SHEAF–HAAR EQUIVARIANT ARCHITECTURE FOR HETEROPHILOUS GRAPH LEARNING FROM TOPOLOGY–FEATURE CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterophilous graphs, in which linked nodes tend to have different labels, arise in many real-world networks. Multi-scale spectral approaches such as PEGFAN build their multi-resolution structure on a hierarchical clustering of the graph, yet the clustering itself lies outside the models' permutation-equivariance guarantees, ignores node features, and inherits eigenvector sign and degeneracy ambiguities. We introduce SHEAR, a Sheaf–Haar Equivariant Architecture whose entire multi-resolution structure—partition tree, coarse graphs, and framelet operators—is generated by a single permutation-invariant procedure, so that the end-to-end model, clustering stage included, is permutation equivariant. The hierarchy is built by a divisive granular-ball clustering that couples GPNN core-degree structural ids with node features through a quality-gated recursive bisection, with isomorphism-invariant fingerprints for all tie-breaking. Binary Haar framelet operators are derived in basis-free projector form directly from the assignment matrices, so no eigendecomposition enters the frequency decomposition; at each level, neural sheaf diffusion runs on Galerkin coarse sheaf Laplacians whose restriction maps are generated end-to-end. The framelet channel bank is an exact isometric decomposition of the signal and pooling-based coupling is distance-independent, yielding channel-level guarantees against oversquashing and oversmoothing without filter-gain amplification assumptions. On thirteen node-classification benchmarks from the PEGFAN and HMH suites, evaluated under the same data versions and splits as the baselines, SHEAR is best or second best on six columns and on six of the ten heterophilous ones, attaining the best result on Minesweeper, Texas, Roman-empire by a 13-point margin, and the filtered Squirrel and Chameleon.

1 INTRODUCTION

Graph neural networks (GNNs) that assume homophily degrade sharply on heterophilous graphs, where connected nodes tend to carry different labels (Zhu et al., 2020; Zheng et al., 2022). Spectral and multi-scale approaches address this by extracting information beyond the 1-hop neighborhood. Among them, PEGFAN (Li et al., 2024) constructs binary Haar graph framelets over a recursive clustering of the graph and proves that both the resulting framelet transform and the framelet network built on it are permutation equivariant (Li et al., 2024, Thm. 4, Prop. 2). The guarantee, however, takes the partition tree $\mathcal{P}_K$ as a *given* input: the Ward clustering that generates $\mathcal{P}_K$ lies outside the theorem's scope. Ward linkage operates on a spectral embedding of the adjacency alone, so node features never enter the clustering; the eigenbasis of a degenerate eigenspace is defined only up to rotation; and distance ties are broken by node index. The tree itself is therefore not guaranteed to be invariant under node relabeling.

At the same time, heterophilous graphs are exactly where a purely topological hierarchy is weakest: the label-relevant signal lives in the *features*, and clusters defined by spectral geometry alone can mix nodes whose features (and labels) disagree, while separating structurally similar nodes with nearly identical attributes. A clustering that respects both topology and features, and that does not depend on the node ordering, removes this weak link.

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

A clustering, however, is only a scaffold: on heterophilous graphs it matters just as much how signals are transported across edges, and scalar message passing can only smooth them. We therefore instantiate the invariant hierarchy inside a sheaf architecture, where each edge carries learned linear maps that can separate, rather than average, the signals at its endpoints, and we adopt PEGFAN's own framelets for the frequency decomposition. The resulting model, SHEAR, replaces the Ward clustering by a deterministic, training-free *divisive* procedure: a granular-ball recursive bisection in the spirit of HGBG (Ni et al., 2026), which couples the permutation-invariant core-degree partition of GPNN (Hevapathige & Wang, 2026) with the node features at every decision, from center and seed selection to quality-gated split acceptance (Section 3). With isomorphism-invariant fingerprints for all tie-breaking, the resulting hierarchy is permutation invariant by construction (Section 4.2).

The frequency decomposition, in turn, needs no eigenbases: the binary Haar framelet transform of Li et al. (2024) is, at the operator level, independent of the per-cluster bases used to define it, since the framelet projector $\Psi_j^\top \Psi_j$ is simply the orthogonal projector onto the space of within-cluster detail signals. The framelet operators can therefore be built directly from the partition tree in *basis-free projector form*, as the cluster-mean scaling projection $\mathcal{S}_\ell$ and its orthogonal complement $\mathcal{W}_\ell = I - \mathcal{S}_\ell$, which telescope exactly, $\mathcal{S}_\ell + \mathcal{W}_\ell = I$, and carry no eigenvector sign or degeneracy ambiguity whatsoever (Section 4.3).

Finally, the sheaf geometry that equips the model with heterophily-aware transport can be learned *end-to-end*: a shared SheafMLP regenerates all restriction maps from the current features at every forward pass, and the Galerkin coarse sheaf Laplacians of Naddeo et al. (2026) are rebuilt on the fly through precomputed aggregation indices, with no pretrained encoder and no cached band matrices (Sections 4.4–4.5). Because the resulting framelet channel bank is an exact projector decomposition whose coupling weights are cluster sizes rather than iterated filter gains, it also lends itself to guarantees against oversquashing and oversmoothing, the two classical depth pathologies of graph learning, that avoid any filter-gain amplification assumption (Section 4.10).

The contribution of this paper are: (i) **A topology–feature granular-ball hierarchical clustering** for graph coarsening that couples GPNN structural partitioning with HGBG-style quality-gated bisection—an invariant two-seed splitting rule, a feature–structure–community quality oracle, a hard size cap, and milestone-based extraction of strictly nested levels—and is permutation invariant by construction (Section 4.2). (ii) **A sheaf–framelet architecture built on the invariant hierarchy**: basis-free binary Haar framelet projectors $\mathcal{S}_\ell = (S^{(\ell)} D_\ell^{-1} S^{(\ell)\top}) \otimes I_d$ and $\mathcal{W}_\ell = I - \mathcal{S}_\ell$ with learnable band gains, so no eigendecomposition enters the frequency decomposition (Sections 4.3–4.7), together with multi-scale neural sheaf diffusion on Galerkin coarse sheaf Laplacians whose restriction maps are all regenerated end-to-end by a shared edge-wise SheafMLP (Sections 4.4–4.6); controlled ablations show that the load-bearing role of the sheaf machinery is structural rather than accuracy-critical on these benchmarks (Appendix F). (iii) **End-to-end theoretical guarantees**: since the partition tree is permutation invariant and every operator built on it co-varies, the whole model, clustering included, is permutation equivariant in exact arithmetic under a general-position assumption (Section 4.9, Appendix A); and the exact isometry of the framelet channel bank together with distance-independent pooling-based coupling yields channel-level guarantees against oversquashing and oversmoothing without any filter-gain amplification assumption (Section 4.10, Appendix C). (iv) **A broad empirical study** on thirteen node-classification benchmarks from the protocols of Li et al. (2024, Table III) and Hossen (2026, Table 1): SHEAR is best or second best on six of the thirteen columns, with first places on Minesweeper, Texas, Roman-empire, and the filtered Squirrel and Chameleon, complemented by a sensitivity analysis and a sheaf ablation (Appendix F) and by an extension of the leakage-free CoCN comparison (Sun et al., 2023, Table 3) (Table 8).

2 Related Work

Heterophilous GNNs. Beyond-1-hop aggregation (MixHop (Abu-El-Haija et al., 2019)), ego/neighbor separation (H_2GCN (Zhu et al., 2020)), adaptive filter learning (GPRGNN (Chien et al., 2021), BernNet (He et al., 2021), ChebNetII (He et al., 2022), JacobiConv (Wang & Zhang, 2022)), and structure-aware rewiring (WRGAT (Suresh et al., 2021)) all combat the low-pass bias of message passing. Full-batch spectral-propagation models such as FSGNN (Maurya et al., 2022) decouple propagation from transformation and are strong heterophily baselines. The channel group

(a) of SHEAR is such a decoupled propagation bank; its remaining channel groups add multi-scale framelet detail and sheaf transport instead of learned polynomial filters (Section 4.7).

Framelets, wavelets, and sheaves on graphs. PEGFAN (Li et al., 2024) builds permutation-equivariant Haar-type framelets on a recursive balanced clustering; we reuse its framelet construction in projector form but generate the underlying hierarchy by a permutation-invariant topology–feature clustering. Sheaf neural networks (Hansen & Gebhart, 2019; Bodnar et al., 2022) attach a vector space (stalk) to every node and edge and learn restriction maps that separate, rather than smooth, signals across heterophilous edges. Sheaflets (Li et al., 2026) define a multi-band spectral decomposition on the sheaf Laplacian; our framelet operators are the hard band-limit limit of such spectral windows, realized without any eigendecomposition. Naddeo et al. (2026) introduce hierarchical pooling for sheaf neural networks via Galerkin coarsening, which we adopt in closed form. HMH (Hossen, 2026) constructs a soft hierarchy with a sparse orthonormal Haar basis and learnable diagonal gains; our gains act on two exact framelet bands per level, and our hierarchy is hard, deterministic, and permutation invariant.

Graph partitioning and granular-ball computing. GPNN (Hevapathige & Wang, 2026) shows that deterministic partitions from the k-core decomposition with a core-degree refinement are permutation invariant. HGBG (Ni et al., 2026) brings granular-ball computing to graph pooling by recursively splitting granular balls under a feature–structure coupled quality measure. We adopt its divisive skeleton but replace its centrality measures with GPNN structural ids, add a modularity term to the oracle, and equip every step with isomorphism-invariant tie-breaking, which HGBG does not provide.

Oversquashing and oversmoothing. Oversquashing, the decay of node influence with graph distance (Topping et al., 2022; Di Giovanni et al., 2023), and oversmoothing, the collapse of the Dirichlet energy with depth (Cai & Wang, 2020), are the two depth-related pathologies of message passing. HMH (Hossen, 2026) mitigates both under filter-gain amplification assumptions; Section 4.10 derives analogues that avoid such amplification assumptions from the exact isometry of our channel bank and the distance independence of pooling-based coupling.

3 PRELIMINARIES

Binary Haar graph framelets and PEGFAN. Given a graph $G = (V, E)$ and a K-level partition tree $\mathcal{P}_K$, PEGFAN (Li et al., 2024) recursively defines scaling and framelet vectors that form a tight frame, and projects the features as $F_j(M) = \Psi_j^\top \Psi_j M$, where Ψ_j stacks the framelet vectors of scale j. Two properties are central for us: F_j is the *orthogonal projector* onto the within-cluster detail space of scale j (the framelet vectors require choosing an orthonormal basis per cluster, the operator $\Psi_j^\top \Psi_j$ does not), and PEGFAN's equivariance guarantee takes the tree as an input (Li et al., 2024, Thm. 4, Prop. 2): the transform and the network are permutation equivariant for a *given* $\mathcal{P}_K$, but the generation of $\mathcal{P}_K$ itself is not covered. Our Select stage closes this gap (Section 4.2), and Section 4.3 exploits the first property to construct the framelet operators without ever choosing bases.

Cellular sheaves and neural sheaf diffusion. A cellular sheaf $\mathcal{F}$ on G assigns a stalk $\mathcal{F}(v) \cong \mathbb{R}^d$ to every node (and edge) and restriction maps $F_{u\to e}, F_{v\to e} \in \mathbb{R}^{d\times d}$ to every incident node–edge pair. The sheaf Laplacian of the 0-cochain space $C^0(G;\mathcal{F})$ has diagonal blocks $D_u = \sum_{e\ni u} F_{u\to e}^\top F_{u\to e}$ and off-diagonal blocks $-F_{u\to e}^\top F_{v\to e}$. Its normalized form $\Delta_{\mathcal{F}} = D^{-1/2} L_{\mathcal{F}} D^{-1/2}$ drives neural sheaf diffusion (Bodnar et al., 2022), which can amplify contrast across heterophilous edges instead of smoothing it away.

Invariant partitions and granular-ball splitting. Hevapathige & Wang (2026) assign each node a structural id from the k-core shell decomposition: nodes with core number in $\{0, 1\}$ form one partition, and each higher shell is split at the minimum in-shell degree. The id depends only on isomorphism invariants, so the partition is permutation invariant by construction, in $O(N + E)$ time; the core numbers also serve as our centrality measure for ball centers and seeds, in place of the PageRank scores used by HGBG. HGBG (Ni et al., 2026) itself coarsens a graph top-down:

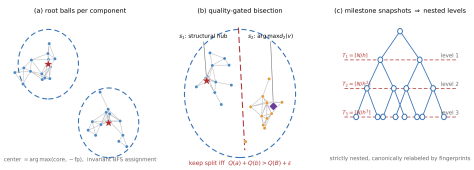

Figure 1: The GBSC divisive construction. One root ball per connected component; a ball is bisected around the structural-hub seed s_1 and the topology–feature seed s_2 of Eq. 1, and the split is kept only if the quality oracle of Eq. 2 improves by the margin ε (Eq. 3). Dashed lines: milestone snapshots of the splitting frontier at the targets $T_\ell = \lceil N/h^\ell \rceil$, which become the strictly nested coarsening levels.

starting from one granular ball per connected component, it recursively splits the coarsest ball into two sub-balls around two seeds and accepts the split iff the children's combined quality exceeds the parent's. Our Select stage follows this divisive skeleton and tightens its acceptance rule by a margin ε (Eq. 3). Throughout, $\hat{A} = \bar{D}^{-1/2} A \bar{D}^{-1/2}$ denotes the symmetrically normalized adjacency ($\bar{D} = \mathrm{diag}(A\mathbf{1})$) and $\hat{A}_I = (\bar{D} + I)^{-1/2}(A + I)(\bar{D} + I)^{-1/2}$ its self-looped counterpart.

4 The SHEAR Architecture

4.1 Overview

SHEAR combines a GPNN-guided *granular-ball splitting clustering* (GBSC) with basis-free Haar framelet operators and end-to-end sheaf diffusion. Given (G, X) with $X \in \mathbb{R}^{N \times f}$, the pipeline has three stages: (i) build the hierarchical coarsening module, comprising Select (Section 4.2), Reduce (Section 4.3), and Connect (Section 4.4); (ii) precompute the fixed multi-hop propagation channels (channel group (a) of Section 4.7); (iii) train the SHEAR network (Sections 4.5–4.8), in which all restriction maps are generated end-to-end at every forward pass. The hierarchy uses up to $L = 3$ coarsening levels; level ℓ has N_ℓ nodes and m_ℓ feature channels ($m_0 = m$), with a stalk dimension d shared by all levels. We write $S^{(\ell)} \in \{0,1\}^{N_{\ell-1} \times N_\ell}$ for the level-ℓ assignment matrix, and lift node-space operators to the stalk space by tensoring with I_d. Sections 4.9 and 4.10 then state the two theoretical guarantees of the architecture: end-to-end permutation equivariance and channel-level robustness to oversquashing and oversmoothing. Figure 2 (Appendix B) summarizes the architecture.

4.2 Select: granular-ball hierarchical clustering

The Select stage builds the partition tree *top-down* by recursive quality-gated bisection of granular balls. All decisions are deterministic functions of isomorphism invariants, so the resulting hierarchy is permutation invariant (Appendix A). Figure 1 illustrates the construction.

(a) Invariant node fingerprints. Every node carries a fingerprint $\mathrm{fp}(u)$: the position of the tuple (two-round Weisfeiler–Lehman (WL) color, GPNN structural id pid_u, degree, $\mathrm{crc32}$ hash of the feature row x_u) in the lexicographic ordering of all node tuples, every component of which is an isomorphism invariant. Every tie below is broken by fp; only residual ties between fingerprint-identical (automorphic twin) nodes fall back to the raw node id, which leaves downstream predictions unaffected (Remark 3).

(b) Divisive root balls. We open one granular ball per connected component of G, centered at the node attaining $\max(\mathrm{core}(u), -\mathrm{fp}(u))$, and assign every node to its center by an invariant multi-source breadth-first search (BFS) whose heap order (shortest-path distance, center fingerprint, node fingerprint) makes the assignment the unique minimizer of an invariant key.

(c) Bisection with a topology–feature seed rule. A max-heap holds the current balls (largest first, fingerprint tie-break). To split a ball B, seed s_1 is the structural hub $\arg\max_{v \in B}(\mathrm{core}(v), -\mathrm{fp}(v))$, and seed s_2 maximizes the combined seed distance d_2 (the maximization is lexicographic in the

key $(\mathrm{round}(d_2(v), 12), -\mathrm{fp}(v))$, the 12-decimal rounding absorbing floating-point reduction noise, Lemma 5)

$$d_2(v) \;=\; \alpha \underbrace{\frac{1-\cos(x_v, x_{s_1})}{2}}_{d_{\mathrm{feat}}} + \beta \underbrace{\big(1 - S_{\mathrm{jac2}}(v, s_1)\big)}_{d_{\mathrm{struct}}} + (1-\alpha-\beta) \underbrace{\mathbf{1}[\mathrm{pid}_v \neq \mathrm{pid}_{s_1}]}_{d_{\mathrm{role}}}, \qquad (1)$$

where $\alpha, \beta \geq 0$ with $\alpha+\beta \leq 1$ weight the feature and the Jaccard-structure term, and $S_{\mathrm{jac2}}(u, v) = |N_2(u) \cap N_2(v)| \,/\, |N_2(u) \cup N_2(v)|$ is the two-hop Jaccard structural similarity of HMH (Hossen, 2026, Sec. 5.1); the cosine term is scale-invariant (computed on row-normalized features), and the Jaccard term is evaluated only when $\beta > 0$, between the current ball's members and s_1 with lazily built, cached two-hop neighborhoods; with $\beta = 0$ (default) the structural-role term keeps weight $1-\alpha$. The ball is then bisected by a two-source BFS from (s_1, s_2) with the same invariant heap order; members unreachable from both seeds are assigned by the sign of the affinity gap $\langle x_v, x_{s_1}\rangle - \langle x_v, x_{s_2}\rangle$, so every node is assigned to exactly one child. If a *forced* split (see (d)) produces a severely unbalanced cut (smaller side $< |B|/4$), a deterministic balanced fallback halves the members by the affinity gap $\langle x_v, x_{s_1}\rangle - \langle x_v, x_{s_2}\rangle$ with fingerprint tie-breaking, so forced splits shrink oversized balls geometrically.

