OPERATOR-RELATIVE ENERGY: A UNIFIED FRAMEWORK FOR UNDERSTANDING AND MITIGATING OVER-SMOOTHING IN HYPERGRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated message passing in hypergraph neural networks (HGNNs) can suppress distinctions between node representations, a phenomenon known as *over-smoothing* that limits the benefits of deeper architectures. The resulting form of representational collapse depends on the underlying propagation rule. However, classical measures of *over-smoothing*, such as the Dirichlet energy, can fail to detect collapse when applied across different propagation rules. To address this limitation, we introduce an operator-relative energy that measures representation variation relative to the collapse behavior of a given propagation rule, while recovering classical Dirichlet energies as special cases. Using the proposed energy, we show that representation variation can decay exponentially with network depth across a broad class of hypergraph propagation operators, leading to *over-smoothing*, whereas initial residuals can prevent complete collapse by preserving a nonzero level of variation. Motivated by these theoretical results, we incorporate initial residuals and identity mappings (II) into five representative HGNN architectures to mitigate *over-smoothing*. Controlled propagation experiments on six real-world datasets demonstrate that the classical Dirichlet energy can fail to detect over-smoothing, while our operator-relative energy correctly detects representation collapse and confirms that initial residuals prevent the energy from vanishing. Node-classification experiments further show that the proposed II variants consistently improve predictive accuracy over their corresponding baseline architectures and sustain their performance as network depth increases.

1 INTRODUCTION

Hypergraphs represent interactions among groups of entities, such as coauthorship teams and sets of molecules participating in a reaction (Zhou et al., 2006). Hypergraph neural networks (HGNNs) learn from these interactions by passing messages between nodes and their incident hyperedges (Feng et al., 2019; Dong et al., 2020; Yadati et al., 2019; Huang & Yang, 2021; Chien et al., 2021). Stacking layers allows information to travel across increasingly distant neighborhoods, but repeated aggregation can also suppress distinctions between node representations. This phenomenon, known as *over-smoothing*, can limit the usefulness of deeper message passing (Oono & Suzuki, 2019; Cai & Wang, 2020; Rusch et al., 2023). Understanding this limitation requires characterizing which distinctions a propagation rule removes as depth increases, and whether they can be preserved.

A common approach to analyzing *over-smoothing* is to track a measure of variation in node representations. In particular, Dirichlet-energy analyses establish conditions under which this variation decays exponentially with depth, providing theoretical guarantees of collapse for graph convolution and HGNN propagation (Cai & Wang, 2020; Chen et al., 2022; Li et al., 2025b). Such an argument depends on what the measure recognizes as a collapsed state: its zero set must coincide with the subspace toward which the propagation converges. However, that subspace depends on the aggregation operator, and need not consist of identical node representations (Oono & Suzuki, 2019; Roth & Liebig, 2024; Roth, 2024). Consequently, an energy calibrated to one propagation rule can remain positive after representations have collapsed under another (Zhang et al., 2026). This mismatch limits the transfer of energy-based analyses across architectures and raises two questions: *(1) How can we construct an energy that identifies collapse for different hypergraph propagation operators?*

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.

(2) Under what conditions can that energy characterize both the loss and preservation of feature variation as depth increases?

We address these questions with an operator-relative energy whose notion of zero variation is determined by the propagation operator. We show that the proposed energy vanishes exactly on the operator's collapse subspace, recovers the classical Dirichlet energy as a special case, and provides a common construction for operators with different collapse directions. We further extend the analysis to full message-passing layers, accounting for both stages of fixed node-to-hyperedge and hyperedge-to-node aggregation, learned channel transformations, and intermediate activations. This framework yields conditions under which representation variation decays exponentially with depth across a broad class of hypergraph propagation operators.

Having characterized how representation variation can vanish with depth, we then study how such collapse can be prevented. Using the proposed energy, we incorporate initial-residual and identity-mapping (II) updates (Chen et al., 2020b; 2022; Huang & Yang, 2021) into hypergraph propagation and analyze their effect on over-smoothing. We show that, under mild conditions, II prevents the energy from vanishing as network depth tends to infinity, thereby preserving representation variation outside the operator's collapse subspace. In the linear setting with identity channel maps, we further establish a lower bound on the preserved energy for arbitrary depth.

We evaluate our framework through two complementary sets of experiments on six real-world hypergraph datasets. First, controlled propagation experiments demonstrate that the classical Dirichlet energy can fail to detect representation collapse, while the proposed operator-relative energy correctly detects it and verifies the predicted decay and preservation behavior. Second, we incorporate II into five representative HGNN architectures, including HGNN (Feng et al., 2019), HNHN (Dong et al., 2020), UniGCN (Huang & Yang, 2021), HyperGCN (Yadati et al., 2019), and AllSetTransformer (Chien et al., 2021), and evaluate their node-classification performance as network depth increases. The II variants improve the best mean test accuracy over the corresponding baselines in 28 of 30 architecture–dataset pairs and sustaintheir predictive performance as network depth increases. These comparisons complement the controlled energy experiments by evaluating the practical behavior of II in trained networks.

In summary, our contributions are threefold. First, we introduce an operator-relative energy that adapts the notion of representation variation to the collapse behavior of different hypergraph propagation operators, providing a unified framework that recovers classical Dirichlet energies as special cases. Second, we establish theoretical guarantees characterizing both over-smoothing and its mitigation: we derive conditions under which representation variation decays exponentially with depth and show that incorporating II updates prevents the energy from vanishing under appropriate conditions. Third, motivated by these theoretical results, we incorporate II into five representative HGNN architectures and demonstrate across six real-world datasets that the resulting models consistently improve predictive performance over their corresponding baselines and remain robust as network depth increases.

2 Related Work

Measuring representation collapse. Existing analyses track collapse through distance to an invariant subspace (Oono & Suzuki, 2019) or decay of Dirichlet energy (Cai & Wang, 2020). These approaches relate representation collapse to propagation dynamics, with the chosen subspace or energy determining which states count as collapsed. Rank-based analyses emphasize the loss of independent feature directions and the role of the aggregation operator (Roth & Liebig, 2024; Roth, 2024), while Bison & Sperduti (2025) examine the suitability of Dirichlet energies as over-smoothing measures. Other diagnostics include cosine-based mean average distance (Chen et al., 2020a) and numerical or effective rank (Zhang et al., 2026). These quantities assess feature alignment or effective dimensionality without prescribing a particular collapse direction. Our objective is an operator-relative seminorm whose associated quadratic energy can be controlled across layers.

