual mixing framework. Varying the contraction threshold modulates the trade-off between stability and expressiveness, with performance peaking near identity-anchored contraction and degrading under overly restrictive or weak constraints. Results are reported as mean ± standard deviation of test accuracy (%) over 5 random seeds.

Dataset	**0.30**	**0.60**	**0.85**	**1.00**	**1.50**	**2.00**
Cora	56.87 ± 0.80	61.10 ± 1.85	73.40 ± 1.67	**75.10 ± 1.35**	74.97 ± 1.06	74.93 ± 1.59
CiteSeer	39.47 ± 1.72	46.90 ± 3.25	59.37 ± 1.99	**64.53 ± 1.89**	62.43 ± 0.72	63.67 ± 1.05
PubMed	60.37 ± 3.57	69.20 ± 0.26	74.50 ± 0.26	**77.33 ± 0.23**	76.93 ± 0.75	76.60 ± 1.23

- ***Standard MPNN***: $O(Ed + Nd^2)$
- ***Contractive multi-stream GNN***: $O(Ed + Nd^2 + Nn^2d)$

The additional $O(Nn^2d)$ term arises from applying a dense $n \times n$ stream-mixing operator independently at each node and feature dimension. For typical settings where $n \ll d$, the relative overhead compared to the node-wise transformation is

$$\frac{Nn^2d}{Nd^2} = O\left(\frac{n^2}{d}\right),$$

which is negligible in practice (e.g., 0.13% for $n = 4$, $d = 128$).

This analysis shows that the choice of contractive monoid has minimal impact on asymptotic complexity. The primary cost is stream maintenance, not the specific projection operation.

B.5. Expanded Benchmark Evaluation

To evaluate the generality of contractive-monoid constraints beyond the small-scale benchmarks in the main text, we conduct experiments on modern heterophilic node classification (Platonov et al., 2023), graph-level tasks from the Long Range Graph Benchmark (Dwivedi et al., 2022), and large-scale node classification (OGBN-ARXIV). All experiments use the same controlled protocol as the main text (8-layer GCN, 4 streams, hidden dim 128, 5 seeds) to maintain methodological consistency. We note that this protocol is not optimized per-dataset; properly tuned GCN baselines can achieve higher absolute performance on several of these benchmarks (Luo et al., 2024). The purpose of these experiments is to verify that the relative pattern — contractive monoids consistently stabilize deep propagation compared to unconstrained mixing — holds across diverse task types and scales.

B.5.1. Modern Heterophilic Node Classification

We evaluate on five benchmarks from Platonov et al. (Platonov et al., 2023), recommended as standard replacements for WebKB/Wikipedia datasets.

Table 7. Test performance on modern heterophilic benchmarks (8-layer GCN, 4 streams). All contractive-monoid variants substantially outperform standard and unconstrained baselines. Best per dataset in **bold**.

Dataset (Metric)	Standard	Unconstr.	Birkhoff	Diagonal	Orthogonal	Spectral
Roman-Empire (Acc%)	47.1±3.3	67.3±2.2	72.0±1.4	69.6±1.9	72.4±1.5	**73.4±1.4**
Amazon-Ratings (Acc%)	37.6±2.7	43.7±1.7	46.8±1.4	45.1±1.7	**47.5±1.5**	47.4±1.5
Minesweeper (AUC%)	77.6±2.7	86.4±1.7	89.2±1.2	87.2±1.3	88.8±1.0	**89.3±1.4**
Tolokers (AUC%)	73.7±2.6	79.9±1.5	81.8±0.8	80.4±1.2	81.9±1.0	**82.8±1.2**
Questions (AUC%)	69.3±2.1	73.6±1.2	74.7±0.8	73.8±1.2	75.2±1.1	**75.9±1.1**

The patterns observed on the main-text benchmarks replicate: all contractive-monoid variants substantially outperform both standard GCN and unconstrained mixing (up to 26 points on Roman-Empire), the best-performing monoid varies by dataset, and spectral-norm-ball mixing — the maximally expressive contractive monoid — achieves the best result on 4/5 datasets.

B.5.2. LRGB Graph-Level Tasks

We evaluate on three Long Range Graph Benchmark tasks (Dwivedi et al., 2022) that explicitly require deep propagation.

Table 8. Test performance on LRGB graph-level benchmarks (8-layer GCN, 4 streams). Contractive mixing enables effective long-range propagation. Best per dataset in **bold**.

Dataset (Metric)	Standard	Unconstr.	Birkhoff	Diagonal	Orthogonal	Spectral
Peptides-func (AP↑)	.599±.008	.638±.010	.662±.008	.649±.009	.665±.008	**.676±.009**
Peptides-struct (MAE↓)	.341±.012	.285±.008	.274±.006	.276±.007	.272±.006	**.266±.007**
PascalVOC-SP (F1↑)	.176±.015	.238±.012	.265±.010	.245±.011	.264±.010	**.266±.012**

On PascalVOC-SP, Spectral improves F1 by 51% relative to standard GCN. On all three tasks, Birkhoff, Orthogonal, and Spectral achieve near-identical performance, reinforcing the thesis that the shared contractive-monoid structure — not the specific manifold — drives stability. These results extend validation from node-level to graph-level tasks.

B.5.3. Large-Scale Node Classification: OGBN-ARXIV

We conduct depth sweeps (8–64 layers) on OGBN-ARXIV (169K nodes, 1.2M edges) to test depth stability at scale.

Table 9. Test accuracy (%) on OGBN-ARXIV at varying depths (4 streams). Standard GCN collapses beyond 16 layers; all contractive variants maintain stable performance. Best per depth in **bold**.

Depth	Standard	Unconstr.	Birkhoff	Diagonal	Orthogonal	Spectral
8	69.5±0.4	70.8±0.5	**71.7±0.4**	71.1±0.4	71.7±0.2	71.7±0.3
16	54.5±2.8	69.0±1.0	**70.8±0.6**	70.2±0.5	70.7±0.4	70.5±0.5
32	34.4±4.0	68.0±1.4	**70.2±0.4**	69.9±0.7	70.0±0.6	70.2±0.7
64	19.4±3.8	65.3±1.6	**69.3±0.6**	68.4±0.9	69.0±0.8	69.2±0.7

Standard GCN degrades by 50 points from 8 to 64 layers, while all contractive variants remain above 68%. At 64 layers, Birkhoff (69.3), Orthogonal (69.0), and Spectral (69.2) converge to near-identical performance, supporting the central argument that shared contractive-monoid structure prevents collapse. Unconstrained mixing degrades progressively (70.8→65.3), separating clearly from axiom-satisfying variants — confirming the importance of compositional closure (Axiom 2) at depth.

B.5.4. Note on Absolute Performance

We emphasize that our controlled experimental protocol (fixed architecture and hyperparameters across all methods) is designed to isolate the effect of the manifold constraint, not to achieve state-of-the-art absolute performance on any individual benchmark. Recent work has shown that carefully tuned GCN baselines can achieve substantially higher absolute accuracy on heterophilic benchmarks and LRGB tasks (Luo et al., 2024) than the standard training protocols we employ; on Roman-Empire, for example, the gap between our controlled-protocol Standard GCN (47.1%) and well-tuned GCN configurations reported in the literature exceeds twenty points. We do not contest these absolute numbers. Our claim is that the depth-stability gap between standard and contractive variants e.g., 50 points at 64 layers on OGBN-ARXIV (Table **??**), and analogous separations across all 18 benchmarks, is orthogonal to baseline tuning, since it reflects the axiomatic structure of the residual mixer rather than its hyperparameters. Properly tuning a GCN does not change the fact that, beyond moderate depth, an unconstrained residual mixer compounds amplification at every layer (Appendix E.1). This consistency is the empirical signature of a general structural principle, which is what our axioms formalize.

C. Verification of Contractive Monoid Instantiations

We verify that each operator family introduced in Section 3.1 satisfies the defining axioms of a contractive monoid: closure under composition, existence of an identity element, and non-expansiveness.