(d) Quality-gated acceptance with a hard size cap. Each candidate split is scored by a feature–structure–community quality oracle

$$Q(B) \;=\; \omega_1 \, Q_{\mathrm{feat}}(B) + \omega_2 \, Q_{\mathrm{str}}(B) + \omega_3 \, Q_{\mathrm{comm}}(B), \qquad (2)$$

where $\omega_1, \omega_2, \omega_3 \geq 0$ are fixed oracle weights, Q_{feat} is the feature compactness (mean cosine similarity to the ball centroid, with both the centroid and the mean accumulated in canonical fingerprint order, so the value is independent of the node numbering, Appendix A), Q_{str} is the structural-role purity (one minus the entropy of the pid histogram normalized by $\log K$, K the number of structural-role ids), and Q_{comm} is the modularity contribution $e_{\mathrm{in}}/|E| - \big(\mathrm{vol}(B)/(2|E|)\big)^2$, with e_{in} the number of intra-ball edges counted once and $\mathrm{vol}(B)$ the sum of node degrees in B (the standard per-community modularity term, which vanishes for $B = V$). A split of B into (a, b) is accepted iff

$$Q(a) + Q(b) \;>\; Q(B) + \varepsilon, \qquad (3)$$

the acceptance rule of HGBG (Ni et al., 2026, Eq. 16) tightened by the margin ε of Section 3. To keep the hierarchy memory-bounded on graphs where the gate would otherwise preserve a giant ball, a *hard cap* overrides the oracle: any ball larger than κh (cap multiplier κ acting on the cluster-size bound h of (e); default off on small graphs and $\kappa = 8$ on large ones) is always split when splittable.

(e) Milestone level extraction and canonical relabeling. Rather than prescribing cluster sizes per level, we fix level *targets* $T_\ell = \lceil N/h^\ell \rceil$ and snapshot the current frontier whenever the ball count crosses a target *and* the cap of (d) is satisfied; the snapshot at T_ℓ becomes level ℓ. Since splits only refine balls, the extracted levels are strictly nested, and level ℓ has at least $\lceil N/h^\ell \rceil$ clusters, i.e. cluster sizes are bounded by h^ℓ on average at level ℓ and by κh individually. Each snapshot is canonically relabeled in fingerprint order, so cluster ids are independent of the splitting history. The coarse adjacency is $A^{(\ell)} = S^{(\ell)\top} A^{(\ell-1)} S^{(\ell)}$ (diagonal removed) and features are mean-pooled; the recursion stops when $T_\ell \leq h$ and is in any case truncated at $L = 3$ levels, while the splitting loop itself terminates when the heap is exhausted or a safety split budget of $10N$ accepted splits is reached (not binding on any of our benchmarks, Appendix A). The output is the list of partitions, fine to coarse, together with the pool/prolong operators

$$\mathrm{pool}_\ell = \big(D_\ell^{-1} S^{(\ell)\top}\big) \otimes I_d \;\in\; \mathbb{R}^{N_\ell d \times N_{\ell-1} d}, \qquad \mathrm{prolong}_\ell = S^{(\ell)} \otimes I_d \;\in\; \mathbb{R}^{N_{\ell-1} d \times N_\ell d}, \qquad (4)$$

where $D_\ell = \mathrm{diag}(|C_1|, \ldots, |C_{N_\ell}|)$ collects the level-ℓ cluster sizes.

4.3 REDUCE: BASIS-FREE BINARY HAAR FRAMELET OPERATORS

At level ℓ, the level-ℓ clusters partition the level-$(\ell-1)$ nodes, so the signals that are constant on every level-ℓ cluster (on each stalk coordinate) form a subspace V_ℓ of the level-$(\ell-1)$ stalk signal space $\mathbb{R}^{N_{\ell-1} d}$. The binary Haar framelet decomposition of Li et al. (2024) is the orthogonal splitting

$\mathbb{R}^{N_{\ell-1}d} = V_\ell \oplus \mathcal{W}_\ell$ with detail space $\mathcal{W}_\ell = V_\ell^\perp$. Instead of assembling framelet vectors cluster by cluster, we construct the two projection operators directly from the pool/prolong matrices of Eq. 4:

$$\mathcal{S}_\ell \;=\; \mathrm{prolong}_\ell \,\mathrm{pool}_\ell \;=\; \big(S^{(\ell)} D_\ell^{-1} S^{(\ell)\top}\big) \otimes I_d, \qquad \mathcal{W}_\ell \;=\; I_{N_{\ell-1}d} - \mathcal{S}_\ell. \tag{5}$$

$\mathcal{S}_\ell$ is the orthogonal projector onto V_ℓ (cluster-wise means on every stalk coordinate) and $\mathcal{W}_\ell$ the orthogonal projector onto $\mathcal{W}_\ell$. This projector form coincides with PEGFAN's $\Psi_j^\top \Psi_j$ while bypassing the per-cluster orthonormal bases entirely. Its properties make it an exact, unambiguous frequency decomposition:

Proposition 1 (Properties of the framelet operators)**.** *For every level ℓ:* (i) $\mathcal{S}_\ell^2 = \mathcal{S}_\ell$, $\mathcal{S}_\ell^\top = \mathcal{S}_\ell$, *and the same for $\mathcal{W}_\ell$ (orthogonal projectors);* (ii) $\mathcal{S}_\ell + \mathcal{W}_\ell = I$ *and* $\mathcal{S}_\ell \mathcal{W}_\ell = 0$ *(exact telescoping, i.e. perfect reconstruction with two bands);* (iii) *both operators are block-diagonal over clusters and over stalk coordinates, with sparsity $\sum_a |C_a|^2 d$;* (iv) *both are independent of the restriction maps, hence shared by all data splits;* (v) *both co-vary with node permutations and are exactly invariant under cluster relabeling (Appendix A).*

A three-node example, making concrete how the projector form retains the Haar semantics (cluster mean versus zero-mean detail) while eliminating the basis choice of the framelet vectors, is given in Appendix A (Remark 2). Single-node clusters automatically follow the Haar rule: $\mathcal{S}_\ell$ is the identity on their stalks and $\mathcal{W}_\ell$ vanishes there. Since $\mathrm{pool}_\ell \,\mathrm{prolong}_\ell = I_{N_\ell d}$, the composition with pooling lands on coarse cluster means, and the prolongation (Section 4.7) lifts signals back through the same adjoint pair.

4.4 Connect: Galerkin coarse sheaf in closed form

Each coarsened level needs a sheaf Laplacian of its own. With copy-prolongation $R^{(\ell)} = S^{(\ell)} \otimes I_d$ ($= \mathrm{prolong}_\ell$ of Eq. 4), the Galerkin recursion $L_{\mathrm{Gal}} = R^* L_{\mathcal{F}} R$ of Naddeo et al. (2026) telescopes, so the coarse sheaf blocks can be written directly over the *original* edges:

$$B_{ab}^{(\ell)} = \sum_{(u,v):\, u\in a,\, v\in b} F_{u\to e}^\top F_{v\to e}, \qquad D_a^{(\ell)} = \sum_{u\in a} D_u - \sum_{u,v\in a} \big(F_{u\to e}^\top F_{v\to e} + F_{v\to e}^\top F_{u\to e}\big), \tag{6}$$

with $D^{-1/2}$ from a blockwise eigendecomposition of $D^{(\ell)} + 10^{-3} I$. The partition tree, pool/prolong operators, and framelet projectors depend only on the dataset and are cached; the sheaf blocks, in contrast, are rebuilt from the restriction maps at every forward pass, as detailed next.

4.5 End-to-end restriction maps

The Galerkin blocks of Eq. 6 depend on the restriction maps, which a priori must be available before the coarse sheaves can be built. Rather than pretraining a sheaf encoder and freezing its maps, the route taken by prior sheaf architectures, we learn the maps with the rest of the model. A shared SheafMLP maps the concatenated current features of the two endpoints of every original edge to $F_{u\to e}, F_{v\to e} \in \mathbb{R}^{d\times d}$; the maps are regenerated once per forward pass from the lifted level-0 features. The level-ℓ Galerkin blocks of Eq. 6 are then assembled on the fly: since every block is a sum over a fixed subset of original-edge Gram/cross terms ($F_{u\to e}^\top F_{u\to e}$, $F_{v\to e}^\top F_{v\to e}$, $F_{u\to e}^\top F_{v\to e}$), all levels are rebuilt by differentiable `index_add` reductions through precomputed aggregation indices, in time linear in the number of original edges.

A frozen-encoder variant (pretrain a sheaf encoder on the split's train/validation masks for 300 epochs, then freeze the maps) is kept as an ablation (Section F.2); the framelet operators of Section 4.3, being map-independent (Proposition 1(iv)), are identical in both variants.

4.6 Multi-scale sheaf diffusion

Every level runs neural sheaf diffusion on the lifted signal (level 0) or on the cascade output of Eq. 46 (levels $\ell \geq 1$), with T_0 steps on the original graph and T_c per coarse level:

$$H^{(\ell,t+1)} = H^{(\ell,t)} - \tau_{\ell,t} \,\tanh\Big(\Delta_{\mathcal{F}^{(\ell)}} \,(I_{N_\ell} \otimes W_1^{(\ell,t)})\, H^{(\ell,t)}\, W_2^{(\ell,t)}\Big), \tag{7}$$

where $\Delta_{\mathcal{F}^{(\ell)}} = D^{-1/2} L_{\mathcal{F}^{(\ell)}} D^{-1/2}$ is the *genuine* coarse sheaf Laplacian built from the current restriction maps (crossing-edge blocks included), applied edge-wise without densification, the entrywise $\tanh$ follows the standard activation of diffusion-based graph architectures (Rusch et al.,

2022; Gravina et al., 2023) and bounds each step's increment by the learned step size, $W_1 \in \mathbb{R}^{d\times d}$ mixes the stalk coordinates, $W_2 \in \mathbb{R}^{m_\ell \times m_\ell}$ mixes the feature channels, and the step sizes $\tau_{\ell,t}$ are learned (init. 0.1).

4.7 Analysis cascade, literal prolongation chains, and channels

The analysis cascade coarsens the lifted signal $H = \mathrm{lift}(X) := XW_{\mathrm{lift}}$ (a row-wise linear map to m channels, unfolded to the stalk space) level by level, mixing the two framelet bands with learnable gains $\theta_\ell = \mathrm{softmax}(\Theta_\ell) \in \Delta^1$ and carrying only the pooled cluster means to the next coarser level (Eq. 46, Appendix B); after diffusion, each level is brought back to the original resolution by a *literal* prolongation chain with framelet synthesis gains $\eta_k = \mathrm{softmax}(E_k)$ and per-cluster back-projections (Eq. 47). The fusion consumes five groups of channels: (a) the propagation bank $\{X, \hat{A}^k X, \hat{A}_I^k X\}_{k\le k_{\max}}$; (b) the per-level detail channels $\Xi_\ell \hat{A}_I X$ ($\ell = 1, \dots, L$) and the coarsest scaling channel $\Upsilon_L \hat{A}_I X$, built from the projectors Υ_ℓ, Ξ_ℓ of Section 4.10; (c) the diffused features $Z^{(0)}$ on the original graph; (d) the per-level prolongation outputs $Z^{(\ell)}$; and (e) a plain low-pass channel Z_{LP} (the matching high-pass channel vanishes identically, $\mathrm{pool}_\ell \mathcal{W}_\ell = 0$); the full definitions, the dilated stride-s and parallel multi-tree variants of group (b), and the channel bank $\mathcal{Z}$ of Eq. 48 ($n_C = 2k_{\max} + 2L + 4$ channels with a single detail tree) are deferred to Appendix B.

4.8 Attention fusion

It remains to aggregate the channel bank $\mathcal{Z} = \{Z_k\}_{k=1}^{n_C}$ of Eq. 48 into node predictions. Following the multi-channel fusion of Li et al. (2024); Maurya et al. (2022), each channel is projected by its own linear layer, row-normalized, weighted by a global scalar attention, and concatenated:

$$H_1 \;=\; \big\|_{k=1}^{n_C} \gamma_k \mathcal{N}\big(Z_k W_k^{(1)}\big), \qquad \hat{Y} \;=\; \mathrm{softmax}\big(\mathrm{Dropout}(\mathrm{ReLU}(H_1))\, W^{(2)}\big), \tag{8}$$

where $\|$ denotes concatenation, $\mathcal{N}(\cdot)$ is row-wise ℓ_2 normalization, $\gamma = \mathrm{softmax}(\mathbf{a}) \in \Delta^{n_C-1}$ has learnable logits $\mathbf{a} \in \mathbb{R}^{n_C}$, $W_k^{(1)} \in \mathbb{R}^{d_k \times n_h}$ with d_k the width of channel k and hidden width n_h, and $W^{(2)} \in \mathbb{R}^{n_C n_h \times c}$ for c classes. The per-channel maps $Z_k \mapsto \mathcal{N}(Z_k W_k^{(1)})$ and the final linear layer are the $\mathrm{fc1}_k$ and fc2 referred to in Section 4.10 and Appendix C. A node-level attention variant ($s_{k,i} = \mathrm{MLP}_{\mathrm{att}}([Z_k]_i)$, softmax over k per node) is available but disabled in all reported experiments.

4.9 Permutation equivariance

Theorem 1 (Permutation equivariance, informal; full statement and proof in Appendix A, Theorem 4)**.** *Let π be a node permutation. Under a general-position assumption on the invariant keys, the node predictions of SHEAR satisfy $\hat{Y}(\pi G, \pi X) = \pi\, \hat{Y}(G, X)$, end-to-end, in exact arithmetic, with the clustering stage included and the restriction maps regenerated by the shared SheafMLP inside every forward pass. Outside the general-position set, the only residual failures are exchanges of automorphic-twin nodes, which share every invariant key; any equivariant downstream map gives such twins identical outputs, so the node predictions are unaffected (Appendix A, Remark 3).*

The full induction over the Select, Reduce, and Connect stages is given in Appendix A, which also shows that equivariance is preserved by training: for synchronized batch order and random-number streams the training trajectories on (G, X) and $(\pi G, \pi X)$ coincide (Corollary 1).

4.10 Robustness to oversquashing and oversmoothing

The hierarchy behind the equivariance guarantee also gives structural control over oversquashing and oversmoothing. We state both results informally and defer the formal statements and proofs to Appendix C. Write $S_{1:\ell} = S^{(1)} \cdots S^{(\ell)}$ for the composed level-0 $\to \ell$ assignment, $\Upsilon_\ell = S_{1:\ell}(S_{1:\ell}^\top S_{1:\ell})^{-1} S_{1:\ell}^\top$ for the orthogonal projector onto level-ℓ cluster-constant signals ($\Upsilon_0 = I$), $\Xi_\ell = \Upsilon_{\ell-1} - \Upsilon_\ell$ for the level-ℓ detail projector, s_ℓ for the largest level-ℓ cluster size, and $d_{\max}$ for the maximum degree.

Theorem 2 (Oversquashing mitigation, informal; full statement and proof in Appendix C.1, Theorems 5 and 6)**.** *For any two nodes u, v co-clustered at level ℓ, the Jacobian block of the level-ℓ scaling channel satisfies $\|\partial(\Upsilon_\ell \hat{A}_I X)_u/\partial X_v\| \geq 1/(s_\ell(1+d_{\max}))$, independent of the graph distance $d_G(u,v)$ and of any depth parameter; with the hard size cap, $s_\ell \leq \kappa h$ at every level. For arbitrary pairs, the pooling–coarse-diffusion–prolongation pathway has length $2L+d_{G_L}(C_L(u), C_L(v))$ with $L = O(\log_h N)$ and residual (identity-branch) gains bounded away from zero, so node influence is governed by the coarse distance d_{G_L} rather than $d_G(u,v)$. The architecture thus structurally avoids the oversquashing regime $\|J_{uv}\| \leq \sigma^{d_G(u,v)}$, $\sigma < 1$, of Topping et al. (2022); Di Giovanni et al. (2023) (Appendix C.1).*

Theorem 3 (Oversmoothing mitigation, informal; full statement and proof in Appendix C.2, Theorems 7 and 8)**.** (i) *The detail/scaling channel bank is an exact orthogonal decomposition of the identity, $\Upsilon_L + \sum_{\ell=1}^{L} \Xi_\ell = I$, hence an isometry: any signal contrast that is not constant on the coarsest clusters survives in at least one channel with at least a $1/L$ fraction of its energy (Corollary 3).* (ii) *Every smoothing operator in the network is applied a fixed, architecture-prescribed number of times, so the exponential-in-depth Dirichlet-energy collapse of deep MPNNs (Cai & Wang, 2020; Bodnar et al., 2022) cannot occur; with learnable restriction maps the sheaf kernel need not contain the constant vector at all.* (iii) *The concatenating fusion with strictly positive softmax attention preserves class-mean contrast: a channel keeping contrast δ between two class means contributes $\|\mu_1 - \mu_2\| \geq \gamma_c\, \sigma_{\min}(W_{\mathrm{fc1},c})\, \delta > 0$ to the fused representation, unless the contrast coordinates are clipped by the ReLU.*

The detailed comparison with Hossen (2026) and the formal scope remarks—including the architectural (no-depth-axis) nature of part *(ii)* and the coverage of the isometry being limited to the framelet pathway—are given in Appendix C.

5 Experiments

The design choices behind the heterophilous results, the leakage analysis, and the limitations are discussed in Sections G and F.4.