Hypergraph over-smoothing. Existing hypergraph studies analyze several propagation settings. Chen et al. (2022) establish energy-based over-smoothing results for HGNN, while Li et al. (2025b) study deep hypergraph networks through framelet constructions. Sheaf hypergraph networks associate diffusion with a sheaf structure (Duta et al., 2023); other approaches study particle-based

message passing (Ma et al., 2026) and continuous-time diffusion and reaction–diffusion dynamics (Zhou et al., 2026). These settings differ in both their propagation rules and the measures of representation variation they track. Comparing their guarantees therefore requires specifying the operator, the measure, and the assumptions on the layer dynamics. We focus on a common energy construction for an explicitly defined class of discrete hypergraph propagation operators.

Initial residuals and identity mapping (II). For graph neural networks, GCNII (Chen et al., 2020b) combines repeated injection of the initial features with channel transformations biased toward the identity. Deep-HGCN (Chen et al., 2022) and UniGCNII (Huang & Yang, 2021) adapt this mechanism to hypergraph learning. For linearized graph networks, Scholkemper et al. (2025) characterize how initial residuals preserve information and determine the representation subspace. In contrast, we analyze II updates through the operator-relative energy, establishing theoretical guarantees for variation preservation across a broad class of hypergraph propagation operators, and empirically evaluate II variants across five representative HGNN architectures and increasing depths.

3 Problem Formulation

Let $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ be a connected hypergraph with n nodes, m nonempty hyperedges, and incidence matrix $H \in \{0,1\}^{n\times m}$. Each hyperedge e has weight $w_e > 0$ and cardinality $|e|$. Define the node degrees $d_v = \sum_{e\ni v} w_e$ and the diagonal matrices $D_V = \mathrm{diag}(d_v)$, $D_E = \mathrm{diag}(|e|)$, and $W = \mathrm{diag}(w_e)$. The rows of $X^{(l)} \in \mathbb{R}^{n\times h}$ and $Z^{(l)} \in \mathbb{R}^{m\times h}$ represent node and hyperedge features at layer l, respectively.

We consider a layer that aggregates node features into hyperedges and then returns the transformed hyperedge features to nodes in the following form:

$$Z^{(l)} = \phi_E\big(AX^{(l)}\Theta_1^{(l)}\big), \qquad X^{(l+1)} = \phi_V\big(BZ^{(l)}\Theta_2^{(l)}\big), \tag{1}$$

where $A \in \mathbb{R}^{m\times n}_{\geq 0}$ and $B \in \mathbb{R}^{n\times m}_{\geq 0}$ have positive entries on incident node–hyperedge pairs and zero entries elsewhere. Their aggregation weights are shared across feature channels. The learnable feature transformations (channel maps) $\Theta_i^{(l)} \in \mathbb{R}^{h\times h}$ may vary across layers, and ϕ_E, ϕ_V apply the same activation map to each row. Define the propagation operator $P = BA$. With identity activations and channel maps, the update reduces to $X^{(l+1)} = PX^{(l)}$.

Operator-dependent collapse. To illustrate collapse under repeated propagation, suppose P has a real eigenbasis $\{\psi_k\}_{k=1}^n$ with $P\psi_k = \lambda_k\psi_k$, $\psi_1 = \psi_P > 0$, and $\lambda_1 > |\lambda_k|$ for $k \geq 2$. Expand the initial features as $X^{(0)} = \sum_{k=1}^n \psi_k c_k^\top$, where $c_k \in \mathbb{R}^h$. Repeated propagation gives $X^{(L)} = \sum_{k=1}^n \lambda_k^L \psi_k c_k^\top$, and dividing by the dominant eigenvalue yields

$$X^{(L)}/\lambda_1^L = \psi_P c_1^\top + \sum_{k=2}^n (\lambda_k/\lambda_1)^L \psi_k c_k^\top \to \psi_P c_1^\top$$

as $L \to \infty$. Every component outside the leading eigenspace decays after normalization, leaving the *collapse subspace* $\mathcal{C}_P = \{\psi_P c^\top : c \in \mathbb{R}^h\}$ in the limit. Dividing by λ_1^L separates this representation collapse from global shrinking when $\lambda_1 < 1$, so that the decay after normalization reflects the loss of components outside $\mathcal{C}_P$.

A motivating example. Consider the classical unnormalized and normalized Dirichlet energies (Zhou et al., 2006):

$$\mathcal{E}_{\mathrm{un}}(X) = \frac{1}{2}\sum_{e\in\mathcal{E}} \frac{w_e}{|e|} \sum_{u,v\in e} \|x_u - x_v\|_2^2, \qquad \mathcal{E}_{\mathrm{n}}(X) = \mathcal{E}_{\mathrm{un}}(D_V^{-1/2}X), \tag{2}$$

where x_v denotes the feature vector at node v. The unnormalized energy $\mathcal{E}_{\mathrm{un}}$ vanishes exactly on constant features, while the normalized energy $\mathcal{E}_{\mathrm{n}}$ vanishes on degree-scaled features. Each therefore assigns positive energy to the other operator's nonzero collapse states. This mismatch motivates an energy $\mathcal{E}_P$ whose zero set is exactly $\mathcal{C}_P$ and whose evolution tracks the variation removed or retained by message passing.

4 MAIN RESULTS

We propose the following operator-relative energy construction. All proofs in this section are deferred to Appendix Section A.

Definition 4.1 (Operator-relative energy)**.** Let P be a fixed nonnegative propagation operator with spectral radius $\lambda_1 > 0$ and positive right and left eigenvectors ψ, ℓ. Its *operator-relative energy* for node features $X \in \mathbb{R}^{n\times h}$ is defined as

$$\mathcal{E}_P(X) = \operatorname{tr}\left(X^\top \Pi \left(I - \frac{P}{\lambda_1} \right) X \right), \tag{3}$$

where $\Pi = \operatorname{diag}(\pi)$ with $\pi_i = \ell_i/\psi_i$, $P\psi = \lambda_1\psi$, and $\ell^\top P = \lambda_1 \ell^\top$.