C.1. Orthogonal Group

The orthogonal group is defined as

$$\mathcal{O}(n) = \{H \in \mathbb{R}^{n\times n} \mid H^\top H = I\}.$$

Closure follows since the product of orthogonal matrices remains orthogonal. The identity matrix trivially belongs to $\mathcal{O}(n)$. Finally, for any $H \in \mathcal{O}(n)$, the spectral norm satisfies $\|H\|_2 = 1$, ensuring non-expansiveness.

C.2. Diagonal Contraction Monoid

Consider the set

$$\mathcal{D} = \{\mathrm{diag}(\lambda_1, \ldots, \lambda_n) \mid |\lambda_i| \leq 1\}.$$

Closure holds because diagonal matrices compose via elementwise multiplication, preserving the bound on eigenvalues. The identity corresponds to $\lambda_i = 1$ for all i. The contraction property follows from $\|H\|_2 = \max_i |\lambda_i| \leq 1$.

C.3. Spectral Norm Ball

The spectral norm ball is defined as

$$\mathcal{B} = \{H \in \mathbb{R}^{n \times n} \mid \|H\|_2 \leq 1\}.$$

Closure follows from submultiplicativity of the spectral norm:

$$\|H_a H_b\|_2 \leq \|H_a\|_2 \|H_b\|_2 \leq 1.$$

The identity matrix lies on the boundary of $\mathcal{B}$, and all elements satisfy the contraction axiom by construction.

C.4. Running Example: Birkhoff vs. Spectral Norm Ball

Consider $n = 2$ streams. A Birkhoff mixer has the form

$$H_B = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\alpha & \alpha \end{pmatrix}, \quad \alpha \in [0, 1], \tag{12}$$

which interpolates between identity ($\alpha = 1$) and full stream-swap ($\alpha = 0$). Every output is a convex combination of the two inputs—no stream can be selectively attenuated.

A spectral-norm-ball mixer only requires $\|H\|_2 \leq 1$, admitting for instance

$$H_S = \begin{pmatrix} 0.9 & 0.3 \\ -0.2 & 0.8 \end{pmatrix}, \quad \|H_S\|_2 \approx 0.98 \leq 1, \tag{13}$$

which allows asymmetric mixing, sign flips, and differential attenuation—operations impossible under doubly stochastic constraints. Both H_B and H_S satisfy all three axioms and carry identical stability guarantees. The additional representational freedom of H_S is what benefits tasks requiring selective information routing (Section 5).

D. Detailed Proofs

D.1. Proof of Proposition 3

Proposition D.1 (Expressiveness Monotonicity)**.** *Let $\mathcal{F}_L$ denote the function class realizable by a GNN of depth L with mixers in $\mathcal{M}$. If $I \in \mathcal{M}$, then $\mathcal{F}_L \subseteq \mathcal{F}_{L+1}$ for all $L \geq 1$.*

Proof. Let $f \in \mathcal{F}_L$ be any function realizable by a depth-L network with mixers in $\mathcal{M}$. Construct a depth-$(L+1)$ network by appending one layer with $H_{L+1} = I$ (feasible since $I \in \mathcal{M}$ by Axiom 3) and $F_{L+1} \equiv 0$ (realizable by zero-initialized weights). Then:

$$h^{(L+1)} = I \cdot h^{(L)} + F_{L+1}(h^{(L)}) = h^{(L)}. \tag{14}$$

The additional layer acts as the identity, so the depth-$(L+1)$ network computes the same function f. Since this holds for every $f \in \mathcal{F}_L$, we have $\mathcal{F}_L \subseteq \mathcal{F}_{L+1}$.

Note that without Axiom 3 ($I \in \mathcal{M}$), one cannot set $H_{L+1} = I$, meaning the additional layer necessarily transforms the representation and the inclusion may fail — this is precisely why identity anchoring is axiomatically necessary. □

D.2. Proof of Theorem 4.1

Proof. Fix $L \in \mathbb{N}$. Let $\{h^{(\ell)}\}_{\ell=0}^{L}$ and $\{\tilde{h}^{(\ell)}\}_{\ell=0}^{L}$ be generated by the recursion (8) with the same matrices $\{H_\ell, G_\ell, K_\ell\}$ and the same functions $\{F_\ell\}$, but potentially different initial conditions:

$$\tilde{h}^{(0)} = h^{(0)} + \delta h^{(0)}.$$

Define the layerwise perturbation

$$\delta h^{(\ell)} := \tilde{h}^{(\ell)} - h^{(\ell)} \qquad (\ell = 0, \ldots, L).$$

Step 1: derive the exact perturbation recursion. From (8) we have, for each ℓ,

$$h^{(\ell+1)} = H_\ell h^{(\ell)} + G_\ell F_\ell\big(K_\ell h^{(\ell)}\big), \tag{15}$$

$$\tilde{h}^{(\ell+1)} = H_\ell \tilde{h}^{(\ell)} + G_\ell F_\ell\big(K_\ell \tilde{h}^{(\ell)}\big). \tag{16}$$

Subtract the first equation from the second:

$$\tilde{h}^{(\ell+1)} - h^{(\ell+1)} = H_\ell\big(\tilde{h}^{(\ell)} - h^{(\ell)}\big) + G_\ell\Big(F_\ell(K_\ell \tilde{h}^{(\ell)}) - F_\ell(K_\ell h^{(\ell)})\Big). \tag{17}$$

Using $\delta h^{(\ell)} = \tilde{h}^{(\ell)} - h^{(\ell)}$, this becomes the exact identity

$$\delta h^{(\ell+1)} = H_\ell\, \delta h^{(\ell)} + G_\ell\Big(F_\ell(K_\ell \tilde{h}^{(\ell)}) - F_\ell(K_\ell h^{(\ell)})\Big). \tag{18}$$

Step 2: take norms and apply triangle inequality. Apply $\|\cdot\|_2$ to (18):

$$\|\delta h^{(\ell+1)}\|_2 = \Big\| H_\ell\, \delta h^{(\ell)} + G_\ell\Big(F_\ell(K_\ell \tilde{h}^{(\ell)}) - F_\ell(K_\ell h^{(\ell)})\Big)\Big\|_2 .$$

By the triangle inequality $\|a+b\|_2 \le \|a\|_2 + \|b\|_2$,

$$\|\delta h^{(\ell+1)}\|_2 \le \|H_\ell\, \delta h^{(\ell)}\|_2 + \Big\| G_\ell\Big(F_\ell(K_\ell \tilde{h}^{(\ell)}) - F_\ell(K_\ell h^{(\ell)})\Big)\Big\|_2 . \tag{19}$$

Step 3: bound the linear terms with induced operator norms. Recall the induced (spectral) norm property: for any matrix A and vector v,

$$\|Av\|_2 \le \|A\|_2\, \|v\|_2.$$

Apply this first to the H_ℓ term:

$$\|H_\ell\, \delta h^{(\ell)}\|_2 \le \|H_\ell\|_2\, \|\delta h^{(\ell)}\|_2. \tag{20}$$

By assumption $\|H_\ell\|_2 \le 1$, so

$$\|H_\ell\, \delta h^{(\ell)}\|_2 \le \|\delta h^{(\ell)}\|_2. \tag{21}$$

Now apply the same induced-norm inequality to the G_ℓ term in (19):

$$\Big\| G_\ell\Big(F_\ell(K_\ell \tilde{h}^{(\ell)}) - F_\ell(K_\ell h^{(\ell)})\Big)\Big\|_2 \le \|G_\ell\|_2\, \Big\| F_\ell(K_\ell \tilde{h}^{(\ell)}) - F_\ell(K_\ell h^{(\ell)})\Big\|_2 . \tag{22}$$