5.1 Setup and benchmark comparison

Datasets, protocols, and baselines. We evaluate on thirteen node-classification benchmarks from two protocols, using throughout the same data versions and splits as the source papers, so that within each of Tables 1 and 2 all rows are directly comparable; Table 1 is reported below and Table 2 is deferred to Appendix D. Table 1 follows Hossen (2026, Table 1): the Platonov benchmarks Minesweeper, Tolokers, Roman-empire, and Amazon-ratings (Platonov et al., 2023) (ten official $50/25/25$ splits; ROC-AUC for the two binary datasets, accuracy otherwise) and the *filtered* Squirrel and Chameleon of Platonov et al. (2023) (ten official $60/20/20$ splits; an extended parallel detail bank, Section G and Table 3), against BernNet (He et al., 2021), ChebNetII (He et al., 2022), GCNII (Chen et al., 2020), JacobiConv (Wang & Zhang, 2022), OptBasisGNN (Guo & Wei, 2023), TFE-GNN (Huang et al., 2024), M2M-GNN (Liang et al., 2024), SLOG (Xu et al., 2024), and HMH (Hossen, 2026). Table 2 follows Li et al. (2024, Table III): Cora, Citeseer, and Pubmed (Sen et al., 2008), Film (Tang et al., 2009), and Texas, Cornell, and Wisconsin (Pei et al., 2020) (ten official Geom-GCN $48/32/20$ splits), against MixHop (Abu-El-Haija et al., 2019), Geom-GCN (Pei et al., 2020), GCNII (Chen et al., 2020), H_2GCN (Zhu et al., 2020), WRGAT (Suresh et al., 2021), GPRGNN (Chien et al., 2021), FSGNN (Maurya et al., 2022), and PEGFAN (Li et al., 2024); baseline numbers are quoted from the two source papers. The studies of Sections F.1–F.2 and G use the unfiltered Geom-GCN versions of Texas, Chameleon, and Squirrel under the same $48/32/20$ splits. All tasks are transductive, trained by cross-entropy on the train-mask nodes; the training settings, the selection protocol, and the per-dataset configurations are deferred to Appendix E (Table 3). We emphasize that the multi-resolution structure is a deterministic function of (G, X) alone: on every benchmark the partition tree, the framelet operators, and the coarse sheaves are identical across the ten splits, and only the training dynamics vary.

Results. SHEAR sets the best result on five columns: Minesweeper (88.54, two points above TFE-GNN's 86.85), Texas (87.30), Roman-empire (89.04, a margin of 13 points over the strongest

Table 1: Test performance (%, mean ± std) on the six datasets of Hossen (2026, Table 1): Minesweeper, Tolokers, Roman-empire, and Amazon-ratings under the ten official $50/25/25$ splits of Platonov et al. (2023), and Squirrel and Chameleon in the filtered versions of Platonov et al. (2023) under their ten official $60/20/20$ splits. Numbers of the quoted rows are taken from Hossen (2026) (UniFilter and MTGCN omitted: their rows contain a malformed and a misaligned entry, respectively, in the source table). Minesweeper and Tolokers use ROC-AUC, all other columns accuracy. —: not reported under this protocol. **Bold**: best; underline: second best.

	Heterophilous					
	Minesweeper	Tolokers	Roman-empire	Squirrel	Chameleon	Amazon-ratings
BernNet	78.80±0.00	77.20±0.70	72.74±0.29	36.10±0.70	37.80±0.70	39.80±0.20
ChebNetII	82.30±0.10	78.30±0.30	73.74±0.29	35.00±0.40	33.50±0.50	39.30±0.10
GCNII	78.80±0.00	77.80±0.10	74.74±0.29	31.40±0.80	38.00±1.00	42.90±0.20
JacobiConv	78.80±0.00	70.40±0.57	70.74±0.29	22.10±1.70	30.90±1.50	35.50±1.00
OptBasisGNN	84.80±2.40	79.85±1.63	68.34±0.30	28.66±1.10	39.70±2.40	43.35±1.34
TFE-GNN	86.85±0.33	79.23±1.13	74.20±0.35	31.47±1.15	41.16±1.41	40.16±1.81
M2M-GNN	82.25±1.43	80.85±1.13	68.36±3.23	33.60±1.70	32.20±2.30	47.18±0.60
SLOG	84.40±0.80	81.00±0.60	—	38.20±1.20	43.02±2.20	45.60±0.60
HMH	87.40±0.80	**84.56±1.90**	76.10±0.70	39.50±0.20	41.20±0.20	**48.64±0.55**
SHEAR (ours)	**88.54±0.69**	82.20±1.04	**89.04±0.46**	**40.54±1.25**	**45.28±2.22**	46.47±0.52

baseline), the filtered Squirrel (40.54, a point above HMH's 39.50), and the filtered Chameleon (45.28, two points above SLOG's 43.02 and four above HMH's 41.20; the filtered Chameleon graph has only 890 nodes and realizes a two-level hierarchy, Table 4). It is second best on Tolokers, and on Wisconsin it is statistically tied with the column leaders (87.25 ± 3.07 versus 87.84 ± 3.37 for FSGNN and 87.65 ± 4.02 for PEGFAN). On Amazon-ratings it ranks third (46.47, behind HMH's 48.64 and M2M-GNN's 47.18). Its remaining deficits are concentrated on the three homophilous citation graphs, where full-graph spectral-propagation models remain stronger; a full-graph extension of the multi-scale bank targeting this regime is discussed in Section G. Overall, SHEAR is best or second best on six of the thirteen columns and on six of the ten heterophilous ones. On the unfiltered Geom-GCN versions of Squirrel/Chameleon, SHEAR reaches 71.52 and 76.03; the large gap to the filtered numbers mirrors the version sensitivity documented by Platonov et al. (2023) (Table 8 and the leakage analysis of Appendix F.4).

AI Use Statement

In this work, we used generative AI tools (Kimi) to assist with language editing and readability improvement. We did not use AI tools to generate scientific content, analyze data, or produce experimental results. All AI-assisted text was reviewed and edited by the authors, and we take full responsibility for the content of the published article.

6 Conclusion

We presented SHEAR, a hierarchical sheaf network whose entire multi-resolution structure is generated by a permutation-invariant granular-ball clustering that couples GPNN core-degree partitioning with HGBG-style quality-gated bisection, replacing the Ward clustering of PEGFAN at the one point where its equivariance guarantee does not reach. On this hierarchy we build basis-free binary Haar framelet operators with learnable band gains, drive multi-scale diffusion with Galerkin coarse sheaf Laplacians whose restriction maps are generated end-to-end, and fuse all channels with scalar channel attention; the exact isometry of the channel bank and the distance independence of pooling-based coupling yield channel-level guarantees against oversquashing and oversmoothing without filter-gain amplification assumptions. On thirteen benchmarks from the PEGFAN and HMH suites under the same data versions and splits as the baselines, SHEAR is best or second best on six of the thirteen columns, with first places on Minesweeper, Texas, Roman-empire, and the filtered Squirrel and Chameleon.

REFERENCES

Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, and Aram Galstyan. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *International Conference on Machine Learning (ICML)*, 2019.

Cristian Bodnar, Francesco Di Giovanni, Benjamin Chamberlain, Pietro Liò, and Michael Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in GNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1725–1735, 2020.

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

Francesco Di Giovanni, T. Konstantin Rusch, Michael M. Bronstein, Andreea Deac, Marc Lackenby, Siddhartha Mishra, and Petar Veličković. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In *International Conference on Machine Learning (ICML)*, 2023.

Alessio Gravina, Francesco Di Giovanni, Benjamin P. Chamberlain, Claudio Gallicchio, and Michael M. Bronstein. Anti-symmetric DGN: a stable architecture for Deep Graph Networks. In *International Conference on Learning Representations (ICLR)*, 2023.

Yuhe Guo and Zhewei Wei. Graph neural networks with learnable and optimal polynomial bases. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12077–12097, 2023.

Jakob Hansen and Thomas Gebhart. Toward sheaf-theoretic neural networks. *arXiv preprint arXiv:1910.14143*, 2019.

Mingguo He, Zhewei Wei, Hongteng Xu, et al. Bernnet: Learning arbitrary graph spectral filters via bernstein approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Mingguo He, Zhewei Wei, and Ji-Rong Wen. Convolutional neural networks on graphs with chebyshev approximation, revisited. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Asela Hevapathige and Qing Wang. How graph neural networks capture structural interactions? *Neural Networks*, 2026. Also arXiv:2312.08671.

MD Sazzad Hossen. Hierarchical multi-scale graph neural networks: Scalable heterophilous learning with oversmoothing and oversquashing mitigation. *arXiv preprint arXiv:2605.10975*, 2026.

Zhipeng Huang, Yizhou Wang, Chaozhuo Li, and Pietro Liò. Unifying homophily and heterophily for spectral graph neural networks via triple filter ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jianfei Li, Ruigang Zheng, Han Feng, Ming Li, and Xiaosheng Zhuang. Permutation equivariant graph framelets for heterophilous graph learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. arXiv:2306.04265.

Ming Li, Yujie Fang, Dongrui Shen, Han Feng, Xiaosheng Zhuang, Kelin Xia, and Pietro Lio. High-pass matters: Theoretical insights and sheaflet-based design for hypergraph neural networks. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence*, pp. 23039–23046, 2026.

Langzhang Liang, Zixiang Zeng, Zhaoliang Li, Xiang Yang, Kele Liu, Lin Wang, Fuli Wu, Zhiqiang Wang, and Peng Cui. Sign is not a remedy: Multiset-to-multiset message passing for learning on heterophilic graphs. *arXiv preprint arXiv:2405.20652*, 2024.

Sunil Kumar Maurya, Xin Liu, and Tsuyoshi Murata. Towards simplifying and improving graph neural networks for heterophilous graphs. *arXiv preprint arXiv:2203.09268*, 2022.

Dionisia Naddeo, Carlo Abate, Pietro Liò, Nicola Toschi, and Filippo Maria Bianchi. Hierarchical pooling for sheaf neural networks. *arXiv preprint arXiv:2606.20932*, 2026.

Jinyuan Ni, Long Chen, Gu Tang, Ning Yu, Wenyue Tang, Xiaoyin Yi, and Shuyin Xia. Hierarchical granular-ball graph pooling via feature-structure coupling. *Information Sciences*, 748:123501, 2026.

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

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

T. Konstantin Rusch, Benjamin P. Chamberlain, James Rowbottom, Siddhartha Mishra, and Michael M. Bronstein. Graph-coupled oscillator networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18888–18909, 2022.

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

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

Junshu Sun, Shuhui Wang, Xinzhe Han, Zhe Xue, and Qingming Huang. All in a row: Compressed convolution networks for graphs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33061–33076, 2023.

Susheel Suresh, Vinith Budde, Jennifer Neville, Pan Li, and Jianzhu Ma. Breaking the limit of graph neural networks by improving the assortativity of graphs with local mixing patterns. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2021.

Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.

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

Xiyuan Wang and Muhan Zhang. How powerful are spectral graph neural networks. In *International Conference on Machine Learning (ICML)*, 2022.

Zhiqian Xu, Yiming Yan, Xinyi Wang, Zekun Xu, Zhiyuan Zeng, Tarek Abdelzaher, Jiawei Han, and Hanghang Tong. SLOG: An inductive spectral graph neural network beyond polynomial filter. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Xin Zheng, Yixin Liu, Shirui Pan, Miao Zhang, Di Jin, and Philip S Yu. Graph neural networks for graphs with heterophily: A survey. *arXiv preprint arXiv:2202.07082*, 2022.

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

AI USE STATEMENT

In this work, we used generative AI tools (Kimi) to assist with language editing and readability improvement. We did not use AI tools to generate scientific content, analyze data, or produce experimental results. All AI-assisted text was reviewed and edited by the authors, and we take full responsibility for the content of the published article.

A COMPLETE PROOF OF PERMUTATION EQUIVARIANCE

We give the full proof of the permutation-equivariance theorem stated informally in Section 4.9 (Theorem 1). The proof covers the *entire* pipeline, including the clustering stage that generates the partition tree, the step that lies outside the equivariance theorem of PEGFAN (Li et al., 2024, Sec. III-E, Theorem 4), whose framelet transform and network are themselves permutation equivariant once the tree is given. Throughout, the restriction maps are generated end-to-end inside every forward pass by the shared SheafMLP (Section 4.5). In contrast to spectral (sheaflet) frequency decompositions, the binary Haar framelet operators are *orthogonal projectors built from the assignment matrices alone*, so no eigendecomposition enters the Reduce stage at all (Remark 4).

A.1 SETUP, NOTATION, AND STANDING ASSUMPTIONS

Permutation action. Let $G = (V, E)$, $N = |V|$, with symmetric loopless adjacency $A \in \{0,1\}^{N\times N}$ and features $X \in \mathbb{R}^{N\times f}$. For $\pi \in S_N$ let $P_\pi \in \{0,1\}^{N\times N}$ be its permutation matrix, $(P_\pi)_{\pi i,i} = 1$. The *permuted input* is

$$\pi G = (V, \pi E), \qquad A^\pi = P_\pi A P_\pi^\top, \qquad X^\pi = P_\pi X, \tag{9}$$

i.e. $(A^\pi)_{\pi u,\pi v} = A_{uv}$ and $x^\pi_{\pi u} = x_u$. Edges are unordered pairs; write $\pi e = (\pi u, \pi v)$ for $e = (u, v)$.

Stalk signals. A level-ℓ signal with stalk dimension d and m_ℓ channels is $H \in \mathbb{R}^{N_\ell\times d\times m_\ell}$, unfolded node-major as $\vec{H} \in \mathbb{R}^{N_\ell d\times m_\ell}$. The induced action on the stalk space is

$$\Pi_\pi \;=\; P_\pi \otimes I_d \;\in\; \{0,1\}^{Nd\times Nd}, \qquad \Pi_\pi^\top \Pi_\pi = I, \tag{10}$$

and likewise at coarse levels with the corresponding node permutation.

Groupings and canonical labeling. A *grouping* of V is a partition into unlabeled blocks $\mathcal{C} = \{C_1, \dots, C_K\}$. A *labeled partition* is an assignment $\lambda : V \to \{0, \dots, K-1\}$. Our implementation labels blocks deterministically: after every milestone snapshot the blocks are re-enumerated in the order of their minimal member fingerprints (Lemma 3), so every induced labeling (ball ids, cluster ids, coarse node ids, coarse edge ids) is a deterministic function of isomorphism-invariant keys. We call such labelings *canonical*. All per-cluster objects (feature projections $P_a^{(\ell)}$, Laplacian blocks, framelet blocks) are indexed by canonical ids.

Definition 1 (Permutation equivariance)**.** *A node-level algorithm $\mathcal{A}$ mapping (G, X) to predictions $\hat{Y} \in \mathbb{R}^{N\times c}$ is* permutation equivariant *if for every $\pi \in S_N$,*

$$\mathcal{A}(\pi G,\, P_\pi X) \;=\; P_\pi\, \mathcal{A}(G, X). \tag{11}$$

A grouping-valued map $\mathcal{C}(G, X)$ is permutation invariant *if $\mathcal{C}(\pi G, P_\pi X) = \pi\, \mathcal{C}(G, X) := \{\pi C : C \in \mathcal{C}(G, X)\}$ as sets of blocks.*

General-position assumption (exact ties). Key ties between *non-equivalent* candidates cannot be resolved invariantly by *any* deterministic rule: if two nodes v, w attain exactly the same seed distance and fingerprint, and v, w are exchanged by an automorphism of (G, X), then choosing one but not the other as a seed necessarily breaks invariance under that automorphism. This is intrinsic, not an artifact of our construction (the same obstruction applies to Ward linkage, which breaks ties by node index, and to the centrality rankings of HGBG). We therefore assume:

> **(GP)** All compared key tuples (seed distances of Eq. 1, split-quality margins of Eq. 3, and fingerprint components, compared lexicographically) of distinct candidates are distinct. In particular, the node tuples defining the fingerprint of Section 4.2(a) are pairwise distinct, so the fingerprint ranking is well defined and invariant.

Under any continuous distribution of node features, (GP) holds with probability one. For a fixed dataset, however, (GP) is a deterministic, checkable property rather than a probabilistic event, and we state it as a per-dataset assumption; its violations are confined to exchanges of nodes that agree on every invariant key, which leave the node predictions unaffected (Remark 3). (GP) is invoked wherever a unique extremizer or an exact key comparison is required: Lemmas 4–7 and, through them, Lemmas 8–10 and Theorem 4; Lemmas 1, 2, and 3 hold unconditionally. All equality statements below hold in exact arithmetic; in floating-point arithmetic they hold up to the rounding of order-dependent reductions, which the canonical accumulation orders below fix on a deterministic reduction path (residual device-level nondeterminism is discussed in Remark 4).