The left and right eigenvectors of P give the following entrywise representation of the energy:

$$\mathcal{E}_P(X) = \frac{1}{2} \sum_{i,j} \frac{\ell_i P_{ij} \psi_j}{\lambda_1} \left\| \frac{x_i}{\psi_i} - \frac{x_j}{\psi_j} \right\|_2^2 \geq 0. \tag{4}$$

This energy measures the differences between rescaled features along the nonzero entries of P. It is zero exactly when every column of X lies in $U_P = \ker(P - \lambda_1 I)$. For later use, write $\Psi = \operatorname{diag}(\psi)$ and $\langle X, Y \rangle_\pi = \operatorname{tr}(X^\top \Pi Y)$. Rescaling the chosen eigenvectors changes the energy only by an overall positive factor. Appendix Section B gives the spectral representation of $\mathcal{E}_P$, shows that it is equivalent to the weighted distance to U_P, and explains why no single seminorm can serve every shared-incidence operator.

Remark 4.2. Under random-walk propagation $P_{\mathrm{RW}} = D_V^{-1} H W D_E^{-1} H^\top$, $\mathcal{E}_{P_{\mathrm{RW}}}(X)$ reduces to the unnormalized Dirichlet energy $\mathcal{E}_{\mathrm{un}}(X)$; under HGNN propagation $P_{\mathrm{HGNN}} = D_V^{-1/2} H W D_E^{-1} H^\top D_V^{-1/2}$, $\mathcal{E}_{P_{\mathrm{HGNN}}}(X)$ reduces to the normalized Dirichlet energy $\mathcal{E}_{\mathrm{n}}(X)$. Thus, both widely used classical Dirichlet energies in Equation (2) are special cases of the proposed operator-relative energy.

To track the energy through eigenmodes, we additionally assume that P is diagonally symmetrizable by the energy weights Π, i.e., $\Pi P = P^\top \Pi$. Under this condition, $\Pi^{1/2} P \Pi^{-1/2}$ is symmetric, and P has real eigenvalues and eigenvectors satisfying $w_i^\top \Pi w_j = \delta_{ij}$. To derive our theoretical results, we consider a class of aggregation operators for which A and B share the factorization

$$B_{ve} = b_v G_{ve} f_e, \qquad A_{ev} = a_e G_{ve} c_v, \tag{5}$$

where $b, c \in \mathbb{R}^n_{>0}$, $a, f \in \mathbb{R}^m_{>0}$, and $G \geq 0$ has the incidence support of H. Here, G contains the shared incidence weights, while a, b, c, f encode the stage-specific scalings. Table 1 lists shared-incidence factors for six representative architectures.

Model	G	b_v	f_e	a_e	c_v	π_v	ψ
HGNN (Feng et al., 2019)	H	$d_v^{-1/2}$	w_e	$1/\lvert e\rvert$	$d_v^{-1/2}$	1	$D_V^{1/2}\mathbf{1}$
RW (Chitra & Raphael, 2019) / HGNN+ (Gao et al., 2023)	H	d_v^{-1}	w_e	$1/\lvert e\rvert$	1	d_v	$\mathbf{1}$
HNHN (Dong et al., 2020)	H	$(\sum_{e\ni v} \lvert e\rvert^\alpha)^{-1}$	$\lvert e\rvert^\alpha$	$(\sum_{u\in e} d_u^\beta)^{-1}$	d_v^β	$d_v^\beta \sum_{e\ni v} \lvert e\rvert^\alpha$	$\mathbf{1}$
UniGCN (Huang & Yang, 2021)	H	$d_v^{-1/2}$	$\bar{d}_e^{-1/2}$	$1/\lvert e\rvert$	1	$d_v^{1/2}$	ψ
GHSC (Zhang et al., 2022)	Q	$\hat{d}_v^{-1/2}$	s_e	1	$\hat{d}_v^{-1/2}$	1	$\hat{D}^{1/2}\mathbf{1}$
TF-HNN (Tang et al., 2025)	H_+	$(d_v^+)^{-1/2}$	1	$1/\lvert e\rvert$	$(d_v^+)^{-1/2}$	1	$D_+^{1/2}\mathbf{1}$

Table 1. Shared-incidence factors, with $B_{ve} = b_v G_{ve} f_e$ and $A_{ev} = a_e G_{ve} c_v$. The HNHN exponents α, β are architecture parameters. For UniGCN, $\bar{d}_e = \lvert e\rvert^{-1} \sum_{v\in e} d_v$. For GHSC Q is a positive weighted incidence matrix, $s_e = w_e \rho(\delta_e) > 0$, $\delta_e = \sum_v Q_{ve}$, and $\hat{d} = Q\operatorname{diag}(s) Q^\top \mathbf{1}$. For TF-HNN, $H_+ = [H, I]$, $d^+ = H_+\mathbf{1}$, and the one-step operator is $D_+^{-1/2}(H D_E^{-1} H^\top + I) D_+^{-1/2}$.

The following proposition shows that the shared-incidence factorization connects the two aggregation stages and provides the spectral structure needed for the energy analysis.

Proposition 4.3. *For (A, B) with the shared-incidence factorization in Equation (5), set $\pi_v = c_v/b_v$, $\pi_{E,e} = f_e/a_e$, $\Pi = \mathrm{diag}(\pi)$, and $\Pi_E = \mathrm{diag}(\pi_E)$. Then*

$$\Pi B = A^\top \Pi_E, \qquad \Pi P = A^\top \Pi_E A \succeq 0, \qquad P = BA. \tag{6}$$

On a connected hypergraph, P has a real eigenbasis and $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, with a corresponding positive eigenvector $\psi = \psi_P$ associated with λ_1.