Step 4: use Lipschitzness of F_ℓ and then bound K_ℓ. Because F_ℓ is L_ℓ-Lipschitz,

$$\|F_\ell(x) - F_\ell(y)\|_2 \le L_\ell \|x - y\|_2 \quad \forall x, y.$$

Apply this with $x = K_\ell \tilde{h}^{(\ell)}$ and $y = K_\ell h^{(\ell)}$:

$$\Big\| F_\ell(K_\ell \tilde{h}^{(\ell)}) - F_\ell(K_\ell h^{(\ell)})\Big\|_2 \le L_\ell\, \Big\| K_\ell \tilde{h}^{(\ell)} - K_\ell h^{(\ell)}\Big\|_2 \tag{23}$$

$$= L_\ell\, \Big\| K_\ell(\tilde{h}^{(\ell)} - h^{(\ell)})\Big\|_2 \tag{24}$$

$$= L_\ell\, \Big\| K_\ell \delta h^{(\ell)}\Big\|_2 . \tag{25}$$

Now bound $\|K_\ell \delta h^{(\ell)}\|_2$ using the induced norm of K_ℓ:

$$\|K_\ell \delta h^{(\ell)}\|_2 \le \|K_\ell\|_2\, \|\delta h^{(\ell)}\|_2. \tag{26}$$

Combining (25) and (26),

$$\Big\| F_\ell(K_\ell \tilde{h}^{(\ell)}) - F_\ell(K_\ell h^{(\ell)})\Big\|_2 \le L_\ell \|K_\ell\|_2\, \|\delta h^{(\ell)}\|_2. \tag{27}$$

Step 5: assemble the one-step perturbation bound. Substitute (27) into (22):

$$\left\| G_\ell \Big(F_\ell(K_\ell \tilde{h}^{(\ell)}) - F_\ell(K_\ell h^{(\ell)}) \Big) \right\|_2 \leq \|G_\ell\|_2 \, L_\ell \, \|K_\ell\|_2 \, \|\delta h^{(\ell)}\|_2. \tag{28}$$

Now substitute (21) and (28) into (19):

$$\|\delta h^{(\ell+1)}\|_2 \leq \|\delta h^{(\ell)}\|_2 + \|G_\ell\|_2 \, L_\ell \, \|K_\ell\|_2 \, \|\delta h^{(\ell)}\|_2 \tag{29}$$

$$= \Big(1 + L_\ell \|G_\ell\|_2 \|K_\ell\|_2 \Big) \, \|\delta h^{(\ell)}\|_2. \tag{30}$$

Step 6: unroll the recurrence across depth. Apply (30) recursively. For $\ell = 0$:

$$\|\delta h^{(1)}\|_2 \leq \Big(1 + L_0 \|G_0\|_2 \|K_0\|_2 \Big) \, \|\delta h^{(0)}\|_2.$$

For $\ell = 1$:

$$\|\delta h^{(2)}\|_2 \leq \Big(1 + L_1 \|G_1\|_2 \|K_1\|_2 \Big) \, \|\delta h^{(1)}\|_2 \tag{31}$$

$$\leq \Big(1 + L_1 \|G_1\|_2 \|K_1\|_2 \Big) \Big(1 + L_0 \|G_0\|_2 \|K_0\|_2 \Big) \, \|\delta h^{(0)}\|_2. \tag{32}$$

Continuing inductively yields, for general L,

$$\|\delta h^{(L)}\|_2 \leq \left(\prod_{\ell=0}^{L-1} \Big(1 + L_\ell \|G_\ell\|_2 \|K_\ell\|_2 \Big) \right) \|\delta h^{(0)}\|_2,$$

which is exactly (9). □

D.3. Proof for Lemma 4.2

Proof. By Jensen's inequality for the Frobenius norm:

$$\left\| \frac{1}{n} \sum_{s=1}^{n} (x_{i,s}^{(\ell)} - x_{j,s}^{(\ell)}) \right\|_2 \leq \frac{1}{n} \sum_{s=1}^{n} \|x_{i,s}^{(\ell)} - x_{j,s}^{(\ell)}\|_2. \tag{33}$$

Applying Cauchy-Schwarz: $\sum_{s=1}^{n} a_s \leq \sqrt{n}\sqrt{\sum_{s=1}^{n} a_s^2}$ with $a_s = \|x_{i,s}^{(\ell)} - x_{j,s}^{(\ell)}\|_2$ yields the result. □

D.4. Proof for Theorem 4.3

Proposition D.2 (Effective Fractional Diffusion via Contractive Composition)**.** *Let $P \in \mathbb{R}^{N\times N}$ be a symmetric stochastic propagation operator with spectral radius $\rho(P) \leq 1$. Consider a sequence of residual mixing operators*

$$H_\ell \;=\; I + \frac{1}{n}(P - I) + E_\ell, \qquad \ell = 1, \ldots, n,$$

where each perturbation satisfies $\|E_\ell\|_2 \leq \varepsilon$ and $\|H_\ell\|_2 \leq 1$. Then, for sufficiently small ε and large n, the composed operator satisfies

$$\prod_{\ell=1}^{n} H_\ell \;=\; P^{\,1+O(\varepsilon n)} \;+\; O(\varepsilon n),$$

where the error is measured in operator norm.

Remark D.3 (Interpretation). Proposition D.2 does not assert that the network explicitly computes a matrix power $P^{1/n}$. Rather, it formalizes that depth-n compositions of near-identity contractive residual mixers approximate the *effective behavior* of fractional diffusion. The use of $P^{1/n}$ in subsequent analyses should therefore be understood as a stylized model capturing the cumulative effect of repeated small contractive updates.

Theory–Implementation Interface Our theoretical analysis studies idealized contractive operators and fractional diffusion processes in order to isolate the structural source of over-smoothing and multi-stream slowdown. The implemented architectures do not explicitly compute operators such as $P^{1/n}$ or exact matrix powers. Instead, they realize a discrete approximation through repeated near-identity residual mixing steps constrained to a contractive operator manifold. The theory should therefore be interpreted as a mechanistic explanation of observed behavior, rather than a literal description of the forward pass.

Restaing the Theorem in details

Theorem D.4 (Multi-stream over-smoothing rate under contractive residual mixers)**.** *Let $P \in \mathbb{R}^{N\times N}$ be a symmetric graph propagation operator with largest eigenvalue $\lambda_1(P) = 1$ and second-largest eigenvalue $\lambda_2(P) \leq 1 - \gamma$ for some $\gamma \in (0, 1]$. Let $\pi \in \mathbb{R}^N$ be the unit eigenvector of P for eigenvalue 1 (the stationary direction), and define the projector onto the mean-zero subspace $\Pi \;=\; I_N - \pi\pi^\top$. Consider a multi-stream GNN with n streams whose per-layer node update (for each feature column) acts on the node dimension as the* fractional diffusion *$P^{1/n}$ and on the stream dimension by a residual mixer $H_\ell^{\text{res}} \in \mathbb{R}^{n\times n}$:*

$$X^{(\ell+1)} \;=\; \big(P^{1/n} X^{(\ell)}\big)\, H_\ell^{\text{res}}, \qquad X^{(\ell)} \in \mathbb{R}^{N\times n\times d}, \tag{34}$$

where $P^{1/n}$ is the principal matrix power (well-defined since P is symmetric). Assume the mixers satisfy the near-identity bound

$$\|H_\ell^{\text{res}}\|_2 \leq 1 + \epsilon \qquad \textit{and} \qquad \|H_\ell^{\text{res}} - I_n\|_2 \leq \epsilon \quad \textit{for all } \ell, \tag{35}$$

for some $0 \leq \epsilon \leq 1$ (e.g., identity-anchored initialization and small drift).

Then for any two nodes $i, j \in \{1, \ldots, N\}$ and any depth $L \geq 0$,

$$\|X_i^{(L)} - X_j^{(L)}\|_F \;\leq\; (1-\gamma)^{L/n}\,(1+\epsilon)^L \;\cdot\; \|\Pi X^{(0)}\|_{(i,j),F}, \tag{36}$$

where $\|\Pi X^{(0)}\|_{(i,j),F}$ denotes the Frobenius norm of the initial pairwise difference after removing the stationary component along π (equivalently, a constant C depending only on initial diversity). In particular, the depth-induced loss of node-to-node diversity is governed by the rate

$$(1-\gamma)^{L/n} \cdot (1+\epsilon)^L,$$

showing an n-fold slowdown in the diffusion-driven contraction term, independent of the specific manifold used for H_ℓ^{res} beyond the contractive monoid / identity-anchoring requirements.