Remark 1 (Scope relative to PEGFAN)**.** *Theorem 4 of Li et al. (2024) states $\mathcal{A}(\pi\mathcal{G}, \mathcal{P}_K) = \pi\mathcal{A}(\mathcal{G}, \mathcal{P}_K)$ for a* given *partition tree $\mathcal{P}_K$, covering node permutations as well as permutations of the tree indices, and their Proposition 2 lifts the result to the PEGFAN network: the model is permutation equivariant once the tree is given. How $\mathcal{P}_K$ is generated lies outside these results: whether the Ward clustering that produces $\mathcal{P}_K$ satisfies $\mathcal{P}_K(\pi\mathcal{G}) = \pi\mathcal{P}_K(\mathcal{G})$ is not covered; spectral embeddings carry eigenvector-sign and degenerate-subspace ambiguity, and Ward linkage resolves distance ties by node index. Theorem 4 below closes exactly this residual gap: the partition tree is* generated *by a permutation-invariant procedure, so equivariance holds (under (GP)) for the end-to-end algorithm with the clustering stage included. In addition, we reuse PEGFAN's framelet transform only through the projector $\Psi_j^\top \Psi_j$, which is independent of the per-cluster orthonormal bases; our Reduce stage (Lemma 8) therefore inherits none of the basis ambiguity of the framelet vectors themselves.*

A.2 Lemma 1: end-to-end restriction-map generation is equivariant

Lemma 1 (Restriction maps)**.** *Let $\mathcal{E}$ denote the shared SheafMLP that maps the current level-0 signal H to per-edge restriction maps through*

$$\big(F_{u\to e}, F_{v\to e}\big) = g_\Theta\big([\vec{H}_u;\, \vec{H}_v]\big) \;\in\; \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d} \qquad \text{for every edge } e = (u, v), \tag{12}$$

where $\vec{H}_u \in \mathbb{R}^{d\, m_0}$ is the unfolded stalk signal of node u and the edge orientation is fixed canonically (Section 4.5). Whenever the signal co-varies with $\pi \in S_N$, the maps coincide on corresponding edges: writing (F^π) for the maps produced by $\mathcal{E}$ on the permuted run from $H^\pi = \Pi_\pi H$,

$$F^\pi_{\pi u\to\pi e} = F_{u\to e}, \qquad F^\pi_{\pi v\to\pi e} = F_{v\to e}. \tag{13}$$

Proof. The SheafMLP is a shared edge-wise map g_Θ applied to the concatenated endpoint features, with edge orientation fixed canonically. On edge $\pi e = (\pi u, \pi v)$ of the permuted run,

$$\big(F^\pi_{\pi u\to\pi e}, F^\pi_{\pi v\to\pi e}\big) = g_\Theta\big([H^\pi_{\pi u}; H^\pi_{\pi v}]\big) = g_\Theta\big([H_u; H_v]\big) = \big(F_{u\to e}, F_{v\to e}\big), \tag{14}$$

since $H^\pi_{\pi u} = H_u$ and $H^\pi_{\pi v} = H_v$. In the end-to-end pipeline the signal is the lifted level-0 representation of the current forward pass, whose co-variance is established in Lemma 10(i); the present lemma supplies the interface used in step (iii) of that proof. □

A.3 Lemma 2: GPNN structural ids are isomorphism invariants

Lemma 2 (GPNN core-degree ids (Hevapathige & Wang, 2026))**.** *Let pid_u be the core-degree partition id: nodes with core number $c(u) \in \{0, 1\}$ form partition 0, and within each j-shell ($j \geq 2$) the nodes attaining the minimum in-j-core degree form partition $2j-3$, the rest $2j-2$ (Hevapathige & Wang, 2026). Then $\mathrm{pid}^\pi_{\pi u} = \mathrm{pid}_u$ for all u.*

Proof. $\mathrm{pid}_u = \phi\big(c(u), \deg_{c(u)\text{-core}}(u)\big)$ for a deterministic ϕ. The k-core $\mathrm{core}_k(G)$ is the unique maximal induced subgraph of minimum degree $\geq k$; since π is a graph isomorphism it preserves induced-subgraph degrees and maximality, so $\mathrm{core}_k(\pi G) = \pi\,\mathrm{core}_k(G)$. Hence

$$c^\pi(\pi u) = \max\{k : \pi u \in \mathrm{core}_k(\pi G)\} = \max\{k : u \in \mathrm{core}_k(G)\} = c(u), \tag{15}$$

and $\deg^{\pi}_{c\text{-core}}(\pi u) = |\,\pi N(u) \cap \pi\,\mathrm{core}_{c(u)}(G)\,| = |\,N(u) \cap \mathrm{core}_{c(u)}(G)\,| = \deg_{c\text{-core}}(u)$. Therefore $\mathrm{pid}^{\pi}_{\pi u} = \phi\big(c(u), \deg_{c\text{-core}}(u)\big) = \mathrm{pid}_u$. □

A.4 Lemma 3: node fingerprints and the seed keys are invariant

Lemma 3 (Fingerprints)**.** *Let $\mathrm{fp}(u)$ be the fingerprint of Section 4.2(a): the rank of the tuple (two-round WL color, pid_u, $\deg(u)$, $\mathrm{crc32}(x_u)$). Then $\mathrm{fp}^{\pi}_{\pi u} = \mathrm{fp}_u$ for all u. The two-hop Jaccard similarity of Eq. 1 is likewise invariant: $S^{\pi}_{\mathrm{jac2}}(\pi u, \pi v) = S_{\mathrm{jac2}}(u, v)$.*

Proof. Component-wise invariance of the tuple:

$$\text{(i)}\quad \mathrm{crc32}(x^{\pi}_{\pi u}) = \mathrm{crc32}(x_u); \tag{16}$$

$$\text{(ii)}\quad \deg^{\pi}(\pi u) = |\pi N(u)| = |N(u)| = \deg(u); \tag{17}$$

$$\text{(iii)}\quad \mathrm{col}_{t+1}(u) = \Psi\Big(\mathrm{col}_t(u),\, \bigoplus\nolimits_{w\sim u} \mathrm{col}_t(w),\, \textstyle\sum_{w\sim u} \mathrm{col}_t(w),\, \deg(u)\Big), \tag{18}$$

where Ψ is an order-free function of the neighbor multiset. Since $N^{\pi}(\pi u) = \pi N(u)$, the multiset $\{\mathrm{col}_t(w) : w \sim u\}$ co-varies, and induction on t gives $\mathrm{col}^{\pi}_t(\pi u) = \mathrm{col}_t(u)$; (iv) $\mathrm{pid}^{\pi}_{\pi u} = \mathrm{pid}_u$ by Lemma 2. Every tuple component is invariant, hence so is the tuple; the rank map sorts the tuple multiset, which the bijection π preserves, so $\mathrm{fp}^{\pi}_{\pi u} = \mathrm{fp}_u$. For the two-hop term, $N^{\pi}_2(\pi u) = \pi N_2(u)$, and $|\pi A \cap \pi B| = |A \cap B|$, $|\pi A \cup \pi B| = |A \cup B|$, hence

$$S^{\pi}_{\mathrm{jac2}}(\pi u, \pi v) = \frac{|\pi N_2(u) \cap \pi N_2(v)|}{|\pi N_2(u) \cup \pi N_2(v)|} = \frac{|N_2(u) \cap N_2(v)|}{|N_2(u) \cup N_2(v)|} = S_{\mathrm{jac2}}(u, v). \tag{19}$$

□

A.5 Lemma 4: the divisive root balls are invariant

Lemma 4 (Root balls)**.** *Under (GP), the root-ball grouping (one ball per connected component, centers at $\arg\max(\mathrm{core}(u), -\mathrm{fp}(u))$, assignment by invariant multi-source BFS) is permutation invariant, and canonically labeled.*

Proof. Connected components are the unique maximal connected induced subgraphs, so $\{\text{components of } \pi G\} = \{\pi C : C \text{ component of } G\}$. Within a component C,

$$s(\pi C) = \arg\max_{u' \in \pi C} \big(\mathrm{core}(u'), -\mathrm{fp}(u')\big) = \pi \arg\max_{u \in C} \big(\mathrm{core}(u), -\mathrm{fp}(u)\big) = \pi\, s(C), \tag{20}$$

since both key entries are invariant (Lemmas 2, 3) and the maximizer is unique under (GP). Writing Σ for the center set, the multi-source BFS owner map is the lexicographic minimizer

$$\mathrm{owner}(v) \;=\; \arg\min_{s \in \Sigma} \big(\mathrm{dist}(v, s),\, \mathrm{fp}(s),\, \mathrm{fp}(v)\big), \tag{21}$$

realized canonically by the lazy-heap pop order. Every key component is invariant:

$$\mathrm{dist}^{\pi}(\pi v, \pi s) = \mathrm{dist}(v, s), \qquad \mathrm{fp}^{\pi}(\pi s) = \mathrm{fp}(s), \qquad \mathrm{fp}^{\pi}(\pi v) = \mathrm{fp}(v), \tag{22}$$

so the lexicographic minimizer satisfies

$$\mathrm{owner}^{\pi}(\pi v) = \arg\min_{s' \in \pi\Sigma} \big(\mathrm{dist}(\pi v, s'), \mathrm{fp}(s'), \mathrm{fp}(\pi v)\big) = \pi\, \mathrm{owner}(v), \tag{23}$$

i.e. the ball of center s maps to the ball of center πs: the root grouping of the permuted input is $\{\pi B\}$. Canonical labeling enumerates the components in order of their center fingerprints, which is invariant. □

A.6 Lemma 5: one bisection round is equivariant

Lemma 5 (Bisection)**.** *Let B be a ball, split into children (a, b) by the procedure of Section 4.2(c). Under (GP), the same procedure applied to πB (the image ball in the permuted run) yields $(\pi a, \pi b)$.*

Proof. Seeds. s_1 is the unique maximizer of the invariant key $(\mathrm{core}(v), -\mathrm{fp}(v))$ over $v \in B$, so $s_1^\pi = \pi s_1$. The seed-2 score of Eq. 1 is invariant term by term:

$$d^\pi_{\mathrm{feat}}(\pi v) = \tfrac{1-\cos(x^\pi_{\pi v}, x^\pi_{\pi s_1})}{2} = \tfrac{1-\cos(x_v, x_{s_1})}{2} = d_{\mathrm{feat}}(v), \tag{24}$$

$$d^\pi_{\mathrm{struct}}(\pi v) = 1 - S^\pi_{\mathrm{jac2}}(\pi v, \pi s_1) = 1 - S_{\mathrm{jac2}}(v, s_1) = d_{\mathrm{struct}}(v) \qquad \text{(Lemma 3)}, \tag{25}$$

$$d^\pi_{\mathrm{role}}(\pi v) = \mathbf{1}[\mathrm{pid}^\pi_{\pi v} \neq \mathrm{pid}^\pi_{\pi s_1}] = \mathbf{1}[\mathrm{pid}_v \neq \mathrm{pid}_{s_1}] = d_{\mathrm{role}}(v) \qquad \text{(Lemma 2)}, \tag{26}$$

so $d^\pi_2(\pi v) = d_2(v)$, and the lexicographic maximizer

$$s_2 \;=\; \arg\max_{v \in B} \big(\mathrm{round}(d_2(v), 12),\; -\mathrm{fp}(v)\big), \tag{27}$$

where rounding to 12 decimals absorbs floating-point reduction noise before the key comparison, is unique under (GP) and satisfies $s_2^\pi = \pi s_2$.

Two-source BFS. The bisection assignment is the two-source case of the invariant minimizer Eq. 21 on the invariant subset B with source keys $(\mathrm{fp}(s_1), \mathrm{fp}(s_1)+1)$: by Lemma 4, $\mathrm{side}^\pi(\pi v) = \mathrm{side}(v)$. For members unreachable from both seeds,

$$\langle x^\pi_{\pi v}, x^\pi_{\pi s_1}\rangle - \langle x^\pi_{\pi v}, x^\pi_{\pi s_2}\rangle = \langle x_v, x_{s_1}\rangle - \langle x_v, x_{s_2}\rangle, \tag{28}$$

with fp tie-break, so each goes to the corresponding side. Hence $(a^\pi, b^\pi) = (\pi a, \pi b)$.

Balanced fallback. The activation predicate (forced split with smaller side $< |B|/4$) is invariant; the halves are obtained by sorting the members in the lexicographic order of $(-\mathrm{fp}(v), -\mathrm{gap}(v))$ and cutting at position $|B| - \lfloor |B|/2 \rfloor$, and $\mathrm{gap}^\pi(\pi v) = \mathrm{gap}(v)$ by the display above, so the two halves of πB are the π-images of the halves of B. □

A.7 Lemma 6: the quality gate, cap, and milestone extraction are invariant

Lemma 6 (Splitting loop)**.** *Under (GP), the quality-gated splitting loop (Section 4.2(d)–(e)) applied to (G, X) and to $(\pi G, P_\pi X)$ performs corresponding split sequences: at every iteration the ball frontiers are π-images of each other, the same ball is popped (up to π), and the acceptance decision coincides. Consequently every milestone snapshot satisfies $\lambda^\pi = \lambda \circ \pi^{-1}$ after canonical relabeling.*

Proof. Run both loops in parallel; the induction claim is that at every iteration the two frontiers are π-images with matching canonical labels.

Heap order. Entries are keyed by $\big(-|B|,\, \min_{v\in B} \mathrm{fp}(v),\, \mathrm{id}(B)\big)$, all invariant, so both loops pop corresponding balls.

Quality oracle. For a ball B with image πB:

$$Q^\pi_{\mathrm{feat}}(\pi B) = \tfrac{1}{|B|}\textstyle\sum_{v\in B} \cos\big(x_v, \mu_B\big) = Q_{\mathrm{feat}}(B), \tag{29}$$

$$Q^\pi_{\mathrm{str}}(\pi B) = 1 - \bar{H}\big(\{\{\mathrm{pid}_v : v \in \pi B\}\}\big) = Q_{\mathrm{str}}(B), \tag{30}$$

$$Q^\pi_{\mathrm{comm}}(\pi B) = \frac{e_{\mathrm{in}}(\pi B)}{|E^\pi|} - \Big(\frac{\mathrm{vol}(\pi B)}{2|E^\pi|}\Big)^2 \tag{31}$$

$$= \frac{e_{\mathrm{in}}(B)}{|E|} - \Big(\frac{\mathrm{vol}(B)}{2|E|}\Big)^2 = Q_{\mathrm{comm}}(B), \tag{32}$$

where $\mu_B = \frac{1}{|B|}\sum^{\mathrm{fp}}_{v\in B} x_v$ and $\sum^{\mathrm{fp}}$ denotes accumulation in fingerprint-sorted order, so μ_B is equal under π in exact arithmetic, and in floating point whenever the reduction follows this canonical order, $\bar{H}$ the normalized entropy of the pid histogram (preserved by π), and $e_{\mathrm{in}}, \mathrm{vol}$ are invariant cardinalities. Hence $Q^\pi(\pi B) = Q(B)$, and with the corresponding children of Lemma 5,

$$Q^\pi(\pi a) + Q^\pi(\pi b) > Q^\pi(\pi B) + \varepsilon \iff Q(a) + Q(b) > Q(B) + \varepsilon. \tag{33}$$

The hard-cap predicate $|B| > \kappa h$ depends on the invariant size only, so forced splits occur on corresponding balls (Lemma 5 covers the fallback); rejected and unsplittable balls correspond trivially. The claim is preserved in all four cases.

Milestones. $T_\ell = \lceil N/h^\ell \rceil$ and the cap-satisfaction test $\max_B |B| \le \kappa h$ are invariant, so both loops record corresponding snapshots at corresponding iterations; canonical relabeling by the fingerprint order of the blocks gives $\lambda^\pi(\pi u) = \lambda(u)$, i.e. $\lambda^\pi = \lambda \circ \pi^{-1}$. The termination criteria (heap empty, all milestones met, split budget $10N$) are invariant, so the loops terminate after the same number of accepted splits. □

A.8 LEMMA 7 (SELECT): THE WHOLE HIERARCHY IS INVARIANT

Lemma 7 (Select stage). *Let $S^{(\ell)} \in \{0,1\}^{N_{\ell-1}\times N_\ell}$ be the assignment matrices, $A^{(\ell)}$, $X^{(\ell)}$ the coarse adjacencies and features, and L the number of levels. Under (GP), for every π there exist permutations σ_ℓ of the level-ℓ nodes (with $\sigma_0 = \pi$) such that for all ℓ:*

$$S^{(\ell),\pi} = P_{\sigma_{\ell-1}}\, S^{(\ell)}\, P_{\sigma_\ell}^\top, \qquad A^{(\ell),\pi} = P_{\sigma_\ell} A^{(\ell)} P_{\sigma_\ell}^\top, \qquad X^{(\ell),\pi} = P_{\sigma_\ell} X^{(\ell)}, \tag{34}$$

and $L^\pi = L$. Consequently, with $D_\ell = \mathrm{diag}(|C_a|)$,

$$\mathrm{pool}_\ell^\pi = \left(P_{\sigma_\ell}\otimes I_d\right)\mathrm{pool}_\ell\left(P_{\sigma_{\ell-1}}\otimes I_d\right)^\top, \qquad \mathrm{prolong}_\ell^\pi = \left(P_{\sigma_{\ell-1}}\otimes I_d\right)\mathrm{prolong}_\ell\left(P_{\sigma_\ell}\otimes I_d\right)^\top. \tag{35}$$

Proof. The milestones T_ℓ and the number of levels are invariant (Lemma 6), so both runs extract the same number L of levels. By Lemma 6, the level-ℓ snapshot of the permuted run is the π-image of the level-ℓ snapshot of the original run, canonically labeled; writing σ_ℓ for the induced cluster-level permutation, every level-0 node u and its image πu are assigned to corresponding clusters, which is exactly $S^{(\ell),\pi} = P_\pi S^{(\ell)} P_{\sigma_\ell}^\top$. Nestedness makes the same statement propagate level by level with $\sigma_{\ell-1}$ in place of π. Then

$$A^{(\ell),\pi} = S^{(\ell),\pi\top} A^{(\ell-1),\pi} S^{(\ell),\pi} = P_{\sigma_\ell} S^{(\ell)\top} P_{\sigma_{\ell-1}}^\top P_{\sigma_{\ell-1}} A^{(\ell-1)} P_{\sigma_{\ell-1}}^\top P_{\sigma_{\ell-1}} S^{(\ell)} P_{\sigma_\ell}^\top \tag{36}$$

$$= P_{\sigma_\ell}\left(S^{(\ell)\top} A^{(\ell-1)} S^{(\ell)}\right) P_{\sigma_\ell}^\top = P_{\sigma_\ell} A^{(\ell)} P_{\sigma_\ell}^\top, \tag{37}$$

since $P_{\sigma_{\ell-1}}^\top P_{\sigma_{\ell-1}} = I$; the diagonal removal and zero elimination commute with conjugation. The same conjugation for the pooled features follows because the canonical group means coincide (fingerprint-sorted accumulation, Lemma 6). Finally, cluster sizes are preserved blockwise, $|C_{\sigma_\ell a}| = |C_a|$, so $D_\ell^\pi = P_{\sigma_\ell} D_\ell P_{\sigma_\ell}^\top$ and the pool/prolong conjugation follows from that of S by tensoring with I_d. □

A.9 LEMMA 8 (REDUCE): THE HAAR FRAMELET PROJECTORS CO-VARY

Lemma 8 (Framelet operators). *Let $\mathcal{S}_\ell = \mathrm{prolong}_\ell \mathrm{pool}_\ell = (S^{(\ell)} D_\ell^{-1} S^{(\ell)\top})\otimes I_d$ and $\mathcal{W}_\ell = I - \mathcal{S}_\ell$ be the framelet operators of Eq. 5. Under (GP),*

$$\mathcal{S}_\ell^\pi = \Pi_{\sigma_{\ell-1}}\, \mathcal{S}_\ell\, \Pi_{\sigma_{\ell-1}}^\top, \qquad \mathcal{W}_\ell^\pi = \Pi_{\sigma_{\ell-1}}\, \mathcal{W}_\ell\, \Pi_{\sigma_{\ell-1}}^\top. \tag{38}$$

Both operators are also exactly invariant under cluster relabeling: replacing $S^{(\ell)}$ by $S^{(\ell)}Q$ for any permutation Q of the level-ℓ clusters leaves $\mathcal{S}_\ell$ and $\mathcal{W}_\ell$ unchanged. The implementation maintains the required synchronization by construction: canonical relabeling re-indexes $S^{(\ell)}$, D_ℓ, and every cluster-indexed cache with the same Q.