To analyze the intermediate activation, we also measure variation in the hyperedge representations. Just as $P = BA$ propagates node features, AB propagates hyperedge features. Applying Definition 4.1 to AB with the weights Π_E from Proposition 4.3 gives, for $Z \in \mathbb{R}^{m\times h}$,

$$\mathcal{E}_E(Z) = \mathrm{tr}\big(Z^\top \Pi_E (I - AB/\lambda_1) Z\big). \tag{7}$$

Its zero set is $\{\psi_E c^\top : c \in \mathbb{R}^h\}$, where $\psi_E = A\psi$. The compatibility between the two energies gives $\mathcal{E}_E(AX) \leq \lambda_2 \mathcal{E}_P(X)$ and $\mathcal{E}_P(BZ) \leq \lambda_2 \mathcal{E}_E(Z)$. These inequalities allow us to track representation variation through the two-stage propagation.

The following theorem recovers the known decay result for HGNN (Chen et al., 2022) and extends it to any fixed aggregation with a shared-incidence factorization, including non-symmetric operators.

Theorem 4.4. *Let (A, B) be fixed with the shared-incidence factorization in Equation (5). In Equation (1), assume ϕ_V, ϕ_E are shared 1-Lipschitz row maps, positively homogeneous unless ψ, ψ_E, respectively, are constant. Then*

$$\mathcal{E}_P(X^{(l+1)}) \leq \lambda_2^2 \|\Theta_1^{(l)}\|_2^2 \|\Theta_2^{(l)}\|_2^2 \mathcal{E}_P(X^{(l)}).$$

Consequently, the operator-relative energy decays exponentially if $\lambda_2 \sup_l(\|\Theta_1^{(l)}\|_2 \|\Theta_2^{(l)}\|_2) < 1$.

To counteract energy decay, we incorporate initial residuals and identity mappings (II) into hypergraph propagation, following GCNII and its hypergraph instances Deep-HGCN and UniGCNII. Specifically, we consider the following two-stage II layer:

$$Z^{(l)} = \phi_E\left(AX^{(l)} M_{E,l}\right), \quad X^{(l+1)} = \phi_V\left(\big((1-\alpha)BZ^{(l)} + \alpha X^{(0)}\big) M_{V,l}\right), \tag{8}$$

where $M_{s,l} = (1-\beta_l)I + \beta_l \Theta_s^{(l)}$ for $s \in \{E, V\}$, $0 < \alpha < 1$, $\lambda > 0$, and $\beta_l = \log(1 + \lambda/(l+1))$. With uniformly bounded learned weights, $M_{s,l} \to I$ as $l \to \infty$.

We next characterize how II preserves operator-relative energy. The following proposition provides a one-layer lower bound that accounts for both aggregation stages and cancellation with the initial residual.

Proposition 4.5. *Under the aggregation and activation assumptions of Theorem 4.4, additionally assume $\|\phi_V(u) - \phi_V(v)\|_2 \geq c_V \|u - v\|_2$ for all $u, v \in \mathbb{R}^h$, where $c_V > 0$. Write $m_{V,l} = \sigma_{\min}(M_{V,l})$ and $s_{E,l} = \|M_{E,l}\|_2$. Then*

$$\mathcal{E}_P(X^{(l+1)}) \geq c_V^2 m_{V,l}^2 \left(\alpha\sqrt{\mathcal{E}_P(X^{(0)})} - (1-\alpha)\lambda_2 s_{E,l}\sqrt{\mathcal{E}_P(X^{(l)})}\right)_+^2, \tag{9}$$

where $(t)_+ = \max\{t, 0\}$.

Because the above bound is not guaranteed to be strictly positive, we therefore show that the limiting energy of the two-stage II update admits a positive lower bound whenever the initial energy is nonzero.

Theorem 4.6. *Let A, B be fixed and satisfy Equation (5), with $P = BA$ and $\lambda_1 \leq 1$. In Equation (8), assume uniformly bounded learned weights $\Theta_E^{(l)}, \Theta_V^{(l)}$ and shared 1-Lipschitz activations. Assume either both activations are identity maps, or $X^{(0)} \geq 0$ and both fix every nonnegative vector. Then*

$$X^{(L)} \xrightarrow{L\to\infty} X^\star = \alpha(I - (1-\alpha)P)^{-1} X^{(0)}, \qquad \alpha^2 \mathcal{E}_P(X^{(0)}) \leq \mathcal{E}_P(X^\star) \leq \mathcal{E}_P(X^{(0)}).$$

The initial residual ensures that the limiting energy is at least α^2 times the initial energy. If $\mathcal{E}_P(X^{(0)}) > 0$, this positive asymptotic floor prevents convergence to the operator's collapse subspace.

5 EXPERIMENTS

In this section, we conduct experiments on six widely used benchmark datasets in hypergraph learning: Cora (Yadati et al., 2019), Citeseer (Yadati et al., 2019), 20News100 (Zhou et al., 2006; Chien et al., 2021), House-Committee (Chien et al., 2021), Pokec (Li et al., 2025c), and NTU2012 (Feng et al., 2019; Chien et al., 2021). These datasets cover citation networks, text documents, political memberships, social relationships, and 3D objects, and have been widely used as benchmark datasets in the literature, including AllSet (Chien et al., 2021), ED-HNN (Wang et al., 2023), and DHG-Bench (Li et al., 2025a). Table 2 describes what the nodes and hyperedges represent in each dataset and reports the corresponding summary statistics. Appendix Section C details the architecture implementations, datasets and preprocessing, the full energy results, and the hyperparameter search spaces.

Dataset	Node	Hyperedge	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	Features	Classes	avg. $\lvert e\rvert$
Cora	Papers	Co-cited papers	2,708	1,579	1,433	7	3.03
Citeseer	Papers	Co-cited papers	3,312	1,079	3,703	6	3.20
20News100	Documents	Documents containing the same word	16,242	100	100	4	654.51
House-Committee	Representatives	Committees	1,290	341	100	2	34.73
Pokec	Users	Friends of a user	14,998	2,406	65	2	2.29
NTU2012	3D objects	Neighborhoods in visual feature space	2,012	2,012	100	67	5.00

Table 2. Benchmark dataset descriptions and summary statistics.