Proof. We prove (36) in four steps:

Step 0: Reduce to linear operators on each feature column. Fix any layer ℓ and any feature coordinate $k \in \{1, \dots, d\}$. Let $Z^{(\ell)}_{\cdot,\cdot,k} \in \mathbb{R}^{N\times n}$ denote the $(N \times n)$ matrix of that feature across nodes (rows) and streams (columns). The update (34) applies identically to each feature:

$$Z^{(\ell+1)} \;=\; (P^{1/n} Z^{(\ell)})\, H_\ell^{\mathrm{res}}, \tag{37}$$

where we suppress the feature index k for readability. Since the Frobenius norm over the d features is the ℓ_2 sum of the Frobenius norms of each feature-slice, it suffices to derive a bound for one feature-slice and then sum over k; equivalently, we can work directly with the Frobenius norm over $(n \times d)$ at a fixed node pair because all steps below are norm-submultiplicative.

Step 1: Contraction of diffusion on the mean-zero subspace. Because P is symmetric, it admits an orthonormal eigen-decomposition

$$P \;=\; U\Lambda U^\top,$$

where U is orthonormal and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$ with $\lambda_1 = 1$ and $\lambda_r \le 1-\gamma$ for all $r \ge 2$ (this is exactly the spectral gap assumption). Let π be the unit eigenvector associated with $\lambda_1 = 1$; then π is the first column of U (up to sign). Define the projector $\Pi = I_N - \pi\pi^\top$, which projects any vector onto the subspace orthogonal to π.

We claim that for any matrix $M \in \mathbb{R}^{N\times m}$ (any number of columns m),

$$\|\Pi P^t M\|_F \;\le\; (1-\gamma)^t\, \|\Pi M\|_F \qquad \text{for all } t \ge 0. \tag{38}$$

To prove (38), note first that since P is symmetric, the principal power is

$$P^t \;=\; U\Lambda^t U^\top, \qquad \Lambda^t = \mathrm{diag}(\lambda_1^t, \dots, \lambda_N^t).$$

Also $\Pi = U\,\mathrm{diag}(0, 1, 1, \dots, 1)\, U^\top$ because it removes the π-component.

Now let M have columns $m_1, \dots, m_m \in \mathbb{R}^N$. Using orthonormal invariance of Euclidean norms,

$$\|\Pi P^t m_j\|_2 = \left\|U\,\mathrm{diag}(0, \lambda_2^t, \dots, \lambda_N^t)\, U^\top m_j\right\|_2 = \left\|\mathrm{diag}(0, \lambda_2^t, \dots, \lambda_N^t)\, U^\top m_j\right\|_2.$$

Since $\max_{r\ge 2} |\lambda_r| \le 1-\gamma$, we have $\max_{r\ge 2} |\lambda_r|^t \le (1-\gamma)^t$. Therefore

$$\|\Pi P^t m_j\|_2 \le (1-\gamma)^t \left\|\mathrm{diag}(0, 1, \dots, 1)\, U^\top m_j\right\|_2 = (1-\gamma)^t \|\Pi m_j\|_2.$$

Square and sum over $j = 1, \dots, m$ to obtain the Frobenius bound:

$$\|\Pi P^t M\|_F^2 = \sum_{j=1}^m \|\Pi P^t m_j\|_2^2 \le \sum_{j=1}^m (1-\gamma)^{2t} \|\Pi m_j\|_2^2 = (1-\gamma)^{2t} \|\Pi M\|_F^2.$$

Taking square roots yields (38).

We will apply (38) with $t = 1/n$ to get the per-layer factor $(1-\gamma)^{1/n}$:

$$\|\Pi P^{1/n} M\|_F \;\le\; (1-\gamma)^{1/n}\, \|\Pi M\|_F. \tag{39}$$

Step 2: Stream mixing increases norms by at most $(1+\epsilon)$. For any matrix $A \in \mathbb{R}^{N\times n}$ and any $H \in \mathbb{R}^{n\times n}$, submultiplicativity of the operator norm gives

$$\|AH\|_F \;\le\; \|A\|_F\, \|H\|_2. \tag{40}$$

(Proof of (40): $\|AH\|_F^2 = \mathrm{tr}(H^\top A^\top AH) \le \lambda_{\max}(H^\top H)\, \mathrm{tr}(A^\top A) = \|H\|_2^2 \|A\|_F^2$.)

Applying (40) with $H = H_\ell^{\mathrm{res}}$ and using (35),

$$\|AH_\ell^{\mathrm{res}}\|_F \;\le\; (1+\epsilon)\,\|A\|_F. \tag{41}$$

Step 3: Combine diffusion contraction and mixing expansion across layers. Take the feature-slice update (37):

$$Z^{(\ell+1)} \;=\; (P^{1/n}Z^{(\ell)})\,H_\ell^{\mathrm{res}}.$$

Project onto the mean-zero subspace in the node dimension by left-multiplying with Π:

$$\Pi Z^{(\ell+1)} = \Pi(P^{1/n}Z^{(\ell)})\,H_\ell^{\mathrm{res}}.$$

Now apply (41) with $A = \Pi(P^{1/n}Z^{(\ell)})$:

$$\|\Pi Z^{(\ell+1)}\|_F \le (1+\epsilon)\,\|\Pi(P^{1/n}Z^{(\ell)})\|_F.$$

Then apply (39) with $M = Z^{(\ell)}$:

$$\|\Pi(P^{1/n}Z^{(\ell)})\|_F \le (1-\gamma)^{1/n}\,\|\Pi Z^{(\ell)}\|_F.$$

Combining the last two displays yields the per-layer recursion

$$\|\Pi Z^{(\ell+1)}\|_F \le (1-\gamma)^{1/n}\,(1+\epsilon)\,\|\Pi Z^{(\ell)}\|_F. \tag{42}$$

Iterate (42) from $\ell = 0$ to $\ell = L-1$:

$$\|\Pi Z^{(L)}\|_F \le \big((1-\gamma)^{1/n}(1+\epsilon)\big)^L\,\|\Pi Z^{(0)}\|_F = (1-\gamma)^{L/n}\,(1+\epsilon)^L\,\|\Pi Z^{(0)}\|_F.$$

This holds for every feature-slice k; equivalently it holds for the full tensor $X^{(L)}$ when the Frobenius norm is taken over the stream and feature dimensions for each node.

Step 4: Convert the global mean-zero bound to a pairwise node bound. Define the node-wise stacked representation (all streams and features) as $R_i^{(\ell)} \in \mathbb{R}^{n\times d}$, the slice of $X^{(\ell)}$ at node i. Consider the pairwise difference $\Delta_{ij}^{(\ell)} = R_i^{(\ell)} - R_j^{(\ell)}$.

The operator Π removes the stationary component along π, i.e., the component that is constant across nodes in the direction of π. In standard over-smoothing analyses, pairwise diversity is controlled by the norm of the mean-zero part, because the stationary component cannot increase node-to-node differences (it is shared). Formally, for any tensor X and any nodes i, j,

$$\|R_i - R_j\|_F \;\le\; \|\Pi X\|_{(i,j),F}, \tag{43}$$

where $\|\Pi X\|_{(i,j),F}$ denotes the Frobenius norm of the difference between node-slices after projection. (One way to see (43) is to write $X = (\pi\pi^\top)X + \Pi X$. The term $(\pi\pi^\top)X$ has all node-slices proportional to π_i times the same global template, so its contribution to $R_i - R_j$ is bounded and in particular does not counteract contraction; bounding pairwise differences by the projected component is standard and yields a constant depending on the initial diversity after removing stationarity.)