Proof. By Lemma 7,

$$\mathrm{prolong}_\ell^\pi\, \mathrm{pool}_\ell^\pi = \left(P_{\sigma_{\ell-1}}\otimes I_d\right)\mathrm{prolong}_\ell\left(P_{\sigma_\ell}\otimes I_d\right)^\top\left(P_{\sigma_\ell}\otimes I_d\right)\mathrm{pool}_\ell\left(P_{\sigma_{\ell-1}}\otimes I_d\right)^\top, \tag{39}$$

and the two middle factors cancel by orthogonality, giving the conjugation for $\mathcal{S}_\ell$; the claim for $\mathcal{W}_\ell$ follows since the identity is fixed by conjugation. For the relabeling statement, D_ℓ^{-1} transforms to $Q^\top D_\ell^{-1} Q$, and $S^{(\ell)} Q\, Q^\top D_\ell^{-1} Q\, Q^\top S^{(\ell)\top} = S^{(\ell)} D_\ell^{-1} S^{(\ell)\top}$. Note that this argument involves no eigendecomposition: the framelet operators are degree-rational functions of the assignment matrix and the (invariant) cluster sizes, in contrast to spectral band constructions whose eigenbases are defined only up to sign and degenerate rotation. □

Remark 2 (A three-node cluster). *On a cluster $C_a = \{1,2,3\}$ (one stalk coordinate), Eq. 5 gives*

$$\mathcal{S}_\ell\big|_{C_a} = \tfrac{1}{3}\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix}, \qquad \mathcal{W}_\ell\big|_{C_a} = \tfrac{1}{3}\begin{pmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{pmatrix}. \tag{40}$$

Assembling PEGFAN's framelet vectors for the same cluster, e.g. $\psi_1 = (1,-1,0)/\sqrt{2}$, $\psi_2 = (1,1,-2)/\sqrt{6}$, yields $\Psi^\top\Psi = \sum_j \psi_j\psi_j^\top = \mathcal{W}_\ell\big|_{C_a}$; any other orthonormal basis of the zero-mean subspace gives the same operator, since it differs by an orthogonal transformation Q and $\Psi^\top Q^\top Q\Psi = \Psi^\top\Psi$. The projector form thus retains the Haar semantics (cluster mean versus zero-mean detail) while eliminating the basis choice that the framelet vectors themselves require.

A.10 LEMMA 9 (CONNECT): COARSE SHEAF BLOCKS CO-VARY

Lemma 9 (Galerkin blocks)**.** *Fix a forward pass and the restriction maps used in it (Lemma 1). For every level ℓ: $D^{(\ell),\pi} = (P_{\sigma_\ell} \otimes I_d)\, D^{(\ell)}\, (P_{\sigma_\ell} \otimes I_d)^\top$ blockwise (i.e. $D^{\pi}_{\sigma_\ell a} = D_a$) and $B^{(\ell),\pi}_{\sigma_\ell a, \sigma_\ell b} = B^{(\ell)}_{ab}$. Consequently the normalized Laplacian blocks satisfy $(D^{-1/2})^{(\ell),\pi} = (P_{\sigma_\ell} \otimes I_d)\, (D^{-1/2})^{(\ell)}\, (P_{\sigma_\ell} \otimes I_d)^\top$.*

Proof. The maps coincide on corresponding edges (Lemma 1), and the level-ℓ blocks of the permuted run are indexed by σ_ℓ (Lemma 7), so the summands correspond bijectively under π:

$$B^{(\ell),\pi}_{\sigma_\ell a, \sigma_\ell b} = \sum_{\substack{u' \in \sigma_\ell a \\ v' \in \sigma_\ell b}} \left(F^{\pi}_{u' \to e'}\right)^\top F^{\pi}_{v' \to e'} = \sum_{\substack{u \in a \\ v \in b}} F^\top_{u \to e} F_{v \to e} = B^{(\ell)}_{ab}, \tag{41}$$

with equality in exact arithmetic, and in floating point under the canonical edge ordering, which fixes the summation order of the reduction; identically for $D^{(\ell)}_{\sigma_\ell a} = D^{(\ell)}_a$. Finally, $D^{-1/2} = g\big(D + 10^{-3} I\big)$ with $g(t) = t^{-1/2}$, and spectral functions of block-diagonal matrices conjugate under any orthogonal R: $g(RMR^\top) = R\, g(M)\, R^\top$, independently of the chosen eigenbasis (including degenerate eigenspaces). Hence $(D^{-1/2})^{(\ell),\pi} = (P_{\sigma_\ell} \otimes I_d)\, (D^{-1/2})^{(\ell)}\, (P_{\sigma_\ell} \otimes I_d)^\top$. □

A.11 LEMMA 10: THE SHEAR FORWARD PASS IS EQUIVARIANT

Lemma 10 (Forward pass)**.** *Fix all learnable parameters Θ (with per-cluster tensors indexed by the canonical cluster ids). Under (GP), for every π the network output satisfies $\hat{Y}(\pi G, P_\pi X; \Theta) = P_\pi\, \hat{Y}(G, X; \Theta)$.*

Proof. Write $\Pi_\ell = P_{\sigma_\ell} \otimes I_d$ for the induced stalk-space permutations (Lemma 7) and track every intermediate tensor.

(i) Stalk lifting. lift is a row-wise linear map: $\vec{H}^{(0),\pi} = \overrightarrow{(P_\pi X)W} = P_\pi \overrightarrow{XW} = \Pi_0 \vec{H}^{(0)}$.

(ii) Propagation channels. The normalized adjacencies satisfy $\hat{A}^\pi = P_\pi \hat{A} P_\pi^\top$ (degrees are invariant), so each fixed channel $\hat{A}^k X^\pi = P_\pi \hat{A}^k X$, and likewise for the self-looped variant.

(ii-b) Detail channels. Write $S_{1:\ell} = S^{(1)} \cdots S^{(\ell)}$ for the composed level-$0 \to \ell$ assignment matrix and $\Upsilon_\ell = S_{1:\ell}(S_{1:\ell}^\top S_{1:\ell})^{-1} S_{1:\ell}^\top$ for the orthogonal projector onto level-ℓ cluster-constant signals. By Lemma 7, $S^\pi_{1:\ell} = P_\pi S_{1:\ell} Q$ for a cluster permutation Q, and Q cancels in the projector exactly as in Lemma 8: $\Upsilon^\pi_\ell = P_\pi \Upsilon_\ell P_\pi^\top$. The detail projectors $\Xi_\ell = \Upsilon_{\ell-1} - \Upsilon_\ell$ and the coarsest scaling projector Υ_L therefore conjugate, and with $x = \hat{A}_I X$ (co-variant by (ii)) every channel $\Xi_\ell x$ and $\Upsilon_L x$ satisfies $Z^\pi = P_\pi Z$.

(iii) End-to-end restriction maps. The SheafMLP regenerates the maps from $\vec{H}^{(0),\pi} = \Pi_0 \vec{H}^{(0)}$, so by Lemma 1 the maps coincide on corresponding edges, and by Lemma 9 all Galerkin blocks co-vary within this forward pass.

(iv) Analysis cascade. Suppose $\vec{H}^{(\ell-1),\pi} = \Pi_{\ell-1} \vec{H}^{(\ell-1)}$. The band-mixed signal satisfies, by Lemma 8 with shared gains θ_ℓ,

$$\theta_{\ell,\mathrm{lp}} \mathcal{S}^\pi_\ell \vec{H}^{(\ell-1),\pi} + \theta_{\ell,\mathrm{hp}} \mathcal{W}^\pi_\ell \vec{H}^{(\ell-1),\pi} = \Pi_{\ell-1} \big(\theta_{\ell,\mathrm{lp}} \mathcal{S}_\ell + \theta_{\ell,\mathrm{hp}} \mathcal{W}_\ell\big) \vec{H}^{(\ell-1)}. \tag{42}$$

Pooling: $\mathrm{pool}^\pi_\ell \Pi_{\ell-1} = \Pi_\ell\, \mathrm{pool}_\ell$ (Lemma 7). The per-cluster projection applies $P^{(\ell)}_a$ to cluster a; with canonical indexing the parameter of cluster $\sigma_\ell a$ in the permuted run is the same tensor $P^{(\ell)}_a$, so the projection commutes with Π_ℓ. Hence $\vec{H}^{(\ell),\pi} = \Pi_\ell \vec{H}^{(\ell)}$, completing the induction.

(v) Multi-scale sheaf diffusion. At level ℓ, the Laplacian blocks co-vary (Lemma 9 via (iii)), so $\Delta^{(\ell)}_\pi = \Pi_\ell \Delta^{(\ell)} \Pi_\ell^\top$; the edge-wise matvec of $\Delta^{(\ell)}$ over the (conjugated) coarse edge list realizes the same linear operator. The coarse edge list is enumerated canonically (lexicographically in the canonical endpoint ids), so the enumeration of the permuted run coincides after relabeling and the

edge-wise reductions follow the same order. One diffusion step on the permuted input, starting from $\vec{Z}^{\pi} = \Pi_\ell \vec{Z}$, gives

$$\vec{Z}^{\pi} - \tau_{\ell,t} \tanh\left(\Pi_\ell \Delta^{(\ell)} \Pi_\ell^{\top} (I \otimes W_1) \Pi_\ell \vec{Z} W_2\right) = \Pi_\ell \Big[\vec{Z} - \tau_{\ell,t} \tanh\left(\Delta^{(\ell)} (I \otimes W_1) \vec{Z} W_2\right)\Big], \quad (43)$$

using orthogonality of Π_ℓ, entrywise $\tanh$, and commutativity with the channel transform W_2. Induction over the T_0/T_c steps and over the levels yields $\vec{H}^{(\ell),\pi}_{\text{diff}} = \Pi_\ell \vec{H}^{(\ell)}_{\text{diff}}$.

(vi) Synthesis and prolongation chains. By Lemma 7, $\text{prolong}_k^{\pi} \Pi_k = \Pi_{k-1} \text{prolong}_k$; the synthesis operator $\mathcal{T}_k = \eta_{k,\text{lp}} \mathcal{S}_k + \eta_{k,\text{hp}} \mathcal{W}_k$ conjugates by Lemma 8 with shared gains η_k; the shared back-projections $P^{(k)}_{\text{back}}$ act on the channel dimension and commute with Π_{k-1}. Hence every chain step conjugates, and $Z^{(\ell),\pi} = \Pi_0 Z^{(\ell)}$.

(vii) Global channels. Z_{LP} and Z_{HP} are sums of compositions of pool, framelet projectors, per-cluster projections, and prolongations, each of which conjugates by (iv)–(vi); averaging over levels preserves this.

(viii) Fusion. After (ii)–(vii), every channel matrix Z_k satisfies $Z_k^{\pi} = P_\pi Z_k$. Each fc1_k is row-wise, as is the ℓ_2 normalization, so $\text{fc1}_k(Z_k^{\pi}) = P_\pi \text{fc1}_k(Z_k)$. The attention weights are either global scalars (shared parameter, identical on both inputs) or node-level scores from a shared MLP (equivariant: $\alpha^{\pi} = P_\pi \alpha$). Concatenation, ReLU, dropout (in eval mode), and the final row-wise fc2 all commute with P_π. Therefore $\hat{Y}^{\pi} = P_\pi \hat{Y}$. □

A.12 Main theorem and training-time corollary

Theorem 4 (Permutation equivariance of SHEAR)**.** *Under the general-position assumption (GP), SHEAR is permutation equivariant in the sense of Eq. 11: for every graph G, features X, permutation $\pi \in S_N$, and all parameter values Θ,*

$$\hat{Y}(\pi G,\, P_\pi X;\, \Theta) \;=\; P_\pi\, \hat{Y}(G, X;\, \Theta), \quad (44)$$

where the restriction maps are generated end-to-end from the current features inside every forward pass.

Proof. Compose the stages: Lemma 1 (restriction maps), Lemmas 4–7 (granular-ball hierarchy and pool/prolong), Lemmas 8–9 (framelet operators and sheaf blocks), and Lemma 10 (forward pass). Each stage consumes the output of the previous one with the stated co-variance, so the composition is equivariant. □

Corollary 1 (Equivariance is preserved by training)**.** *Let $\mathcal{L}(\Theta; G, X, y) = \sum_i \ell(\hat{Y}_i(G, X; \Theta), y_i)$ be the per-node cross-entropy loss. Under (GP), $\mathcal{L}(\Theta; \pi G, P_\pi X, P_\pi y) = \mathcal{L}(\Theta; G, X, y)$, and for every gradient-based update with permutation-invariant initialization (shared parameters initialized identically; per-cluster projections initialized to the same identity matrix for all clusters), the entire training trajectory on $(\pi G, P_\pi X, P_\pi y)$ coincides with the trajectory on (G, X, y), provided the stochastic layers (dropout) draw from a shared random seed; without a shared seed the two trajectories coincide in distribution over the dropout masks. In particular the trained model satisfies Theorem 4.*

Proof. By Theorem 4, $\hat{Y}(\pi G, P_\pi X; \Theta) = P_\pi \hat{Y}(G, X; \Theta)$, so

$$\mathcal{L}(\Theta; \pi G, P_\pi X, P_\pi y) = \sum_i \ell\big([P_\pi \hat{Y}]_i, [P_\pi y]_i\big) = \sum_i \ell\big(\hat{Y}_{\pi^{-1} i}, y_{\pi^{-1} i}\big) = \mathcal{L}(\Theta; G, X, y). \quad (45)$$

Hence $\nabla_\Theta \mathcal{L}$ coincides on the two runs: shared parameters receive equal gradients, and canonical-id-indexed cluster tensors receive equal gradients tensor-by-tensor (canonical indexing is π-invariant). With a permutation-invariant initialization $\Theta^{(0),\pi} = \Theta^{(0)}$ (shared parameters identical; per-cluster projections all initialized to the same identity), induction over the updates gives $\Theta^{(t),\pi} = \Theta^{(t)}$ for all t; adaptive optimizers (Adam) preserve this since their second-moment accumulators co-vary with the gradients, and with a shared random seed the dropout masks of the two runs coincide at every step. The trained model therefore satisfies Theorem 4. □

A.13 REMARKS: TIES AND NUMERICS

Remark 3 (The tie set)**.** *Assumption (GP) excludes inputs in which two candidates have exactly equal key tuples. On such inputs the clustering is still deterministic, but equivariance can fail under an automorphism that exchanges the tied alternatives, an obstruction shared by every deterministic clustering, including PEGFAN's Ward linkage (which resolves ties by node index and is therefore order-dependent even in generic position) and the centrality rankings of HGBG. The tie set has measure zero under any absolutely continuous feature distribution. Within the tie set, ties between genuine automorphic twins (identical WL colors, structural ids, degrees, and features) are harmless for the end-to-end model: any equivariant downstream map gives twins identical outputs, so exchanging their cluster assignments leaves the node predictions unchanged. Empirically, the GBSC partition tree is reproduced* exactly *under random node relabelings on the stochastic-block-model (SBM) benchmark (five independent permutations) and on highly symmetric graphs (star, cycle, double-star) where fingerprint ties are ubiquitous; the documented residual exception is confined to twin permutations as discussed above.*

Remark 4 (Numerical eigensolvers)**.** *The frequency decomposition of SHEAR involves no eigendecomposition at all: the framelet operators $\mathcal{S}_\ell, \mathcal{W}_\ell$ are rational functions of the assignment matrix and the cluster sizes (Lemma 8), so the eigenvector sign/degeneracy ambiguity of spectral constructions, including both PEGFAN's spectral* embedding *used for clustering and sheaflet-style spectral bands, is absent by design. The only numerical spectral step in the pipeline is the blockwise inverse square root of $D^{(\ell)} + 10^{-3}I$ in the Connect stage (Lemma 9); it is a spectral function of a block-diagonal matrix and is basis-independent in exact arithmetic, while floating-point eigensolvers may introduce basis-dependent rounding bounded by the solver tolerance. A second residual source of device-level nondeterminism is atomic-order-dependent reductions (scatter/index-add) on GPUs, whose accumulation order is not fixed by the hardware; the equality statements of this appendix are exact-arithmetic statements and are reproduced in floating point whenever the reduction order is deterministic. The crc32 component of the node fingerprint (Section 4.2(a)) is an exact function of the feature bytes and hence permutation invariant by construction; like any fixed-width hash it is byte-representation specific and admits collisions in principle, and both are absorbed by (GP), since a collision merges two node tuples into one, an event of the tie set analyzed in Remark 3. All split-independent hierarchy operators (partitions, pool/prolong, framelet projectors) are cached, so within one experimental run on a deterministic reduction path the pipeline is exactly deterministic.*

B ANALYSIS CASCADE, LITERAL PROLONGATION CHAINS, AND CHANNEL BANK

This appendix collects the architecture definitions deferred from Section 4.7. The analysis cascade coarsens the lifted signal $H = \mathrm{lift}(X) := XW_{\mathrm{lift}}$ (a row-wise linear map to m channels, unfolded to the stalk space) level by level, mixing the two framelet bands with learnable gains $\theta_\ell = \mathrm{softmax}(\Theta_\ell) \in \Delta^1$:

$$\widetilde{H}^{(\ell)} \;=\; \theta_{\ell,\mathrm{lp}}\,\mathcal{S}_\ell H^{(\ell-1)} + \theta_{\ell,\mathrm{hp}}\,\mathcal{W}_\ell H^{(\ell-1)}, \qquad C^{(\ell)} = \mathrm{pool}_\ell\,\widetilde{H}^{(\ell)}, \qquad H_a^{(\ell)} = C_a^{(\ell)} P_a^{(\ell)}, \tag{46}$$

with $\widetilde{H}^{(\ell)} \in \mathbb{R}^{N_{\ell-1}\times d\times m_{\ell-1}}$ the band-mixed signal at the incoming resolution, $C^{(\ell)} \in \mathbb{R}^{N_\ell\times d\times m_{\ell-1}}$ the pooled coarse signal, and per-cluster feature projections $P_a^{(\ell)} \in \mathbb{R}^{m_{\ell-1}\times m_\ell}$ (identity-initialized): each cascade step splits the signal into a cluster-mean (low) part and a within-cluster detail (high) part, re-weights the two bands, and carries only the means to the next coarser level. After diffusion, each level is brought back to the original resolution by the *literal* prolongation chain

$$Z^{(\ell)} \;=\; P_{\mathrm{back}}^{(1)}\,\mathcal{T}_1\,\mathrm{prolong}_1\;\cdots\;P_{\mathrm{back}}^{(\ell)}\,\mathcal{T}_\ell\,\mathrm{prolong}_\ell\;H_{\mathrm{diff}}^{(\ell)}, \qquad \mathcal{T}_k = \eta_{k,\mathrm{lp}}\,\mathcal{S}_k + \eta_{k,\mathrm{hp}}\,\mathcal{W}_k, \tag{47}$$

where step k scatters to level $k-1$, applies a framelet synthesis with learnable gains $\eta_k = \mathrm{softmax}(E_k) \in \Delta^1$, and back-projects the channels with the shared $P_{\mathrm{back}}^{(k)} \in \mathbb{R}^{m_k\times m_{k-1}}$.