5.1 DIAGNOSING COLLAPSE AND TRACKING ENERGY UNDER FIXED PROPAGATION

We first examine whether classical Dirichlet energies correctly identify representation collapse and whether the operator-relative energy captures energy decay and preservation under II propagation. We isolate these effects using fixed linear propagation, for which the collapse subspace and spectral bounds are explicitly available.

We select HGNN, HNHN, and UniGCN to cover distinct collapse directions: degree-scaled features for HGNN, constant features for HNHN, and features proportional to the Perron eigenvector for UniGCN. Their operators admit the shared-incidence factorization in Proposition 4.3. On each dataset, the operators use the same largest connected component and original node features. With identity channel maps and activations, we run 256 steps of

$$X^{(\ell+1)} = PX^{(\ell)}, \qquad Y^{(\ell+1)} = (1-\alpha)PY^{(\ell)} + \alpha X^{(0)}, \qquad Y^{(0)} = X^{(0)}.$$

Appendix C.2 provides additional details on the six datasets, component sizes, Gaussian-initialization control, and numerical checks. To distinguish changes in representation shape from global shrinking, we consider three scale-invariant metrics:

$$R_{\mathrm{un}}(X) = \frac{\mathcal{E}_{\mathrm{un}}(X)}{\mathrm{tr}(X^\top X)}, \qquad R_{\mathrm{n}}(X) = \frac{\mathcal{E}_{\mathrm{n}}(X)}{\mathrm{tr}(X^\top X)}, \qquad R_P(X) = \frac{\mathcal{E}_P(X)}{\mathrm{tr}(X^\top \Pi X)}.$$

Figure 1(a) compares these three metrics along the same base propagation trajectory. Figure 1(b) reports the normalized operator-relative energy, $\mathcal{E}_P(X^{(\ell)})/\mathcal{E}_P(X^{(0)})$, under base and II propagation. Together, the two panels examine whether the proposed energy correctly identifies operator-specific collapse and captures the predicted energy decay and preservation.

Diagnostic mismatch. As shown in Figure 1(a), for HGNN, R_{n} coincides with R_P, while R_{un} retains a positive residual. For HNHN, the unnormalized energy shares the operator's constant collapse direction, whereas the normalized energy assigns positive energy to nonzero constant features on an irregular hypergraph. For UniGCN, neither classical energy has a zero set guaranteed to match its Perron direction, and both classical diagnostics can retain a residual. These results demonstrate that classical Dirichlet energies can fail to detect representation collapse when their zero sets do not match the operator-specific collapse direction, whereas the proposed operator-relative energy consistently captures the collapse across different propagation operators and datasets.

The nearly flat HNHN curves on 20News100 warrant additional explanation. They reflect slow propagation rather than a failure to detect decay. The second-largest eigenvalue of the HNHN propagation operator is $\lambda_2 \approx 0.99999$, so the slowest nonconstant mode retains a fraction $\lambda_2^{512} \approx 0.9955$

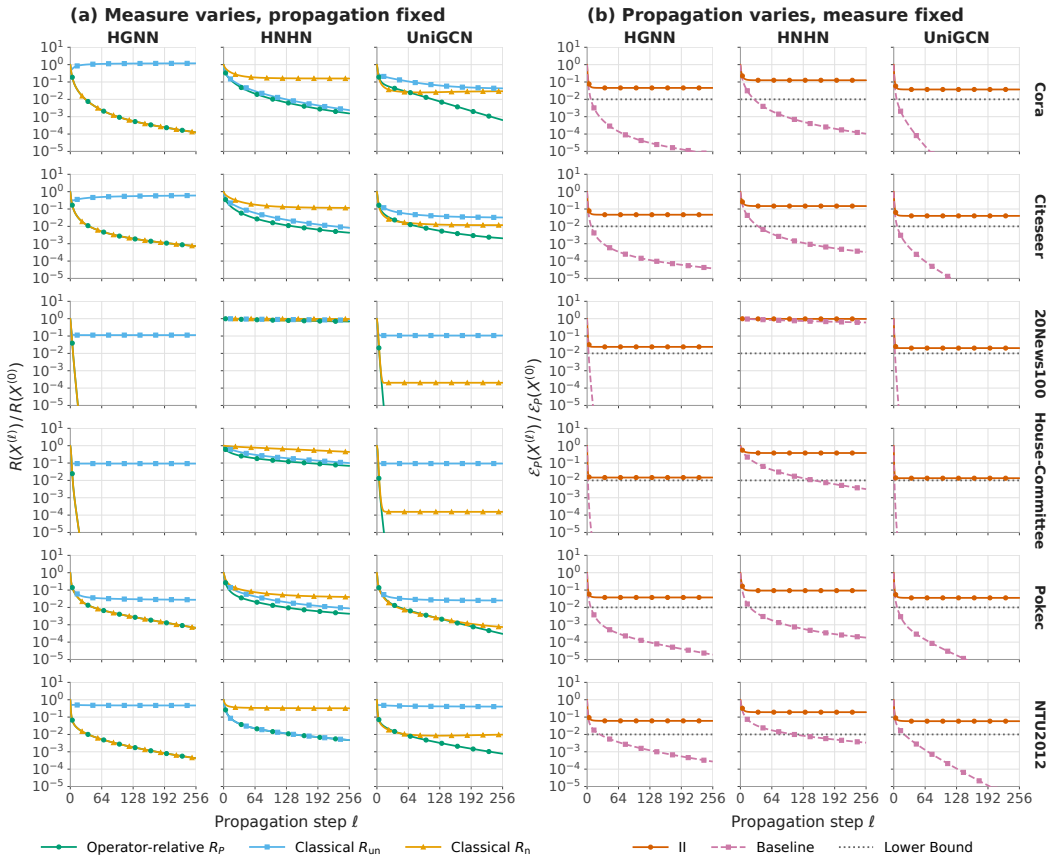


Figure 1. Energy trajectories under fixed linear propagation, normalized by their initial values. **(a)** Three scale-invariant energy measures under the same base propagation, comparing classical Dirichlet energies with the proposed operator-relative energy. **(b)** Operator-relative energy under base and II propagation; dotted lines indicate the theoretical lower bound at $\alpha^2 = 0.01$ with $\alpha = 0.1$. Rows correspond to six datasets and columns to propagation operators.

of its energy after 256 steps. The measured operator-relative energy falls to 0.604 of its initial value, while the scale-invariant diagnostic falls to $R_P = 0.687$; both decreases appear small on the shared logarithmic axis.