Applying the bound from Step 3 to the full representation gives

$$\|\Pi X^{(L)}\|_{(i,j),F} \le (1-\gamma)^{L/n}(1+\epsilon)^L\,\|\Pi X^{(0)}\|_{(i,j),F}.$$

Combining with (43) yields (36), completing the proof. □

Remark D.5 (Local dilution vs. global diffusion slowdown). The two results describe complementary mechanisms. Stream aggregation produces a one-shot reduction in representation differences through averaging, while fractional diffusion distributes the graph's spectral contraction across streams, yielding an exponential slowdown in depth. Importantly, neither mechanism depends on doubly stochastic structure: the Birkhoff polytope is a sufficient instantiation, but any contractive monoid with identity anchoring induces the same asymptotic behavior.

D.5. Proof of Theorem 4.5

Proof of Theorem 4.5. We proceed in four steps: (1) rewrite $\mathcal{D}(x)$ as a centered Frobenius norm, (2) show mean preservation implies centering commutes with H, (3) apply a sharp operator-norm lower bound from near-identity, and (4) conclude.

Step 1: Express diversity via the centering matrix. Let $\mathbf{1} \in \mathbb{R}^n$ denote the all-ones vector and define the averaging and centering matrices

$$P := \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \qquad C := I_n - P. \tag{44}$$

Note that P is the orthogonal projector onto $\mathrm{span}\{\mathbf{1}\}$ and C is the orthogonal projector onto the subspace orthogonal to $\mathbf{1}$ (i.e., the zero-mean subspace). Indeed:

$$P^2 = P, \quad P^\top = P, \qquad C^2 = C, \quad C^\top = C, \qquad CP = PC = 0. \tag{45}$$

Write $x \in \mathbb{R}^{n\times d}$ as a matrix whose s-th row is $x_s^\top$. The stream mean $\bar{x} \in \mathbb{R}^d$ satisfies

$$\bar{x}^\top = \frac{1}{n}\mathbf{1}^\top x. \tag{46}$$

The matrix whose every row equals $\bar{x}^\top$ is Px, because

$$(Px)_{s,:} = \left(\frac{1}{n}\mathbf{1}\mathbf{1}^\top x\right)_{s,:} = \frac{1}{n}\mathbf{1}_s(\mathbf{1}^\top x) = \frac{1}{n}(\mathbf{1}^\top x) = \bar{x}^\top. \tag{47}$$

Therefore, the centered matrix of deviations from the mean is

$$x - Px = (I_n - P)x = Cx. \tag{48}$$

Now compute $\mathcal{D}(x)$:

$$\mathcal{D}(x) = \frac{1}{n}\sum_{s=1}^{n}\|x_s - \bar{x}\|_2^2 = \frac{1}{n}\sum_{s=1}^{n}\|(Cx)_{s,:}\|_2^2 \tag{49}$$

$$= \frac{1}{n}\|Cx\|_F^2, \tag{50}$$

where $\|\cdot\|_F$ is the Frobenius norm (sum of squared row norms).

Step 2: Mean-preserving implies $HC = CH$. The assumptions $H\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top H = \mathbf{1}^\top$ imply that H preserves the $\mathbf{1}$-subspace and also preserves averages.

First, observe that

$$HP = H\left(\frac{1}{n}\mathbf{1}\mathbf{1}^\top\right) = \frac{1}{n}(H\mathbf{1})\mathbf{1}^\top = \frac{1}{n}\mathbf{1}\mathbf{1}^\top = P. \tag{51}$$

Similarly,

$$PH = \left(\frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)H = \frac{1}{n}\mathbf{1}(\mathbf{1}^\top H) = \frac{1}{n}\mathbf{1}\mathbf{1}^\top = P. \tag{52}$$

Thus, $HP = PH = P$. Using $C = I_n - P$,

$$HC = H(I_n - P) = H - HP = H - P, \tag{53}$$

$$CH = (I_n - P)H = H - PH = H - P. \tag{54}$$

Therefore

$$HC = CH. \tag{55}$$

Step 3: Lower bound $\|Hy\|_2$ for near-identity H. Let $E := H - I_n$. The assumption $\|H - I_n\|_2 \leq \epsilon$ is exactly $\|E\|_2 \leq \epsilon$. For any vector $y \in \mathbb{R}^n$,

$$Hy = (I_n + E)y = y + Ey. \tag{56}$$

Apply the triangle inequality in the form $\|a + b\|_2 \geq \|a\|_2 - \|b\|_2$:

$$\|Hy\|_2 = \|y + Ey\|_2 \geq \|y\|_2 - \|Ey\|_2. \tag{57}$$

Then use the definition of the operator norm: $\|Ey\|_2 \leq \|E\|_2\|y\|_2 \leq \epsilon\|y\|_2$. Hence

$$\|Hy\|_2 \geq (1 - \epsilon)\|y\|_2. \tag{58}$$

Step 4: Apply the bound column-wise to Cx and conclude. Let $z := Cx \in \mathbb{R}^{n\times d}$ be the centered stream matrix. Let $z^{(j)} \in \mathbb{R}^n$ denote its j-th column (for feature dimension $j = 1, \ldots, d$). Then

$$\|z\|_F^2 = \sum_{j=1}^{d} \|z^{(j)}\|_2^2. \tag{59}$$

Using Step 2 ($HC = CH$),

$$C(Hx) = (CH)x = (HC)x = H(Cx) = Hz. \tag{60}$$

Thus

$$\mathcal{D}(Hx) = \frac{1}{n}\|C(Hx)\|_F^2 = \frac{1}{n}\|Hz\|_F^2. \tag{61}$$

Expand $\|Hz\|_F^2$ by columns:

$$\|Hz\|_F^2 = \sum_{j=1}^{d} \|Hz^{(j)}\|_2^2. \tag{62}$$

Now apply the vector lower bound (58) to each column $z^{(j)}$:

$$\|Hz^{(j)}\|_2 \geq (1 - \epsilon)\|z^{(j)}\|_2 \quad \Longrightarrow \quad \|Hz^{(j)}\|_2^2 \geq (1 - \epsilon)^2\|z^{(j)}\|_2^2. \tag{63}$$

Summing over j yields

$$\|Hz\|_F^2 = \sum_{j=1}^{d} \|Hz^{(j)}\|_2^2 \geq \sum_{j=1}^{d}(1 - \epsilon)^2\|z^{(j)}\|_2^2 = (1 - \epsilon)^2 \sum_{j=1}^{d} \|z^{(j)}\|_2^2 = (1 - \epsilon)^2\|z\|_F^2. \tag{64}$$

Substituting back $z = Cx$ and $\mathcal{D}(x) = \frac{1}{n}\|Cx\|_F^2$ from Step 1, we conclude

$$\mathcal{D}(Hx) = \frac{1}{n}\|Hz\|_F^2 \geq \frac{1}{n}(1 - \epsilon)^2\|z\|_F^2 = (1 - \epsilon)^2\, \mathcal{D}(x). \tag{65}$$

Finally, since $(1 - \epsilon)^2 = 1 - 2\epsilon + \epsilon^2 \geq 1 - 2\epsilon$ for all $\epsilon \in [0, 1)$, we also have $\mathcal{D}(Hx) \geq (1 - 2\epsilon)\mathcal{D}(x)$.

This completes the proof. □

Remark (Why mean preservation matters; approximate variants). The definition $\mathcal{D}(x)$ measures variance about the *stream mean*. To compare $\mathcal{D}(Hx)$ to $\mathcal{D}(x)$, one must control how H acts on the mean direction $\mathrm{span}\{\mathbf{1}\}$. The conditions $H\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top H = \mathbf{1}^\top$ ensure exact mean preservation and make the bound sharp. For general contractive H (e.g., spectral-norm ball) one can obtain a similar bound by adding an explicit *approximate mean preservation* condition such as $\|\mathbf{1}^\top H - \mathbf{1}^\top\|_2 \leq \delta$, which introduces an additive $O(\delta)$ slack term in the final inequality.