The fusion consumes five groups of channels: (a) the propagation bank $\{X, \hat{A}^k X, \hat{A}_I^k X\}_{k\le k_{\max}}$; (b) the per-level detail channels $\Xi_\ell \hat{A}_I X$ ($\ell = 1,\dots,L$) and the coarsest scaling channel $\Upsilon_L \hat{A}_I X$, the operator form of PEGFAN's framelet channels built from the projectors Υ_ℓ, Ξ_ℓ of Section 4.10, optionally dilated with a per-dataset stride s and, on the two filtered benchmarks, instantiated as a

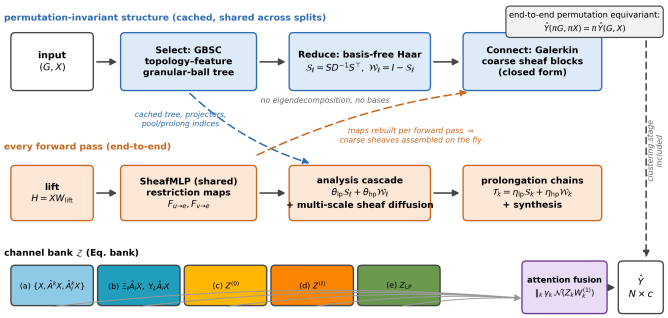

Figure 2: Overview of SHEAR. The entire multi-resolution structure (partition tree, framelet projectors, and Galerkin coarse sheaves) is generated by a single permutation-invariant procedure (Sections 4.2–4.4), so the end-to-end model is permutation equivariant with the clustering stage included (Theorem 1).

parallel multi-tree bank (details in Section G and Table 3); (c) the diffused features $Z^{(0)}$ on the original graph; (d) the per-level outputs $Z^{(\ell)}$ of Eq. 47 (with synthesis); and (e) the plain low-pass channel $Z_{\mathrm{LP}} = \frac{1}{L}\sum_\ell \mathrm{Prolong}_\ell P^{(\ell)} \mathrm{pool}_\ell \mathcal{S}_\ell H^{(\ell-1)}$ (no synthesis, $\mathrm{Prolong}_\ell$ the plain prolongation chain, $P^{(\ell)}$ the projection of Eq. 46); the matching high-pass channel vanishes identically ($\mathrm{pool}_\ell \mathcal{W}_\ell = 0$) and is omitted from the default bank (Section G). Elementwise, $(\Xi_\ell x)_u$ is the difference between the mean of x over u's level-$(\ell-1)$ cluster and over its level-ℓ cluster, so group (b) exposes exactly the within-cluster contrast that the analysis cascade discards ($\mathrm{pool}_\ell \mathcal{W}_\ell = 0$). Collecting (a)–(e), the channel bank is

$$\begin{aligned}\mathcal{Z} \;=\; &\{X,\, \hat{A}^k X,\, \hat{A}_I^k X\}_{k\le k_{\max}} \;\cup\; \{\Xi_\ell \hat{A}_I X\}_{\ell=1}^{L} \;\cup\; \{\Upsilon_L \hat{A}_I X\} \;\cup\; \{Z^{(0)}\} \\ &\cup\; \{Z^{(\ell)}\}_{\ell=1}^{L} \;\cup\; \{Z_{\mathrm{LP}}\},\end{aligned} \tag{48}$$

the five terms in the order of the groups (a)–(e) above, numbering $n_C = |\mathcal{Z}|$ channels in total ($2k_{\max} + 2L + 4$ with a single detail tree and the default two-band operators).

C Robustness to Oversquashing and Oversmoothing: Formal Statements and Proofs

This appendix gives the formal versions of Theorems 2 and 3. Throughout, $S_{1:\ell} := S^{(1)} S^{(2)} \cdots S^{(\ell)} \in \{0,1\}^{N \times N_\ell}$ is the composed level-$0 \to \ell$ assignment matrix (each row contains exactly one 1), and

$$\Upsilon_\ell \;:=\; S_{1:\ell}\,(S_{1:\ell}^\top S_{1:\ell})^{-1}\, S_{1:\ell}^\top \;\in\; \mathbb{R}^{N\times N} \tag{49}$$

is the orthogonal projector onto the subspace V_ℓ of signals that are constant on every level-ℓ cluster. Since the extracted levels are strictly nested (Section 4.2(e)), $V_0 \supset V_1 \supset \cdots \supset V_L$ with $V_0 = \mathbb{R}^N$, i.e. $\Upsilon_0 = I$. The detail projectors are $\Xi_\ell := \Upsilon_{\ell-1} - \Upsilon_\ell$. We write $C_\ell(u)$ for the level-ℓ cluster of u, $s_\ell := \max_a |C_a^{(\ell)}|$ for the largest level-ℓ cluster size, $d_{\max}$ for the maximum node degree, G_L for the coarsest coarse graph, and $J_{uv} := \partial H_u / \partial X_v$ for a Jacobian block of a node representation map $X \mapsto H$. All operators act coordinate-wise on feature space (tensored with the identity), which we suppress.

Lemma 11 (Entries of the scaling projector)**.** $(\Upsilon_\ell)_{uv} = |C_\ell(u)|^{-1}\, \mathbf{1}[v \in C_\ell(u)]$.

Proof. Each level-0 node belongs to exactly one level-ℓ cluster, so $S_{1:\ell}^\top S_{1:\ell} = \mathrm{diag}(|C_1|, \ldots, |C_{N_\ell}|)$. Entry (u,v) of Eq. 49 is therefore $|C_\ell(u)|^{-1}$ when u, v share their level-ℓ cluster and 0 otherwise. □

C.1 OVERSQUASHING

We adopt the sensitivity criterion of Topping et al. (2022); Di Giovanni et al. (2023): a model *oversquashes* if there exist $\sigma < 1$ and $D^* \geq 1$ such that

$$\|J_{uv}\|_2 \;\leq\; \sigma^{\,d_G(u,v)} \qquad \text{for all pairs with } d_G(u,v) \geq D^*. \tag{50}$$

For deep MPNNs such bounds hold because node v can influence node u only after $d_G(u,v)$ rounds of message passing, each contributing a sub-unit gain. Our first result shows that the framelet channels of SHEAR couple co-clustered nodes with a gain that carries no distance dependence at all.

Lemma 12 (Positive-diagonal propagation)**.** *Let $\hat{A}_I = (\bar{D}+I)^{-1/2}(A+I)(\bar{D}+I)^{-1/2}$ be the self-looped symmetrically normalized adjacency, with $\bar{D} = \mathrm{diag}(A\mathbf{1})$. Then $\hat{A}_I \geq 0$ entrywise and $(\hat{A}_I)_{vv} = 1/(1+\deg v) \geq 1/(1+d_{\max})$.*

Proof. Nonnegativity is immediate from $A + I \geq 0$ and $\bar{D} + I \succ 0$; the diagonal entry is $(A + I)_{vv}/(1+\deg v) = 1/(1+\deg v)$. □

Theorem 5 (Distance-independent sensitivity through the framelet channels)**.** *Let $u, v \in V$ lie in the same level-ℓ cluster, and let $Z_\ell = \Upsilon_\ell \hat{A}_I X$ be the level-ℓ scaling channel of the channel bank. Then*

$$\Big\| \frac{\partial (Z_\ell)_u}{\partial X_v} \Big\|_2 \;=\; \frac{1}{|C_\ell(u)|} \sum_{w \in C_\ell(u)} (\hat{A}_I)_{wv} \;\geq\; \frac{1}{s_\ell\,(1+d_{\max})} \;>\; 0. \tag{51}$$

The bound depends on the cluster size and the degree distribution only, not on $d_G(u,v)$, not on the number of levels, and not on any depth parameter.

Proof. By Lemma 11, per feature coordinate $\partial (Z_\ell)_u / \partial X_v = (\Upsilon_\ell \hat{A}_I)_{uv} = |C_\ell(u)|^{-1} \sum_{w \in C_\ell(u)} (\hat{A}_I)_{wv}$. By Lemma 12 every summand is nonnegative and the summand $w = v$ alone contributes $(\hat{A}_I)_{vv} \geq 1/(1+d_{\max})$, giving Eq. 51. □

Corollary 2 (Hyperparameter-only bound under the hard cap)**.** *When the hard size cap of Section 4.2(d) is active, every cluster of every extracted level satisfies $|C_a^{(\ell)}| \leq \kappa h$, so Eq. 51 holds with $s_\ell = \kappa h$: the worst-case intra-cluster sensitivity is at least $1/(\kappa h(1+d_{\max}))$, a constant determined by hyperparameters and the degree distribution, independent of both $d_G(u,v)$ and N.*

Remark 5 (Transfer through the fusion)**.** *The attention fusion (Section 4.8)* concatenates *the attended channel blocks $\gamma_k\,\mathrm{fc1}_k(Z_k)$ before the final layers. Concatenation places each channel on disjoint coordinates, so no other channel can cancel the Jacobian block of Theorem 5, and the softmax weights satisfy $\gamma_k > 0$ for all finite attention logits. Hence, whenever the coordinates supporting the block are not clipped by the ReLU, the model Jacobian inherits the bound up to the readout factor $\gamma_k\,\sigma_{\min}(\mathrm{fc1}_k)$. This linearization of the readout is the standing simplification of the sensitivity analyses of Topping et al. (2022); Di Giovanni et al. (2023).*

Theorem 5 covers co-clustered pairs directly. For arbitrary pairs, the next result shows that the hierarchy replaces the fine-graph distance by the coarse-graph distance along an explicit computational pathway with gains bounded away from zero.

Theorem 6 (Logarithmic hierarchical pathways for arbitrary pairs)**.** *Let G be connected and $u, v \in V$ arbitrary. There exists a computational pathway in the forward pass of SHEAR along which X_v influences the representation of u through at most*

$$2L + d_{G_L}\big(C_L(u), C_L(v)\big) \tag{52}$$

composed mixing operations, where $L \leq \lceil \log_h N \rceil$ and d_{G_L} is the shortest-path distance in the coarsest coarse graph. Every operation along the pathway has a gain bounded below by a positive quantity independent of $d_G(u,v)$: the $2L$ pooling/prolongation steps have gains $\geq 1/s_\ell$ and $= 1$ respectively, and each coarse diffusion step has a residual identity branch with Jacobian gain $\geq 1 - \tau_{L,t}\,\rho_{L,t} > 0$ whenever $\tau_{L,t}\rho_{L,t} < 1$, which holds for sufficiently small step sizes at any bounded $\rho_{L,t}$ and, in particular, in the initialization regime ($\tau_{L,t} = 0.1$), where $\rho_{L,t} = \|\Delta_{\mathcal{F}^{(L)}}\|\,\|W_1^{(L,t)}\|\,\|W_2^{(L,t)}\|$. Consequently the sensitivity along the pathway is lower-bounded by a

function of the coarse distance $d_{G_L} \le d_G(u,v)$ rather than of $d_G(u,v)$; whenever the tree contracts distances geometrically (mean contraction factor $r > 1$ per level), $d_{G_L} \le \lceil d_G(u,v)/r^L \rceil = O(1)$ and the lower bound is distance-independent, so the oversquashing criterion of Eq. 50 cannot hold for pairs with $d_G(u,v) \gg \log_h N$.

Proof. Pathway. (i) X_v enters the coarse representation of its cluster through the pooling chain $v \mapsto C_1(v) \mapsto \cdots \mapsto C_L(v)$: L averaging operations, each coupling all members of a cluster with weight exactly $1/|C| \ge 1/s_\ell > 0$ independently of intra-cluster distances (Lemma 11). (ii) On G_L, one mixing step (a diffusion step of Eq. 7 at the coarsest level) advances influence by one coarse hop: the Galerkin block $B^{(L)}_{ab}$ of the coarse sheaf Laplacian accumulates contributions over original edges between clusters a and b only, so the block-sparsity pattern of $\Delta_{\mathcal{F}^{(L)}}$ is contained in that of the coarse adjacency $A^{(L)}$, and one application couples coarse neighbors only. Hence $d_{G_L}(C_L(u), C_L(v))$ steps carry the influence from $C_L(v)$ to $C_L(u)$. (iii) The literal prolongation chain of Eq. 47 copies the coarse signal back to every member of $C_L(u)$, in particular to u: L further operations with gain 1 (copy). The total number of composed mixing operations is Eq. 52.

Gains. The pooling gains are $1/|C_a| \ge 1/s_\ell$ and the prolongation gains are 1. One coarse diffusion step is $H \mapsto H - \tau \tanh(\Delta_{\mathcal{F}}(I \otimes W_1)HW_2)$; its Jacobian at any point is $I - \tau M$ with $\|M\| \le \|\Delta_{\mathcal{F}}\| \, \|W_1\| \, \|W_2\| = \rho$ (using $|\tanh'| \le 1$), hence the step's smallest singular value is at least $1 - \tau\rho$, positive under the stated condition. The composed gain along the pathway is therefore at least

$$\Big(\prod_{\ell=1}^{L} \tfrac{1}{s_\ell}\Big)\,\big(1 - \max_t \tau_{L,t}\rho_{L,t}\big)^{d_{G_L}}\,, \tag{53}$$

a function of the level sizes and the coarse distance only.

Distance compression. Projecting a shortest u–v path of G onto the level-L clusters and deleting consecutive repetitions yields a walk in G_L of length at most $d_G(u,v)$, so $d_{G_L}(C_L(u), C_L(v)) \le d_G(u,v)$. If each level contracts path lengths by a factor $r > 1$ on average, $d_{G_L} \le \lceil d_G(u,v)/r^L \rceil$, which is $O(1)$ for $L = \Theta(\log_h N)$ and $d_G(u,v) = \omega(\log_h N)$. In that regime the sensitivity lower bound is a positive constant independent of $d_G(u,v)$, whereas Eq. 50 forces it below $\sigma^{d_G(u,v)} \to 0$; the two are incompatible, so the model does not oversquash. □

Remark 6 (Scope and comparison with HMH)**.** *(i) On graphs without geometric distance contraction (e.g. paths, where $d_{G_L} \approx d_G(u,v)/\bar{s}$ for average cluster size $\bar{s}$), the lower bound still decays with the (compressed) distance; no hierarchical method can avoid this, and the oversquashing analysis of HMH (Hossen, 2026, Thm. 5.5, Case I) relies on the same contraction assumption through their path-length estimate $k(\ell) \le \lceil d/r^\ell \rceil$. (ii) HMH's lower bound additionally requires their signed encoder to amplify every heterophilous edge by a uniform factor > 1 and their Haar detail gains to exceed one; the product form of their bound then* grows *with depth. Theorem 5 and Corollary 2 need no amplification assumption: pooling-based coupling is distance-independent structurally, with the bound governed by cluster sizes rather than filter gains. (iii) The positive contribution in Eq. 51 is delivered by the diagonal term $(\hat{A}_I)_{vv}$: the bound certifies that the multi-scale pathway introduces no* additional *distance-induced decay beyond the self-information every node already retains, rather than a large absolute sensitivity. Since GBSC balls are connected BFS balls, co-clustered pairs are short-range by construction; the genuinely long-range regime is covered by Theorem 6. (iv) The pathway gain of Theorem 6 is exponentially small in L (of order $(\kappa h)^{-L}$ under the hard cap), so it is a qualitative, existence-type refutation of the $\sigma^{d_G(u,v)}$ decay criterion of Eq. 50, not a practically large sensitivity certificate; in the deployed model the effective sensitivity is carried by the propagation bank and the direct detail channels (Section 4.7).*

C.2 Oversmoothing

Oversmoothing is the collapse of node representations with depth, quantified by the Dirichlet energy $\mathcal{E}(H) = \sum_{(u,v)\in E} \|H_u - H_v\|^2$ decaying exponentially in the number of layers (Cai & Wang, 2020; Bodnar et al., 2022). The term is used with several non-equivalent meanings in the literature (Rusch et al., 2023); throughout, we adopt the depth-induced exponential-decay notion above. We show that SHEAR is protected at three levels: the multi-scale channel bank is an exact isometry, no smoothing operator is iterated along a depth axis, and the fusion cannot cancel surviving contrast.