Energy decay and preservation. Figure 1(b) reports operator-relative energy normalized by its initial value, without feature-norm normalization. Across all six datasets, base propagation satisfies $\mathcal{E}_P(X^{(\ell)}) \leq \lambda_2^{2\ell}\mathcal{E}_P(X^{(0)})$ at every step within numerical tolerance, while II propagation remains above $\alpha^2 = 0.01$ of the initial energy (Table 4). These observations support both the predicted energy decay under base propagation and the theoretical positive energy lower bound under II.

5.2 Predictive Performance at Increasing Depth

Having shown that II preserves representation variation under controlled propagation, we next examine whether this translates into improved robustness to network depth in trained HGNNs. Specifically, we compare the node classification accuracy of five baseline propagation methods and their corresponding II variants across different network depths.

We consider five methods HGNN, HNHN, UniGCN, HyperGCN, and AllSetTransformer. The first three provide different fixed aggregation operators covered by our framework; the latter two test the empirical applicability of II beyond the fixed-aggregation assumptions. Each method and its II variant are evaluated at depths ($L \in \{2, 4, 8, 16, 32, 64\}$) on the same six real-world datasets (Table 2). All models are trained in a full-batch setting for 200 epochs using Adam and the negative

log-likelihood loss. Each seed uses a randomly drawn 50/25/25 train/validation/test split, and model selection is based on validation accuracy. To ensure a fair comparison across depths, we tune hyperparameters separately for every (model, dataset, depth) combination using Optuna with at least 30 trials, evaluating each trial over 3 random seeds; the search spaces are summarized in Tables 5 and 6. The selected configuration is then retrained across 10 random seeds, and we report the resulting test performance.

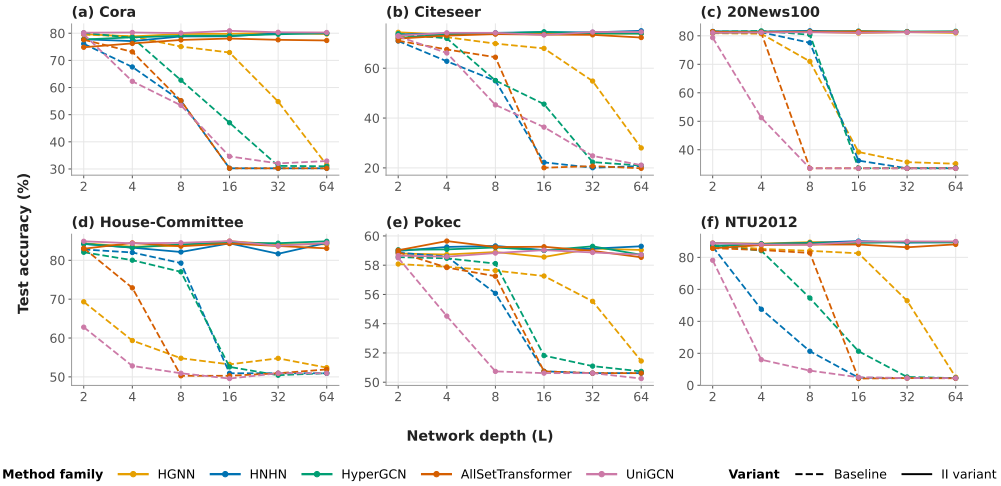


Figure 2. Test accuracy against network depth for the five baseline architectures and their II variants, on (a) Cora, (b) Citeseer, (c) 20News100, (d) House-Committee, (e) Pokec, and (f) NTU2012. Colors denote method families; dashed and solid lines denote baselines and II variants, respectively. Each point reports the mean test accuracy over 10 seeds using the per-(model, dataset, depth) configuration selected by Optuna.

Performance against network depth. Figure 2 shows a clear separation between the baseline and II variants. At shallow depths, the baseline and II variants generally achieve comparable accuracy. As depth increases, however, the baseline models deteriorate sharply across most method-dataset pairs, often approaching near-random accuracy. In contrast, the II variants remain stable even at the largest depth considered. Averaged over all 30 method–dataset pairs, the baseline accuracy drops by 42.94 percentage points from $L = 2$ to $L = 64$, whereas the II variants show no degradation, even achieving an average gain of 0.66 points.

Model	Variant	Cora	Citeseer	20News100	House-Committee	Pokec	NTU2012
HGNN	Base	79.51±1.09 (2)	72.95±0.86 (2)	80.91±0.40 (2)	69.32±2.18 (2)	58.08±0.87 (2)	87.67±1.35 (2)
	II	80.18±1.23 (32)	74.40±0.94 (2)**	81.38±0.36 (8)*	84.92±2.27 (16)***	59.14±0.68 (32)*	89.72±1.39 (64)**
HNHN	Base	77.02±1.90 (2)	72.19±1.65 (2)	81.54±0.47 (2)	82.48±2.16 (4)	58.90±0.81 (2)	86.68±0.84 (2)
	II	80.55±0.77 (64)***	74.55±1.14 (64)**	81.64±0.55 (16)	84.92±2.55 (32)*	59.63±0.73 (2)	**90.20±1.22 (16)*****
HyperGCN	Base	80.13±1.14 (2)	73.83±0.95 (2)	**81.72±0.44 (4)**	82.07±2.19 (2)	58.55±0.60 (2)	87.40±1.56 (2)
	II	80.04±1.29 (64)	**74.58±1.11 (16)**	81.66±0.50 (16)	84.89±2.38 (64)*	59.29±0.73 (32)*	89.52±1.50 (32)**
AllSetTransformer	Base	77.61±1.39 (2)	71.04±0.98 (2)	81.48±0.36 (2)	83.13±2.02 (2)	58.89±0.62 (2)	85.77±1.95 (2)
	II	78.05±1.06 (16)	73.67±0.98 (8)***	81.52±0.38 (16)	84.46±2.10 (4)	**59.65±0.76 (4)***	88.13±1.47 (64)*
UniGCN	Base	79.65±1.02 (2)	72.75±1.44 (2)	79.35±1.23 (2)	62.79±4.88 (2)	58.50±0.61 (2)	78.19±2.53 (2)
	II	**80.90±1.30 (16)***	74.54±1.01 (64)**	81.27±0.37 (32)**	**84.95±2.49 (16)*****	59.03±0.64 (16)	89.98±1.28 (32)***