D.6. Proof of Theorem 4.6

Architecture. Fix a number of streams $K \geq 2$. Each node $v \in V$ carries a stacked state $H^{(\ell)}(v) \in \mathbb{R}^{K\times d}$ at layer ℓ, where the k-th stream is $h_k^{(\ell)}(v) \in \mathbb{R}^d$.

A single layer consists of:

$$\tilde{h}_k^{(\ell)}(v) = \phi\left(W_\ell\, h_k^{(\ell-1)}(v) \;+\; \sum_{u\in N(v)} \psi_\ell\big(h_k^{(\ell-1)}(u)\big)\right) \in \mathbb{R}^d, \tag{66}$$

$$H^{(\ell)}(v) = \mathrm{Mix}_\ell\Big(\tilde{H}^{(\ell)}(v)\Big) := M_\ell\, \tilde{H}^{(\ell)}(v), \tag{67}$$

where $\tilde{H}^{(\ell)}(v) \in \mathbb{R}^{K\times d}$ stacks $\tilde{h}_k^{(\ell)}(v)$ by rows, ϕ and ψ_ℓ are MLPs applied row-wise, and the stream mixer $M_\ell \in \mathcal{M} \subseteq \mathbb{R}^{K\times K}$ acts only on the stream index.

We assume $\mathcal{M}$ is a *contractive monoid*: it contains I_K, is closed under multiplication, and all its elements are non-expansive (e.g., $\|M\| \leq 1$ in an operator norm). Additionally, assume $\mathcal{M}$ contains a non-trivial permutation matrix $P \neq I_K$.

Lemma D.6 (Exact stream reindexing and copying)**.** *If $\mathcal{M}$ contains a permutation matrix $P \neq I_K$, then for any layer ℓ the map $H \mapsto PH$ is realizable as a mixer. In particular, if P is a transposition swapping streams a and b, then P swaps the corresponding stream states at every node:*

$$(PH)(v)_a = H(v)_b, \qquad (PH)(v)_b = H(v)_a, \qquad (PH)(v)_k = H(v)_k \;\; \forall k \notin \{a,b\}.$$

Moreover, since $\mathcal{M}$ is closed under multiplication and contains I_K, any finite composition of such permutation mixers and identity mixers is realizable.

Proof. A permutation matrix P is a linear operator on the stream index. By assumption $P \in \mathcal{M}$, so we may select $M_\ell = P$ at layer ℓ. Then for any node v, $H^{(\ell)}(v) = P\,\tilde{H}^{(\ell)}(v)$ holds by definition of Mix_ℓ. The componentwise identities follow from the definition of a permutation matrix. Closure under multiplication implies any product of mixers from $\mathcal{M}$ remains in $\mathcal{M}$, so any finite composition of such permutations and identities is realizable. □

Lemma D.7 (A single layer can compute a coupled self–neighbor polynomial)**.** *Consider two streams ($K = 2$). There exists a choice of ϕ, ψ_ℓ, W_ℓ and mixers $M_\ell \in \{I_2, P\}$ (with P swapping the two streams) such that for each node v the layer output contains the following information (as part of some coordinates of $h_1^{(\ell)}(v)$):*

$$\Big(\, h_1^{(\ell-1)}(v),\;\; \sum_{u\in N(v)} h_1^{(\ell-1)}(u),\;\; h_1^{(\ell-1)}(v) \odot \sum_{u\in N(v)} h_1^{(\ell-1)}(u) \,\Big),$$

where $\odot$ denotes elementwise product.

Proof. We construct the layer explicitly.

Step 1 (copy self into both streams). Choose message passing so that $\tilde{h}_1(v) = \tilde{h}_2(v) = h_1^{(\ell-1)}(v)$ by setting $\psi_\ell \equiv 0$, W_ℓ to select stream-1 input, and ϕ to be the identity on the relevant coordinates. This is feasible because ϕ is an MLP and can implement coordinate selection and identity on a subspace.

Step 2 (aggregate neighbors into stream 2 only). Now choose message passing so that in the next affine+sum stage, stream 2 receives neighbor sums while stream 1 remains unchanged:

$$\tilde{h}_1(v) = h_1^{(\ell-1)}(v), \qquad \tilde{h}_2(v) = \sum_{u\in N(v)} h_1^{(\ell-1)}(u).$$

This can be done by setting W_ℓ appropriately for each stream (block-diagonal with different blocks per stream) and choosing ψ_ℓ to map $h_1^{(\ell-1)}(u)$ into the coordinates accumulated by the neighborhood sum for stream 2, while mapping to zero for stream 1.

Step 3 (form a coupled interaction). Finally, apply ϕ (row-wise MLP) on the concatenation of the coordinates within each stream. Since ϕ is an MLP, it can output simultaneously: (1) the identity of $h_1^{(\ell-1)}(v)$ (present in stream 1), (2) the

identity of $\sum_{u\in N(v)} h_1^{(\ell-1)}(u)$ (present in stream 2), and (3) their elementwise product by implementing multiplication via a standard MLP construction (e.g., using the fact that MLPs can approximate products on bounded domains; with sufficient width this can be made exact on discrete one-hot domains, or arbitrarily accurate otherwise).

To make both terms accessible in one stream's coordinates, use the permutation mixer P to swap streams when needed and concatenate outputs across layers (or reserve disjoint coordinate blocks to store each quantity). Thus the claimed triple can be embedded into the output coordinates of $h_1^{(\ell)}(v)$. □

Lemma D.8 (Multi-scale walk signatures under shared-parameter multi-stream mixing)**.** *Consider the architecture in Eqs. (70)–(71) with $K \geq 2$ streams, shared parameters $(W_\ell, \psi_\ell, \phi)$ across streams, and per-layer stream mixing matrices $M_\ell \in \mathcal{M}$ (a contractive monoid). Assume ϕ and ψ_ℓ are sufficiently expressive (e.g., MLPs), and let* AGG *be a permutation-invariant neighborhood sum as in Eq. (70). Then for any depth L, there exists a choice of parameters and readout such that, for each node v, the final representation $H^{(L)}(v)$ determines a* multi-scale signature *of the form*

$$\mathcal{S}_L(v) \;=\; \big\{\, s_{k,t}(v) \;:\; k = 1,\dots,K,\; t = 0,\dots,L \,\big\}, \qquad s_{k,t}(v) \;=\; \sum_{u\in V} \big(A^t\big)_{vu}\, \eta_k(x_u),$$

where A is the adjacency matrix of the graph, x_u are initial node features, and η_k are (learned) feature maps (potentially distinct across k via stream content). In particular, $H^{(L)}(v)$ can encode and combine multiple hop-scales (0-hop, 1-hop, …, L-hop) through contractive mixing across streams.

Proof. We give a constructive parameter setting.

Step 1 (encode per-stream feature maps). Initialize streams so that $h_k^{(0)}(v) = \eta_k(x_v)$ for learned maps η_k, which is realizable by the input embedding layer (or by taking x_v already embedded).

Step 2 (realize linear neighborhood propagation). Choose ψ_ℓ and ϕ so that Eq. (70) implements, on a designated coordinate block, the update

$$\tilde{h}_k^{(\ell)}(v) = \sum_{u\in N(v)} h_k^{(\ell-1)}(u) \quad \text{and} \quad \tilde{h}_k^{(\ell)}(v) \text{ also retains } h_k^{(\ell-1)}(v) \text{ on a disjoint block.}$$

This is achieved by letting ψ_ℓ be the identity on the propagation block and zero elsewhere, setting W_ℓ to route $h_k^{(\ell-1)}(v)$ into the retention block, and letting ϕ be the identity on these blocks.

Step 3 (induction on hop-length). By construction, the propagation block of $\tilde{h}_k^{(\ell)}(v)$ equals $\sum_{u\in N(v)} \tilde{h}_k^{(\ell-1)}(u)$, which corresponds to multiplying by A. Thus after t propagation steps, the resulting block equals $\sum_u (A^t)_{vu}\eta_k(x_u)$. The retention blocks allow storing $t = 0,\dots,L$ simultaneously across depth using disjoint coordinates.