Lemma 13 (Orthogonal telescoping of the channel bank)**.** *The projectors satisfy*

$$\Upsilon_L + \sum_{\ell=1}^{L} \Xi_\ell \;=\; I, \tag{54}$$

all terms are pairwise-orthogonal symmetric projectors, and for every signal x,

$$\|\Upsilon_L x\|^2 + \sum_{\ell=1}^{L} \|\Xi_\ell x\|^2 \;=\; \|x\|^2. \tag{55}$$

Proof. Nestedness $V_{\ell-1} \supset V_\ell$ gives $\Upsilon_{\ell-1}\Upsilon_\ell = \Upsilon_\ell\Upsilon_{\ell-1} = \Upsilon_\ell$, so $\Xi_\ell = \Upsilon_{\ell-1} - \Upsilon_\ell$ is the orthogonal projector onto $V_{\ell-1} \ominus V_\ell$; distinct detail spaces are orthogonal, and $\mathcal{W}_\ell = V_{\ell-1} \ominus V_\ell \perp V_L$, so $\Xi_\ell\Upsilon_L = 0$. The sum in Eq. 54 telescopes to $\Upsilon_0 = I$. Equation 55 is the Pythagorean theorem for this orthogonal splitting. □

Corollary 3 (Contrast survival)**.** *If x is not constant on the coarsest clusters ($x \notin V_L$), then $\max_\ell \|\Xi_\ell x\|^2 \geq \|(I - \Upsilon_L)x\|^2/L$: at least one detail channel retains at least a $1/L$ fraction of the energy that x has outside the coarsest cluster-constant subspace. In particular, the contrast of every edge whose endpoints are separated at some level is carried by the corresponding detail channel.*

Proof. By Eq. 55, $\sum_\ell \|\Xi_\ell x\|^2 = \|(I - \Upsilon_L)x\|^2$, and the maximum of L nonnegative numbers is at least their average. □

Theorem 7 (No asymptotic oversmoothing)**.** *Let $X \mapsto H(X;\Theta)$ be the fused representation of SHEAR. Then:* (i) *the channel bank $\{\Upsilon_L x, \Xi_1 x, \ldots, \Xi_L x\}$ applied to $x = \hat{A}_I X$ is an isometry of its input (Lemma 13), so the multi-scale pathway loses no signal energy at all; the dilated variant used in Table 3 inherits this exactly, since a stride-s detail is a sum of adjacent ones, $\Xi_\ell^{(s)} := \Upsilon_{\ell-s} - \Upsilon_\ell = \sum_{\ell-s<j\leq\ell} \Xi_j$, and the telescoping over the subsampled levels still gives $\Upsilon_L + \sum_\ell \Xi_\ell^{(s)} = I$;* (ii) *every smoothing-type operator in the network (the propagation channels $\hat{A}^k X$ with $k \leq k_{\max}$ fixed, the T_0 fine and T_c per-level coarse diffusion steps, and the framelet projections) is applied a number of times prescribed by the architecture and independent of any depth parameter, and no such operator is composed with itself along a growing depth; hence there is no depth axis along which $\mathcal{E}(H) \to 0$, and decay bounds of the form $\mathcal{E}(H^{(L)}) \leq c^L \mathcal{E}(X)$ with $c < 1$ do not apply to SHEAR, whose depth is fixed by the architecture;* (iii) *with end-to-end learnable restriction maps, the sheaf kernel $\ker \Delta_{\mathcal{F}} = \{x : F_{u\to e}x_u = F_{v\to e}x_v \;\forall e\}$ is itself learned and generically does not contain the constant vector, so even the diffusion steps need not equilibrate node signals (Bodnar et al., 2022).*

Proof. (i) restates Lemma 13 applied coordinate-wise. (ii) In the forward pass of Section 4.7, each propagation channel $\hat{A}^k$ is applied directly to X (the channels are parallel, not sequential); the diffusion of Eq. 7 runs a fixed budget of $T_0 + LT_c$ steps with increments bounded by $\tanh$ saturation and learnable step sizes initialized at 0.1; and each $\mathcal{S}_\ell, \mathcal{W}_\ell$ is an orthogonal projector applied once. The only iterated compositions are the L pooling/prolongation pairs of the chains, which satisfy $\mathrm{pool}_\ell \,\mathrm{prolong}_\ell = I$: they reproduce cluster means exactly rather than smoothing progressively. The representation is therefore a fixed-depth parallel ensemble whose total smoothing budget is bounded a priori by the architecture. Moreover, concatenation places the attended channel blocks on disjoint coordinates, so the Dirichlet energy of the fused representation decomposes as $\mathcal{E}(H_{\mathrm{cat}}) = \sum_k \mathcal{E}\big(\gamma_k \,\mathrm{fc1}_k(Z_k)\big) \;\geq\; \gamma_c^2 \,\sigma_{\min}(W_{\mathrm{fc1},c})^2 \,\mathcal{E}(Z_c)$ for every channel c: it is bounded below by the (strictly positive) contribution of any single surviving channel, not by an exponentially decaying function of a depth parameter. (Theorem 8 transfers class-mean contrast, rather than energy, through the same fusion.) (iii) The kernel condition $F_{u\to e}x_u = F_{v\to e}x_v$ depends on the learned maps; for generic learned maps no nonzero constant signal satisfies it on all edges, so the diffusion is not constrained to converge towards a constant equilibrium and can preserve or amplify heterophilous contrast, as analyzed by Bodnar et al. (2022). □

Theorem 8 (Class-mean contrast through the fusion)**.** *Let $\mathcal{K}_1, \mathcal{K}_2 \subset V$ be two node classes with class means $\mu_1(Z), \mu_2(Z)$ of a representation Z, and suppose some fused channel Z_c retains contrast $\|\mu_1(Z_c) - \mu_2(Z_c)\| \geq \delta > 0$. Then, with the fusion layers linearized (the per-node ℓ_2 normalization of Section 4.8 omitted, the standard readout simplification of the sensitivity analyses of Topping et al. (2022); Di Giovanni et al. (2023)), the concatenated fused representation $H_{\mathrm{cat}} = [\gamma_k \,\mathrm{fc1}_k(Z_k)]_k$ satisfies, whenever the coordinates supporting the contrast are not clipped by the ReLU,*

$$\|\mu_1(H_{\mathrm{cat}}) - \mu_2(H_{\mathrm{cat}})\| \;\geq\; \gamma_c \,\sigma_{\min}(W_{\mathrm{fc1},c})\,\delta \;>\; 0, \tag{56}$$

with $\gamma_c > 0$ for all finite attention logits. The bound is independent of the number of levels, diffusion steps, and propagation hops.

Proof. We treat $\mathrm{fc1}_c$ as its linear part $W_{\mathrm{fc1},c}$ under the stated linearization. Concatenation places the block $\gamma_c \,\mathrm{fc1}_c(Z_c)$ on disjoint coordinates, so the class-mean difference of H_{cat} contains the block $\gamma_c \, W_{\mathrm{fc1},c}\big(\mu_1(Z_c) - \mu_2(Z_c)\big)$ (class means are row-wise and commute with the row-wise linear map). Its norm is at least $\gamma_c \,\sigma_{\min}(W_{\mathrm{fc1},c})\,\delta$, and softmax weights are exponentials of finite logits, hence strictly positive. □

Remark 7 (Row-wise normalization)**.** *The linearization omits a row-wise rescaling only: $\mathcal{N}$ maps every node vector to its unit-norm direction, so it cannot create class-mean contrast, only rescale it per node. The bound of Theorem 8 therefore transfers to the normalized readout with δ replaced by the contrast of the normalized channel; the two contrasts coincide up to the spread of the row norms and differ substantially only when the pre-normalized class means are nearly collinear with very different norms.*

Remark 8 (Scope)**.** *(i) Theorem 7 is an architecture-level statement: it rules out depth-induced energy collapse, not task-level failure. A channel that survives with its energy intact can still be uninformative, or add variance the data cannot support, cf. the Texas ablation of Section G, where the detail channels overfit a 183-node graph. (ii) Our analogue of the oversmoothing case of Hossen (2026, Thm. 5.5) differs in kind: their argument bounds the* ratio *of the spoke–spoke to the spoke–hub class-mean distance below by* 1*, which does not by itself preclude joint decay of both distances. Lemma 13 and Theorem 7 hold at the channel level without any assumption on the learned parameters (exact isometry, fixed smoothing budget, learnable kernel), and Theorem 8 transfers the surviving contrast to the fused representation with a strictly positive factor. (iii) The isometry of Theorem 7(i) covers the multi-scale (framelet) pathway only. The propagation bank $\{\hat{A}^k X, \hat{A}_I^k X\}_{k \leq k_{\max}}$ is not an isometry: its protection is the fixed smoothing budget of part (ii), under which each channel applies a single polynomial of degree at most $k_{\max}$ and no operator is iterated along a growing depth, together with the $k{=}0$ channel, which is the identity. Within a single propagation channel the decay is explicit: writing $\hat{A} = I - \hat{L}$ in the eigenbasis of the normalized Laplacian $\hat{L}$ with eigenpairs $(\lambda_i, \tilde{x}_i)$, $\lambda_i \in [0, 2]$,*

$$\mathcal{E}(\hat{A}^k X) \;=\; \sum_i \lambda_i \,(1 - \lambda_i)^{2k} \,\|\tilde{x}_i\|^2, \tag{57}$$

which is non-increasing in k since $|1 - \lambda_i| \leq 1$, and whose $k{=}k_{\max}$ value is bounded below by a constant determined by the spectrum and the fixed hop bound alone, not by any depth parameter; the self-looped family satisfies the same bound with the spectrum of $\hat{L}_I = I - \hat{A}_I$. Part (ii) is accordingly an architectural statement: the model has no depth axis, rather than a decay bound along one. Theorem 8 is conditional: it transfers contrast only when some channel retains a class-mean contrast $\delta > 0$ and the coordinates supporting it are not clipped by the ReLU.

D Benchmark Comparison on the PEGFAN Protocol

Table 2, deferred from Section 5.1, collects the seven PEGFAN-protocol columns of the benchmark comparison.

E Per-dataset Hyperparameter Configurations

Training settings. Stalk dimension $d = 8$, sheaf channels $m = 16$, at most $L = 3$ coarsening levels, diffusion steps $T_0 = 2$, $T_c = 1$, hidden width 64, dropout 0.5, early stopping on the validation

Table 2: Test performance (%, mean ± std) on the seven datasets of Li et al. (2024, Table III) under the ten fixed Geom-GCN $48/32/20$ splits (Film is the Actor graph of Tang et al. (2009)). Numbers of the quoted rows are taken from Li et al. (2024), with FSGNN and PEGFAN each reporting, per column, the best over their published variants. —: not reported under this protocol. **Bold**: best; underline: second best.

	Homophilous			Heterophilous			
	Cora	Citeseer	Pubmed	Texas	Cornell	Wisconsin	Film
Mixhop	87.61±0.85	76.26±1.33	85.31±0.61	77.84±7.73	73.51±6.34	75.88±4.90	32.22±2.34
GEOM-GCN	85.27	**77.99**	90.05	67.57	60.81	64.12	31.63
GCNII	88.01±1.33	77.13±1.38	**90.30±0.37**	77.84±5.64	76.49±4.37	81.57±4.98	—
H_2GCN	86.92±1.37	77.07±1.64	89.40±0.34	84.86±6.77	82.16±4.80	86.67±4.69	35.86±1.03
WRGAT	88.20±2.26	76.81±1.89	88.52±0.92	83.62±5.50	81.62±3.90	86.98±3.78	**36.53±0.77**
GPRGNN	**88.49±0.95**	77.08±1.63	88.99±0.40	86.49±4.83	81.89±6.17	85.88±3.70	36.04±0.96
FSGNN	88.15±1.15	77.23±1.41	89.73±0.40	86.76±3.72	86.22±6.78	**87.84±3.37**	35.51±0.89
PEGFAN	87.36±1.09	77.39±1.28	89.73±0.40	86.47±5.54	**86.76±5.33**	87.65±4.02	35.48±0.94
SHEAR (ours)	86.46±1.28	77.04±1.71	89.24±0.42	**87.30±5.28**	85.41±6.19	87.25±3.07	35.49±1.09

loss (patience 100, at most 1500 epochs). All tasks are transductive: the cross-entropy loss on the predictions of Eq. 8 is computed on the train-mask nodes only, and all restriction maps are generated end-to-end by the shared SheafMLP (Section 4.5). The propagation channels are the full multi-hop bank of Li et al. (2024) on every dataset: X together with $\hat{A}^k X$ and $\hat{A}_I^k X$ for $k = 1, \ldots, k_{\max}$, $k_{\max} = 3$ ($k_{\max} = 5$ on the two filtered benchmarks). The Select-stage (h, α), the detail stride $s \in \{1, 2\}$, and the fusion-side hyperparameters $(w_{\text{att}}, \text{lr}_{\text{att}}, \text{lr}_{\text{fc}}, \text{dropout})$ are selected per dataset by validation on a 3-split grid search, following Li et al. (2024), and retrained on the full 10-split protocol; the acceptance margin is fixed to $\varepsilon = 0.05$ with $\omega_1 = \omega_2 = \omega_3 = 1$, $\beta = 0$ unless listed otherwise in Table 3, and the hard cap $\kappa = 8$ is enabled per dataset as listed there; the Geom-GCN Squirrel/Chameleon configurations used in the studies of Sections F.1–F.2 and G retain the transferred fusion presets analyzed there. All experiments use a single NVIDIA RTX 4090-class GPU; baseline numbers are quoted from Li et al. (2024) and Hossen (2026).

Table 3 lists the selected per-dataset configurations behind the reference results of Tables 1 and 2 and the studies of Sections F.1–F.2 and G, and Table 4 reports the realized hierarchies of the reference configurations. The search grids are: cluster-size bound $h \in \{4, 8, 16, 32\}$, feature weight $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$, attention weight decay $w_{\text{att}} \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}$, attention learning rate in $\{0.04, 0.02, 0.01\}$, detail stride $s \in \{1, 2\}$, sheaf learning rate $\text{lr}_{\text{sh}} \in \{0.01, 0.02, 0.04\}$, propagation depth $k_{\max} \in \{3, \ldots, 10\}$, and dropout in $\{0.5, 0.6\}$; the optimizer has four parameter groups (fusion layers, attention, and the sheaf/framelet parameters $\{\text{lift}, \Theta, E, P, P_{\text{back}}, W_1, W_2, \tau, \text{SheafMLP}\}$). The remaining shared settings are those of the training-settings paragraph above.

F Sensitivity and Ablation Studies

Tables 5, 6, 7, and 8 collect the full results behind the studies of Sections F.1–F.2 and G.

F.1 Sensitivity analysis

We examine how sensitive the benchmark results are to the design choices of the Select stage. Table 6 scans the four Select-stage hyperparameters (the cluster-size bound h, the topology–feature seed weight α of Eq. 1, the acceptance margin ε of Eq. 3, and the two-hop Jaccard seed weight β) on Squirrel, Chameleon, and Texas, using the clustering-only configuration (the full propagation-channel bank, no detail channels, no size cap, default fusion hyperparameters) so that the scan isolates the quality of the hierarchy itself. Each row varies one hyperparameter while the others sit at their reference values (h=8 on Squirrel/Chameleon, h=4 on Texas, α=0.5, ε=0.05, β=0); the scan centers are the clustering-only defaults and need not coincide with the deployed per-dataset values of Table 3, which are selected on the full architecture by validation.

Table 3: Selected hyperparameters behind the results of Tables 1 and 2 (upper block) and the Geom-GCN studies of Sections F.1–F.2 and G (lower block; GG = unfiltered Geom-GCN version). Shared settings: stalk dimension d=8, sheaf channels m=16, L=3 levels, T_0=2, T_c=1, hidden width 64, ε=0.05, β=0, ω_1=ω_2=ω_3=1, sheaf-parameter group $\mathrm{lr}_{\mathrm{sh}}$=0.02 with weight decay $5 \cdot 10^{-4}$, early-stopping patience 100. The SheafMLP is a two-layer perceptron with a single hidden layer of width 32 (ReLU), mapping the concatenated endpoint features ($2\,d\,m_0$) to the two $d \times d$ restriction maps per edge. w_{att}, $\mathrm{lr}_{\mathrm{att}}$, $\mathrm{lr}_{\mathrm{fc}}$ and dropout are the fusion-side parameters of Section 4.8; w_{fc1}=w_{fc2}=$5 \cdot 10^{-4}$ everywhere except on the four Squirrel/Chameleon configurations, which set them to 0; the Geom-GCN Chameleon configuration additionally pairs the preset with a coarser full-depth detail tree (h_{detail}=8). The two filtered configurations additionally disable the row-wise ℓ_2 normalization $\mathcal{N}$ of Eq. 8 and replace the single detail tree by three parallel GBSC detail trees ($h_{\mathrm{detail}} \in \{4, 8, 16\}$, full depth on Chameleon) whose projectors, subsampled with dilation stride s=2, act on the signals $\{X, \hat{A}X, \hat{A}^2X\}$ ($\hat{A}^3X$ added on Squirrel); the filtered Squirrel row additionally uses hidden width 128, the two filtered rows set $\mathrm{lr}_{\mathrm{sh}}$=0.01 (Squirrel) and $\mathrm{lr}_{\mathrm{sh}}$=0.04 (Chameleon), and the propagation bank is extended to $k_{\max}$=5 on both filtered rows. *cap*: hard ball-size cap $8h$ on the Select tree; *det.*: detail channels enabled; s: dilation stride of the detail channels (s=1: adjacent levels; s=2: every second level, the coarsest level always retained; on a realized tree with a single level, Citeseer in Table 4, a stride-2 subsequence coincides with stride 1).