Table 3. Best test accuracy (%) across depths for each method on each dataset. Reported as mean ± population standard deviation over 10 seeds at the best-performing depth (shown in parentheses); ties in mean accuracy select the shallower depth. Depth values include $L \in \{2, 4, 8, 16, 32, 64\}$. Bold marks the best result per dataset and underline marks the second-best result, with ties retained. Stars on an II variant denote approximate significance of its improvement over the corresponding base variant under a two-sided Welch's t-test reconstructed from the rounded means and standard deviations, converting population to sample standard deviations, with 10 seeds per model: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Best-of-depth comparison. Across the 30 method–dataset pairs, the best-performing II variant outperforms the corresponding best baseline in mean test accuracy in 28 cases, with 20 comparisons reaching an approximate significance level of $p < 0.05$ based on two-sided Welch's t-tests. The results are summarized in Table 3. The two no-improvement exceptions are small, each far below one reported standard deviation. More notably, the performance optima shift markedly toward deeper architectures. Among the baselines, 28 of 30 best accuracies occur at $L = 2$, and the remaining two occur at $L = 4$. Among the II variants, only two best accuracies occur at $L = 2$; 23 of 30 occur at $L \in \{16, 32, 64\}$. Thus, II not only avoid the severe performance degradation observed in deep baseline models, but also allows deeper architectures to achieve their best predictive performance.

6 Discussion and Future Work

A central challenge in studying over-smoothing is that different hypergraph propagation operators can drive node representations toward different collapse states. Consequently, classical Dirichlet energies may fail to detect over-smoothing when their notion of zero variation does not match the collapse behavior of the propagation operator. We address this problem by introducing an operator-relative energy whose zero set matches the operator-specific collapse subspace, providing a unified measure of representation variation across different propagation rules. Our analysis and controlled experiments demonstrate that this operator-specific construction correctly captures representation collapse where classical energies can fail. We further establish conditions under which operator-relative energy decays with network depth and under which II preserves a positive level of representation variation. Across five HGNN architectures, our predictive experiments show that the II variants remain robust as network depth increases, while the corresponding baseline models exhibit substantial performance degradation. Together, these results provide a unified view of both the emergence and mitigation of over-smoothing under different hypergraph propagation rules.

Several limitations of the current framework motivate directions of future work. First, our theoretical analysis characterizes representation collapse rather than predictive performance. A positive operator-relative energy guarantees that variation remains outside the collapse subspace, but does not establish that this variation is useful for prediction. Although our classification experiments provide empirical evidence that II can simultaneously preserve variation and improve predictive performance, a theoretical connection between these two effects remains open. Future work could establish conditions under which preserved variation contains task-relevant signal, for example by relating operator-relative energy to class separation or prediction error. Second, this distinction between preserved variation and predictive information also motivates more selective mechanisms for retaining initial features. Rather than inject all feature channels with a common residual strength, future architectures could selectively retain task-relevant information while allowing irrelevant variation to be smoothed. Controlled experiments that vary feature noise and the agreement between hyperedges and class labels could help identify when residual injection preserves useful distinctions and when it obstructs beneficial smoothing. Separating energy contributions within and between classes could support this investigation, with any label-guided training rule restricted to training labels. Third, our analysis assumes fixed propagation operators, whereas attention and learned incidence weights can change propagation across layers. Extending the framework to these settings requires distinguishing changes in the representations from changes in the collapse subspace and the energy used to measure them. Studying slowly varying operators, i.e., $\|P_{l+1} - P_l\| \leq \epsilon$, could provide a starting point for understanding their cumulative effect.

AI use statement

We used generative AI tools to assist with literature searches, code development, and manuscript editing for clarity and readability. The authors checked the AI-assisted work by reading the suggested references running unit tests on the code, and revising all draft text. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with AI assistance.

Reproducibility statement

The complete source code and associated datasets supporting this study are publicly accessible at https://anonymous.4open.science/r/hygII-2421

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Anna Bison and Alessandro Sperduti. Analysis of Dirichlet energies as over-smoothing measures. *arXiv preprint arXiv:2512.09890*, 2025.

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

Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3438–3445, 2020a.

Guanzi Chen, Jiying Zhang, Xi Xiao, and Yang Li. Preventing over-smoothing for hypergraph neural networks. *arXiv preprint arXiv:2203.17159*, 2022.

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

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are allset: A multiset function framework for hypergraph neural networks. *arXiv preprint arXiv:2106.13264*, 2021.

Uthsav Chitra and Benjamin Raphael. Random walks on hypergraphs with edge-dependent vertex weights. In *International conference on machine learning*, pp. 1172–1181. PMLR, 2019.

Yihe Dong, Will Sawin, and Yoshua Bengio. Hnhn: Hypergraph networks with hyperedge neurons. *arXiv preprint arXiv:2006.12278*, 2020.

Iulia Duta, Giulia Cassarà, Fabrizio Silvestri, and Pietro Liò. Sheaf hypergraph networks. *Advances in Neural Information Processing Systems*, 36:12087–12099, 2023.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3558–3565, 2019.

Yue Gao, Yifan Feng, Shuyi Ji, and Rongrong Ji. Hgnn+: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3181–3199, 2023. doi: 10.1109/TPAMI.2022.3182052.

Jing Huang and Jie Yang. Unignn: a unified framework for graph and hypergraph neural networks. *arXiv preprint arXiv:2105.00956*, 2021.

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: graph neural networks meet personalized PageRank. In *International Conference on Learning Representations*, 2019.

Fan Li, Xiaoyang Wang, Wenjie Zhang, Ying Zhang, and Xuemin Lin. Dhg-bench: A comprehensive benchmark for deep hypergraph learning. *arXiv preprint arXiv:2508.12244*, 2025a.