Step 4 (combine scales across streams). Eq. (71) applies M_ℓ to mix streams at each node. Because $\mathcal{M}$ is a monoid, compositions of the M_ℓ remain in $\mathcal{M}$, so the network can stably combine the per-stream multi-hop summaries into a joint signature without affecting permutation-equivariance.

Therefore $H^{(L)}(v)$ can be parameterized to determine $\mathcal{S}_L(v)$ as claimed. □

Edge-lift architecture. For each directed edge $e = (u \to v)$, define the K-stream state $E^{(\ell)}(u \to v) \in \mathbb{R}^{K\times d}$. A layer updates:

$$\tilde{e}_k^{(\ell)}(u \to v) = \phi\left(W_\ell\, e_k^{(\ell-1)}(u \to v) + \sum_{w\in N(u)\setminus\{v\}} \psi_\ell\big(e_k^{(\ell-1)}(w \to u)\big) \right), \tag{68}$$

$$E^{(\ell)}(u \to v) = M_\ell\, \tilde{E}^{(\ell)}(u \to v), \qquad M_\ell \in \mathcal{M}. \tag{69}$$

A node embedding can be obtained by aggregating its incoming edge states, and a graph embedding by pooling over nodes.

Lemma D.9 (Edge-lift realizes pair-refinement statistics locally)**.** *In the edge-lift architecture above, the state on a directed edge $(u \to v)$ after ℓ layers can be parameterized to encode a function of the rooted, directed ℓ-neighborhood of the ordered pair (u, v) (equivalently, the ℓ-neighborhood of the node $(u \to v)$ in the directed line graph). Consequently, the model can refine representations based on ordered pairs (u, v) rather than only single nodes.*

Proof. The update for $(u \to v)$ aggregates messages from $(w \to u)$ for $w \in N(u) \setminus \{v\}$, which are exactly the incoming neighbors of $(u \to v)$ in the directed line graph. Thus the update is a standard message-passing step on the lifted graph, and by induction $E^{(\ell)}(u \to v)$ depends only on the radius-ℓ neighborhood of $(u \to v)$ in that lift. Since $(u \to v)$ represents an ordered pair, this yields pair-dependent refinement. □

Proof of Theorem 4.6. We use the strictness of the WL hierarchy: there exist graphs G, H that are indistinguishable by 1-WL but distinguishable by 2-WL. It is standard that 2-WL can be viewed as 1-WL operating on ordered pairs (the 2-tuple graph), where the color of a pair (u, v) is refined using the multiset of colors of pairs (u, w) and (w, v).

By Lemma D.9, the edge-lift network performs local message passing on directed edges, which are ordered pairs constrained to adjacency. By using two streams and a non-trivial permutation mixer in $\mathcal{M}$, the model can maintain and exchange (i) a "left" context summary (messages into u) and (ii) a "right" context summary (messages out of v via the next layer's aggregation after swapping), thereby implementing a pair-color refinement that distinguishes the same separating graphs as a suitable fragment of 2-WL. Finally, pooling these refined edge representations back to node/graph level yields different outputs on G and H.

Thus the architecture strictly exceeds 1-WL. □

E. Theoretical Remarks

E.1. Why the Threshold is Exactly 1

If instead $\|H_\ell\|_2 \leq k$ for some $k > 1$, the perturbation bound (Eq. 9) becomes

$$\|\delta h^{(0)}\|_2 \prod_\ell \left(k + L_\ell \|G_\ell\|_2 \|K_\ell\|_2\right). \tag{70}$$

Even when the GNN transformation contributes no amplification ($L_\ell \|G_\ell\|_2 \|K_\ell\|_2 = 0$, e.g., at initialization), the residual path alone introduces a factor k^L. At $L = 128$ with $k = 1.01$, this exceeds $3.5\times$ — an exponential instability channel independent of the backbone.

The constraint $\|H_\ell\|_2 \leq 1$ is the **exact threshold** at which the residual mixer ceases to be an independent source of exponential amplification. This is confirmed empirically: at 128 layers on Cora (Table 3), unconstrained GCN degrades to 14.47% while all contractive variants remain above 72%.

E.2. Choice of operator norm

The contractive monoid definition (Definition 3.1) is norm-agnostic — any submultiplicative norm yields a valid instantiation. We use the spectral norm ($\|H\|_2$) due to its tight connection to worst-case perturbation amplification. The infinity norm ($\|H\|_\infty \leq 1$, constraining maximum absolute row sum) provides a viable alternative that controls per-node feature bounds rather than global energy — potentially natural for graphs with heterogeneous degree distributions.

Notably, the Frobenius norm is *not* submultiplicative, so the set $\{H : \|H\|_F \leq 1\}$ violates compositional closure (Axiom 3) and cannot define a contractive monoid. The axioms thus rule out inappropriate norms *a priori*, without requiring empirical trial-and-error.

E.3. Directed and non-symmetric graphs

The contractive monoid axioms (Definition 3.1) and the stability guarantee (Theorem 4.1) make no reference to graph symmetry and apply to directed graphs without modification. The specific over-smoothing rate $(1-\gamma)^{L/n}$ in Theorem 4.3 assumes a symmetric propagation operator P for its spectral decomposition (Appendix D.4, Step 1).

For diagonalizable non-symmetric P with $|\lambda_2(P)| \leq 1 - \gamma$, the same n-fold slowdown holds via Schur decomposition. For non-diagonalizable P, Jordan blocks introduce polynomial prefactors eventually dominated by exponential contraction — the qualitative conclusion persists.

The architectural prescription — constrain residual mixers to a contractive monoid — is thus graph-agnostic.

E.4. Detailed Comparison with Operator Constraint Methods

Spectral normalization (Miyato et al., 2018) constrains individual weight matrices to $\|W\|_2 \leq 1$, addressing non-expansiveness at the single-layer level. However, it does not ensure that admissible matrices are closed under multiplication or that identity initialization lies within the constraint set. These compositional requirements are absent when normalizing isolated weights but become essential when constraining residual mixers, whose effective transformation after L layers is $\prod_{\ell=1}^{L} H_\ell$.

Orthogonal and unitary RNNs (Arjovsky et al., 2016; Helfrich et al., 2018) use norm-preserving recurrence for gradient stability in sequential models. Our orthogonal instantiation (Section 3.1) brings this idea to multi-stream GNN residual mixing, where the failure mode is graph-spectral contraction rather than vanishing gradients in time.

Lipschitz-bounded networks (Anil et al., 2019) enforce global Lipschitz bounds for certified robustness. Our non-expansiveness constraint serves a different purpose: it prevents the residual mixer from amplifying perturbations across streams, decoupling mixer-induced instability from the diffusion inherent in GNN layers (Theorem 4.1). PairNorm and DropEdge operate on different components entirely (node features and graph topology) and are complementary to our framework.

F. Practitioner's Guide

The containment hierarchy

$$\text{Diagonal} \subset \text{Orthogonal} \subset \text{Spectral Norm Ball} \supseteq \text{Birkhoff}$$

yields concrete design guidance:

Design Priority	Recommended Monoid	Rationale
Max expressiveness	Spectral Norm Ball	Maximally expressive contractive monoid; richest class of residual transformations
Exact norm preservation	Orthogonal Group	$\|Hx\|_2 = \|x\|_2$; zero information loss per layer
Computational efficiency	Diagonal Contraction	$O(n)$ vs. $O(n^2)$ mixing cost; negligible overhead
Mean + mass conservation	Birkhoff Polytope	$H\mathbf{1} = \mathbf{1}$, $\mathbf{1}^\top H = \mathbf{1}^\top$; strongest structural guarantees

The theory predicts that less restrictive monoids offer broader representational capacity at the cost of weaker structural properties.

G. Representation Diagonistics

G.1. Over-smoothing Analysis via Representation Collapse Metrics

To understand how different residual mixing constraints affect representation collapse in deep graph networks, we analyze the evolution of node embeddings across depth using two complementary diagnostics. These metrics quantify the extent to which node representations become indistinguishable as depth increases, a phenomenon commonly referred to as *over-smoothing*.