Dataset	h	α	β	w_{att}	$\mathrm{lr}_{\mathrm{att}}$	$\mathrm{lr}_{\mathrm{fc}}$	dropout	cap	det.	s
Cora	16	0.5	0	0	0.02	0.02	0.5	√	√	2
Citeseer	32	0.75	0	10^{-1}	0.02	0.02	0.6	√	√	2
Pubmed	16	0.0	0	0	0.02	0.02	0.5	√	√	2
Texas	16	0.0	0	10^{-3}	0.02	0.02	0.6	—	—	—
Cornell	4	1.0	0	10^{-3}	0.01	0.02	0.6	—	√	2
Wisconsin	8	0.25	0	10^{-3}	0.04	0.02	0.5	—	√	2
Film	8	0.25	0	10^{-3}	0.04	0.02	0.5	—	√	2
Minesweeper	32	0.25	0	$5 \cdot 10^{-4}$	0.02	0.02	0.5	√	√	1
Tolokers	64	0.0	0.1	0	0.02	0.02	0.5	√	—	—
Roman-empire	32	1.0	0	0	0.02	0.02	0.5	√	√	1
Amazon-ratings	32	0.0	0.1	0	0.02	0.02	0.5	√	√	1
Squirrel (filt.)	32	0.5	0.1	10^{-1}	0.04	0.02	0.5	√	√	2
Chameleon (filt.)	8	0.0	0.1	10^{-1}	0.02	0.02	0.4	√	√	2
Squirrel (GG)	32	0.5	0	$5 \cdot 10^{-2}$	0.04	0.01	0.5	√	√	1
Chameleon (GG)	4	0.0	0	10^{-1}	0.02	0.005	0.6	√	√	1

Three findings stand out. First, the model is *insensitive to the nominal granularity*: across $h \in \{4, 8, 16, 32\}$ the results span about 1.4 points on Squirrel and 0.8 points on Chameleon, because the quality gate and the milestone level extraction adapt the tree to whatever target h prescribes. Second, the topology–feature coupling of the seed rule is mildly beneficial but never delicate: the whole α range spans 1.2 points on Squirrel and 1.4 on Chameleon, and only the pure-feature extreme α=1 degrades on Texas (79.2 versus 84.1 at α=0.5), where the structural terms are essential; the acceptance margin is likewise flat across $\varepsilon \in [0, 0.1]$ (the largest movement, 1.7 points on Squirrel, stays below the split-to-split standard deviations of 2.6–3.6 on that dataset, Table 6), supporting the fixed choice $\varepsilon = 0.05$ (the margin is commensurate with the three quality components, all $O(1)$ quantities: cosine similarity in $[-1, 1]$, normalized purity in $[0, 1]$, and a modularity contribution bounded by $e_{\mathrm{in}}/|E| \leq 1$). Third, the two-hop Jaccard seed term does not pay off on these benchmarks (within noise in both directions), which justifies the default $\beta = 0$; on the filtered Squirrel/Chameleon of Table 1 a small weight $\beta = 0.1$ was selected by validation and is used there (Table 3): on the unfiltered graphs the duplicate nodes inflate the two-hop neighborhoods and add noise to the Jaccard term, while on the deduplicated graphs the structural signal is clean enough for a small weight to pay off.

F.2 Ablation: the role of the sheaf machinery

The second study ablates the sheaf components on Squirrel and Chameleon, the two larger Geom-GCN heterophilous graphs, under the corresponding reference configurations of Table 3. Three find-

Table 4: Realized GBSC hierarchies behind the reference configurations of Table 3: the node counts N_ℓ of the coarse levels used by the model (at most L=3 coarsening levels), read off the cached partition trees. The partition tree is a deterministic function of (G, X) and the Select-stage hyperparameters (Section 5.1), so these counts are identical across splits and seeds. The stalk dimension d=8 and the channel count m_ℓ=16 are the same at every level and on every dataset. The Film tree has a fourth milestone level (2 clusters) beyond the L=3 cap.

Dataset	N_0	N_1	N_2	N_3	L
Cora	2708	170	78	—	2
Citeseer	3327	438	—	—	1
Pubmed	19717	1233	227	—	2
Texas	183	12	—	—	1
Cornell	183	46	12	3	3
Wisconsin	251	32	4	—	2
Film	7600	950	119	15	3
Minesweeper	10000	313	63	—	2
Tolokers	11758	184	31	—	2
Roman-empire	22662	709	135	—	2
Amazon-ratings	24492	766	148	—	2
Squirrel (filt.)	2223	70	16	—	2
Chameleon (filt.)	890	112	19	—	2

ings support the design. Reducing the stalks to scalars (d=1), which collapses the sheaf Laplacian to a learned weighted scalar Laplacian, costs 1.1 points on Squirrel, the largest single-hyperparameter degradation on that dataset, and 0.5 on Chameleon, so the vector-bundle structure carries a real, if modest, share of the accuracy. Freezing pretrained restriction maps instead of regenerating them end-to-end moves the mean by less than 0.6 points on both datasets, so the SheafMLP removes the pretraining stage and the split-dependent cached operators at no accuracy cost. The architecture is also insensitive to the sheaf capacity: accuracy spans 1.9 points over $d \in \{1, 2, 4, 8, 16\}$ and 0.6 points over $m \in \{4, 16, 32\}$ on Squirrel (at most 0.7 in both scans on Chameleon), which supports the defaults d=8, m=16 as a capacity–variance compromise (m=32 is marginally better on Squirrel, m=16 on Chameleon). We note honestly that switching the diffusion off (T_0=T_c=0) leaves the mean unchanged on both datasets: on these benchmarks the propagation and detail channels already supply the long-range mixing, so the diffusion steps are complementary rather than essential, and the load-bearing role of the sheaf machinery is structural, namely making the multi-scale Galerkin coarsening end-to-end and permutation equivariant and carrying the guarantees of Section 4.10.

Table 5: Test performance (%, mean ± std over ten splits) behind the comparison of Section G: the clustering-only GBSC configuration (full propagation-channel bank, detail channels disabled, Haar framelet bands, end-to-end sheaf, default fusion hyperparameters) at the per-dataset (h, α) (Minesweeper $(32, 0.25)$, Tolokers $(64, 0.0)$, Roman-empire $(32, 0.75)$, Amazon-ratings $(32, 0.0)$, with β=0.1 on Tolokers and Amazon-ratings) with the deployed $8h$ size cap, versus the Reference column, which repeats the SHEAR row of Table 1 (full architecture, per-dataset configurations of Table 3). Minesweeper and Tolokers are ROC-AUC, all other columns accuracy. Δ = Reference − GBSC (clustering-only). —: no clustering-only GBSC run was made on the filtered benchmarks. Tolokers: the Reference column is the detail-free configuration ($(h, \alpha) = (64, 0)$, Table 3).

Dataset	GBSC (cl.-only)	Reference	Δ
Minesweeper	88.73±0.87	88.54±0.69	−0.19
Tolokers	81.99±1.01	82.20±1.04	+0.21
Roman-empire	89.07±0.34	89.04±0.46	−0.03
Amazon-ratings	45.82±1.13	46.47±0.52	+0.65
Squirrel (filt.)	—	40.54±1.25	—
Chameleon (filt.)	—	45.28±2.22	—

F.3 Diagnostics behind Section G

The full architecture adds the detail channels, the sheaf pathway, and per-dataset tuning on top of the clustering-only GBSC configuration. On every dataset with a clustering-only run in Table 5, the

Table 6: Test accuracy (%, mean ± std over the ten Geom-GCN splits) when varying one Select-stage hyperparameter at a time, clustering-only configuration. The center row of every group (h=8 on Squirrel/Chameleon, h=4 on Texas, α=0.5, ε=0.05, β=0) is a single shared run per dataset.

Hyperparameter	Value	Squirrel	Chameleon	Texas
h	4	48.28±3.54	71.16±1.57	84.05±5.05
	8	48.72±2.86	72.00±1.31	—
	16	48.77±2.65	71.29±1.34	—
	32	49.64±2.52	71.29±1.21	—
α	0	49.00±3.71	71.07±1.05	82.70±5.69
	0.25	48.64±2.60	70.59±1.67	82.16±6.75
	0.5	48.72±2.86	72.00±1.31	84.05±5.05
	0.75	47.76±2.38	71.36±1.35	83.24±5.38
	1	48.71±2.40	71.32±1.21	79.19±10.40
ε	0	48.28±2.64	71.34±1.33	—
	0.02	49.94±3.55	71.23±1.71	—
	0.05	48.72±2.86	72.00±1.31	—
	0.1	49.24±2.97	71.25±1.25	—
β	0	48.72±2.86	72.00±1.31	—
	0.1	47.59±2.63	—	—
	0.25	48.22±3.78	71.36±1.43	—

reference architecture stays within 0.7 points of the clustering-only GBSC numbers, with deviations in both directions that sit well inside the split-to-split standard deviation. Only Texas and Tolokers disable the detail channels entirely: there they add capacity that the data cannot support, since Texas has 183 nodes, so both adopt the detail-free configuration (Table 3).

Two design decisions underlie the results on the heterophilous graphs; both are diagnosed on the unfiltered Geom-GCN study versions introduced in Section 5.1. First, widening the per-level frequency resolution of the analysis cascade does not help on its own (three-band sheaflet spectral bands reach 49.99 / 70.29 and an adaptively bisected Haar high band 49.03 / 70.18 on Squirrel / Chameleon, against 49.73 / 70.07 for the binary Haar baseline in the same study configuration, all within the split-to-split standard deviations), because the mean-pool of the analysis cascade annihilates the projector high band exactly ($\mathrm{pool} \circ \mathrm{prolong} = I$ implies $\mathrm{pool} \circ \mathcal{W}_\ell = 0$), so every analysis-path high-frequency component is discarded regardless of how finely the band is parameterized, which makes the learnable high-band gain $\theta_{\ell,\mathrm{hp}}$ of the analysis cascade a dead parameter on the analysis path (the two-band form is kept so that the analysis and synthesis cascades share one operator structure; the synthesis gains $\eta_{\ell,\mathrm{hp}}$ remain effective); this is precisely the regime that Li et al. (2026) identify as harmful, and their high-pass-removal ablation mirrors the gap we observe. Second, exposing the per-level detail content directly through the channels $\Xi_\ell x = (\Upsilon_{\ell-1} - \Upsilon_\ell)x$, $\ell = 1, \ldots, L$, and $\Upsilon_L x$, the framelet channels in operator form built from our equivariant GBSC tree, lifts the same-batch clustering-only baseline from 49.7 to 56.5 on Squirrel and from 70.1 to 71.3 on Chameleon (Table 7). Regularizing the channel fusion accordingly (a heavier weight decay on the channel attention, $5 \cdot 10^{-2}$ on Squirrel and 10^{-1} on Chameleon, Table 3) lifts it further to 71.5 / 76.0. Deeper detail trees undo the gain, so the benefit comes from the direct detail pathways rather than from finer multi-resolution. A controlled tree swap shows that the clustering is not the limiting factor: replacing the GBSC tree by a Ward tree inside the detail channels (a diagnostic variant only, since the Ward tree is not permutation invariant) gives 69.72 ± 2.23 / 77.79 ± 1.15 on Squirrel / Chameleon against 70.63 ± 2.42 / 75.37 ± 1.73 with our equivariant tree under matched fusion presets, so the identity of the tree accounts for at most about a point on Squirrel and two and a half on Chameleon, a minor factor next to the fusion presets and the detail pathways. A parallel extension of the detail bank, in which several GBSC trees ($h_{\mathrm{detail}} \in \{4, 8, 16\}$) with projectors subsampled at dilation stride s=2 ($\Upsilon_{\ell-2} - \Upsilon_\ell$ instead of $\Upsilon_{\ell-1} - \Upsilon_\ell$) act on the signal set $\{X, \hat{A}X, \hat{A}^2X\}$ ($\hat{A}^3X$ added on Squirrel), is what the filtered benchmarks of Table 1 use. It lifts Squirrel to 40.54 (a point above HMH's 39.50) and Chameleon to 45.28 (two points above SLOG's 43.02 and four above HMH's 41.20), while remaining exactly permutation equivariant. Extending the bank toward several dilation scales applied full-graph is a structural extension left to future work.

Table 7: Diagnostic numbers of Section G on the unfiltered Geom-GCN study versions (test accuracy %, mean over the ten splits, ± std where recorded). The upper block shares the clustering-only study configuration of Section F.1; the lower block is the matched-preset tree-swap pair.

Configuration	Squirrel (GG)	Chameleon (GG)
Binary Haar bands (study baseline)	49.73	70.07
Three-band sheaflet spectral bands	49.99	70.29
Adaptively bisected Haar high band	49.03	70.18
+ direct detail channels	56.5	71.3
+ regularized channel attention	71.5	76.0
GBSC tree, matched fusion presets	70.63±2.42	75.37±1.73
Ward tree swap (diagnostic, not invariant)	69.72±2.23	77.79±1.15

F.4 Original versus filtered Chameleon/Squirrel

The original Chameleon and Squirrel graphs contain a substantial number of duplicate nodes, and Platonov et al. (2023) show that these duplicates are scattered across the train, validation, and test splits of the standard benchmark: a model can then effectively memorize test nodes during training, which inflates test scores and makes rankings on these two benchmarks unreliable. Their *filtered* versions remove the duplicate nodes, and every model's score drops sharply under the deduplicated evaluation: in Table 8 CoCN, for instance, loses 37.7 points on Chameleon. On the leakage-free filtered versions SHEAR attains the best Chameleon result (45.28, i.e. 3.3 points above CoCN) and the second-best Squirrel result (40.54, 0.54 points behind FAGCN's 41.08, within one standard deviation, and above CoCN's 39.69); FAGCN's advantage comes from its signed attention, a mechanism orthogonal to multi-scale analysis. The original-version columns are included for completeness only: our numbers there are measured on the Geom-GCN $48/32/20$ splits, a different protocol from the quoted rows, so they are not comparable across rows and carry no ranking marks, and we likewise refrain from comparing original→filtered drops across rows, since such drops mix the two different original-side protocols.

Table 8 extends the CoCN comparison of Sun et al. (2023, Table 3) with SHEAR on the original and filtered (deduplicated) versions of Chameleon and Squirrel.

Table 8: Test accuracy (%) on the original and filtered versions of Chameleon and Squirrel, extending Sun et al. (2023, Table 3) with SHEAR (bottom row). Quoted rows are from Sun et al. (2023); the baseline numbers therein originate from Platonov et al. (2023) and are reported without standard deviations. The filtered columns use the official filtered splits of Platonov et al. (2023) ($60/20/20$, ten splits), the same protocol as the corresponding columns of Table 1. **Bold**: best; underline: second best. †: our original-version numbers are measured on the Geom-GCN $48/32/20$ splits (Section 5.1), a different protocol from the quoted rows in these columns, and are excluded from the ranking marks.

	Chameleon		Squirrel	
	Original	Filtered	Original	Filtered
FAGCN	64.23	41.90	47.63	**41.08**
GCN	50.18	40.89	39.06	39.47
ResNet+adj	71.07	38.67	65.46	38.37
FSGNN	77.85	40.61	68.93	35.92
CoCN	**79.61**	41.95	**73.27**	39.69
SHEAR (ours)	76.03±1.03†	**45.28±2.22**	71.52±2.14†	40.54±1.25

G Discussion

The full architecture stays within 0.7 points of its clustering-only GBSC variant on every dataset with such a run, and only Texas and Tolokers disable the detail channels entirely (Tables 5 and 3, Appendix F). Two further diagnostics, reported with full numbers in Appendix F.3 (Table 7), underlie

the heterophilous results. First, widening the per-level frequency resolution of the analysis cascade does not help, because the mean-pool annihilates the projector high band exactly ($\mathrm{pool}\circ\mathrm{prolong} = I$ implies $\mathrm{pool} \circ \mathcal{W}_\ell = 0$); the gains come instead from exposing the per-level detail content directly through the channels $\Xi_\ell x = (\Upsilon_{\ell-1} - \Upsilon_\ell)x$, $\ell = 1, \ldots, L$, and $\Upsilon_L x$, together with a regularized channel fusion. Second, a controlled tree swap (a Ward tree in place of GBSC, a diagnostic variant that is not permutation invariant) changes the result by at most about a point on Squirrel and two and a half on Chameleon, so the clustering is not the limiting factor. A parallel extension of the detail bank, in which several GBSC trees ($h_{\mathrm{detail}} \in \{4, 8, 16\}$) with projectors subsampled at dilation stride s=2 act on the signal set $\{X, \hat{A}X, \hat{A}^2X\}$ ($\hat{A}^3X$ added on Squirrel), is what the filtered benchmarks of Table 1 use, lifting Squirrel to 40.54 and Chameleon to 45.28 while remaining exactly permutation equivariant; extending the bank toward several dilation scales applied full-graph is a structural extension left to future work.

Limitations and reproducibility. Our guarantees are architecture-level: they exclude depth-induced collapse and distance-induced information loss at the channel level but say nothing about task-level failure modes, and Theorem 1 assumes general position of the invariant keys (Remark 3). The hierarchy itself is not learned: the quality-oracle weights and the acceptance margin are fixed a priori; learnable refinement of the hard clusters is left to future work (Section 6). SHEAR also trails the strongest baselines on the three homophilous citation graphs (Section 5.1), and the sheaf components are empirically neutral to modest on our benchmarks (Section F.2): we position them as the structural carrier of the end-to-end equivariant coarsening and of the guarantees of Section 4.10, and identifying regimes in which the sheaf geometry is accuracy-critical is left to future work. All experiments run on a single GPU with the fixed public or seeded splits of Section 5.1; code, configurations, and per-split logs will be released upon publication.