Let $Z^{(l)} \in \mathbb{R}^{N \times d}$ denote the matrix of node embeddings after the l-th layer, where N is the number of nodes and d is the hidden dimension.

Metric 1: Mean Pairwise Cosine Similarity. Over-smoothing is often characterized by node embeddings collapsing toward a single dominant direction in feature space. To quantify this effect, we compute the mean pairwise cosine similarity between normalized node embeddings at each layer.

We first normalize each node embedding:

$$\hat{z}_i^{(l)} = \frac{z_i^{(l)}}{\|z_i^{(l)}\|_2}, \quad i = 1, \ldots, N. \tag{71}$$

The mean pairwise cosine similarity at layer l is then defined as:

$$\mathrm{CosSim}(l) = \frac{1}{N(N-1)} \sum_{i \neq j} \left(\hat{z}_i^{(l)}\right)^\top \hat{z}_j^{(l)}. \tag{72}$$

This quantity lies in $[-1, 1]$ and increases as embeddings become more aligned. Values approaching 1 indicate severe directional collapse, whereas values near 0 correspond to diverse representations.

For computational efficiency, we use the equivalent formulation:

$$\mathrm{CosSim}(l) = \frac{\left\|\sum_{i=1}^N \hat{z}_i^{(l)}\right\|_2^2 - N}{N(N-1)}, \tag{73}$$

which avoids explicit $\mathcal{O}(N^2)$ pairwise computations.

Metric 2: Feature Variance Across Nodes. While cosine similarity captures directional alignment, it does not directly measure the magnitude of representation collapse. We therefore complement it with a variance-based metric that quantifies the dispersion of node features across the graph.

For each feature dimension $k \in \{1, \ldots, d\}$, we compute the variance across nodes:

$$\mathrm{Var}_k^{(l)} = \frac{1}{N} \sum_{i=1}^N \left(Z_{i,k}^{(l)} - \mu_k^{(l)}\right)^2, \quad \mu_k^{(l)} = \frac{1}{N} \sum_{i=1}^N Z_{i,k}^{(l)}. \tag{74}$$

The mean feature variance at layer l is then defined as:

$$\mathrm{VarMean}(l) = \frac{1}{d} \sum_{k=1}^d \mathrm{Var}_k^{(l)}. \tag{75}$$

A decreasing $\mathrm{VarMean}(l)$ indicates that node embeddings are collapsing toward a low-variance subspace or a single point, a hallmark of over-smoothing.

Usage in Analysis. We compute this metrics at each layer for different residual mixing constraints and report their evolution as a function of depth. $\mathrm{CosSim}(l)$ provide a fine-grained view of representation collapse, capturing both directional alignment and magnitude contraction of node embeddings.

H. Additional Experimental Details

H.1. Complete Hyperparameter Table

Here we provides complete hyperparameter settings for all experiments

H.2. Dataset Statistics

Table 11 provides detailed statistics for all datasets used in our experiments Homophily ratio is computed as the fraction of edges connecting nodes of the same class:$h = \frac{1}{|E|} \sum_{(i,j)\epsilon E} I[yi = yj]$. Datasets with $h > 0.5$ are considered homophilic, while $h < 0.5$ indicates heterophilic graphs where nodes tend to connect to dissimilar nodes.

Table 10. Complete hyperparameter configuration for all experiments.

Hyperparameter	Main Experiments	Abalation Analysis
Hidden dimension	128	128
Dropout rate	0.5	0.5
Learning rate	0.001	0.001
Weight decay	0.0005	0.0005
Max epochs	500	500
Early stopping patience	100	100
Learning rate scheduler	Yes (ReduceLROnPlateau)	Yes
LR scheduler patience	50	50
LR scheduler factor	0.5	0.5
Contractive Monoid specific		
Sinkhorn temperature (τ)	0.1	0.1
Sinkhorn iterations	10	10
Pre-mixing	Dynamic	Dynamic
Residual mixing	Dynamic + Static	Dynamic + Static
Training details		
Batch processing	Full-batch	Full-batch
Optimizer	Adam	Adam
Random seeds	42, 123, 456, 789, 1024	42, 123, 456, 789, 1024
Mixed precision (AMP)	No	No
Hardware	4× NVIDIA A6000 Ada	4× NVIDIA A6000 Ada

Table 11. Dataset statistics for homophilic and heterophilic benchmarks

Dataset	Nodes	Edges	Features	Classes	Homophily
Homophilic					
Cora	2,708	5,429	1,433	7	0.81
CiteSeer	3,327	4,732	3,703	6	0.74
PubMed	19,717	44,338	500	3	0.80
Heterophilic					
Chameleon	2,277	36,101	2,325	5	0.23
Actor	7,600	33,544	931	5	0.22
Texas	183	309	1,703	5	0.11
Wisconsin	251	499	1,703	5	0.21
Cornell	183	295	1,703	5	0.30

H.3. Architecture Details

We evaluate four standard GNN architectures as baselines:

Evaluation protocol:
We report mean test accuracy and standard deviation across 5 random seeds (42, 123, 456,789, 1024) for each configuration.
Computational resources.
All experiments were conducted on a server with:

- 4× NVIDIA A6000 Ada GPUs (48GB memory each)
- AMD EPYC 7763 CPU (128 cores)
- 512GB RAM
- CUDA 12.1, PyTorch 2.1.0, PyTorch Geometric 2.4.0

Training time.
Average wall-clock time per experiment:

- Small datasets (Cora, CiteSeer, WebKB): 5-15 minutes for 8-layer models
- Medium datasets (PubMed, Actor, Chameleon): 10-30 minutes
- Depth analysis (16 layers): 15-25 minutes

The overhead of mHC-GNN compared to baseline is 6-8in wall-clock time, consistent with our theoretical complexity analysis.

H.4. Reproducibility

To ensure reproducibility, we:

1. Fix random seeds for all experiments (PyTorch, NumPy, CUDA)
2. Use deterministic algorithms where possible (torch.use deterministic algorithms)
3. Report results across 5 different random seeds
4. Provide complete hyperparameter configurations
5. Anonymized code is at: https://anonymous.4open.science/r/mhc-gnn-7F22

H.5. Statistical Significance Testing

For all reported improvements, we perform paired t-tests across the 5 random seeds. We report improvements as statistically significant when $p < 0.05$.

I. Hyperparameters and Implementation Details

I.1. Model Hyperparameters

I.2. Computational Environment

All experiments were conducted on NVIDIA RTX A6000 GPUs with 48GB memory. Training time per configuration ranged from 30 seconds (Texas) to 80 seconds (ogbn-arxiv) per seed for 8-layer models. Depth experiments ranged from 120 seconds (16 layers) to approximately 15 minutes (128 layers) per seed.

I.3. Reproducibility

Complete code, data splits, and trained models are available in the supplementary materials. All random seeds are fixed for reproducibility (seeds: 42, 123, 456, 789, 2024).

J. Broader Impact Statement

This work is a foundational contribution to the design of deep graph neural network architectures. By identifying minimal algebraic conditions for stable residual propagation, we expand the design space available to practitioners building GNNs for applications in drug discovery, materials science, social network analysis, and recommendation systems. We do not introduce new datasets, deploy models in sensitive domains, or collect user data. The primary societal impact is indirect: enabling more reliable deep GNNs may improve downstream applications, but the benefits and risks are mediated by how these architectures are deployed. We do not foresee negative societal consequences specific to this architectural contribution.

Table 12. Hyperparameters used across all experiments.

Hyperparameter	Value
Architecture	
Number of layers (main)	8
Number of layers (depth)	2, 4, 8, 16, 32, 64, 128
Hidden dimension	128
Dropout rate	0.5
Expansion rate n	2, 4, 8
Optimization	
Optimizer	Adam
Learning rate	0.001
Weight decay	5×10^4
Max epochs	500
Early stopping patience	100
Sinkhorn-Knopp	
Iterations T	10
Temperature τ	0.1
Data Splits	
Train/Val/Test	60/20/20
Random seeds	5 per config