Ming Li, Yujie Fang, Yi Wang, Han Feng, Yongchun Gu, Lu Bai, and Pietro Lio. Deep hypergraph neural networks with tight framelets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18385–18392, 2025b.

Ming Li, Yongchun Gu, Yi A Wang, Yujie Fang, Lu Bai, Xiaosheng Zhuang, and Pietro Lio. When hypergraph meets heterophily: New benchmark datasets and baseline. In *AAAI Conference on Artificial Intelligence*, 2025c. URL https://api.semanticscholar.org/CorpusID:277750213.

Yixuan Ma, Kai Yi, Pietro Liò, Shi Jin, and Yuguang Wang. How particle system theory enhances hypergraph message passing. *Advances in Neural Information Processing Systems*, 38:135304–135335, 2026.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947*, 2019.

Andreas Roth. Simplifying the theory on over-smoothing. *arXiv preprint arXiv:2407.11876*, 2024.

Andreas Roth and Thomas Liebig. Rank collapse causes over-smoothing and over-correlation in graph neural networks. In *Learning on Graphs Conference*, pp. 35–1. PMLR, 2024.

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

Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael Schaub. Residual connections and normalization can provably prevent oversmoothing in gnns. In *International Conference on Learning Representations*, volume 2025, pp. 43581–43610, 2025.

Bohan Tang, Zexi Liu, Keyue Jiang, Siheng Chen, and Xiaowen Dong. Training-free message passing for learning on hypergraphs, 2025. URL https://arxiv.org/abs/2402.05569.

Peihao Wang, Shenghao Yang, Yunyu Liu, Zhangyang Wang, and Pan Li. Equivariant hypergraph diffusion neural operators. In *International Conference on Learning Representations (ICLR)*, 2023.

Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. Hypergcn: A new method for training graph convolutional networks on hypergraphs. *Advances in neural information processing systems*, 32, 2019.

Jiying Zhang, Fuyang Li, Xi Xiao, Tingyang Xu, Yu Rong, Junzhou Huang, and Yatao Bian. Hypergraph convolutional networks via equivalency between hypergraphs and undirected graphs, 2022. URL https://arxiv.org/abs/2203.16939.

Kaicheng Zhang, Piero Deidda, Desmond Higham, and Francesco Tudisco. Are we measuring oversmoothing in graph neural networks correctly? In *International Conference on Learning Representations*, volume 2026, pp. 143581–143609, 2026.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. *Advances in neural information processing systems*, 19, 2006.

Zhiheng Zhou, Mengyao Zhou, Yancheng Chen, Dengyi Zhao, Xingqin Qi, and Guiying Yan. From diffusion to reaction-diffusion: A dynamical-systems view of oversmoothing in hypergraph neural networks. *arXiv preprint arXiv:2607.15773*, 2026.

APPENDIX

The appendix is organized as follows. Section A provides the proofs of the main results. Section B develops supplementary properties of fixed propagation, including spectral and stochastic representations, limitations of classical energy measures, and residual fixed-point bounds. Section C presents the architecture formulations and implementation details, datasets and preprocessing, the full energy experiment protocol and results, and the classification protocol and hyperparameter search.

A PROOFS OF THE MAIN RESULTS

This section proves the results of Section 4 in order: the recovery of the classical Dirichlet energies (Remark 4.2), the spectral structure of shared-incidence aggregation (Proposition 4.3), two-stage energy decay (Theorem 4.4), the one-layer lower bound under II (Proposition 4.5), and the asymptotic energy floor (Theorem 4.6), followed by a finite-depth version of the floor for learned channel maps (Theorem A.1). Throughout this section, $\|X\|_\pi^2 = \mathrm{tr}(X^\top \Pi X)$ and $\|Z\|_{\pi_E}^2 = \mathrm{tr}(Z^\top \Pi_E Z)$.

Proof of Remark 4.2. We use the normalizations $\Pi = D_V$ for random-walk propagation and $\Pi = I$ for HGNN. Write $K = HWD_E^{-1}H^\top$. Since $K\mathbf{1} = D_V\mathbf{1}$,

$$\mathcal{E}_{\mathrm{un}}(X) = \mathrm{tr}(X^\top(D_V - K)X), \qquad \mathcal{E}_{\mathrm{n}}(X) = \mathrm{tr}\Big(X^\top(I - D_V^{-1/2}KD_V^{-1/2})X\Big).$$

For $P_{\mathrm{RW}} = D_V^{-1}K$, choose $\psi = \mathbf{1}$ and $\ell = d$, so $\lambda_1 = 1$ and $\Pi = D_V$. Equation equation 3 then gives $\mathcal{E}_{P_{\mathrm{RW}}} = \mathcal{E}_{\mathrm{un}}$. For $P_{\mathrm{HGNN}} = D_V^{-1/2}KD_V^{-1/2}$, choose $\ell = \psi = D_V^{1/2}\mathbf{1}$, so $\Pi = I$ and $\mathcal{E}_{P_{\mathrm{HGNN}}} = \mathcal{E}_{\mathrm{n}}$. The same calculation recovers the normalized graph Dirichlet energy (Cai & Wang, 2020). □

Proof of Proposition 4.3. From Equation (5),

$$A = \mathrm{diag}(a)G^\top \mathrm{diag}(c), \qquad B = \mathrm{diag}(b)G\,\mathrm{diag}(f).$$

With $\Pi = \mathrm{diag}(c/b)$ and $\Pi_E = \mathrm{diag}(f/a)$, direct multiplication gives

$$\Pi B = \mathrm{diag}(c)G\,\mathrm{diag}(f) = A^\top\Pi_E, \qquad \Pi P = A^\top\Pi_E A \succeq 0.$$

Consequently, $\Pi^{1/2}P\Pi^{-1/2}$ is symmetric positive semidefinite. The support of P connects nodes that share a hyperedge, and every diagonal entry is positive. Connectedness therefore makes P irreducible and aperiodic. Perron–Frobenius theory gives a simple positive leading eigenvalue and $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. □

Proof of Theorem 4.4. We first record the properties of the hyperedge energy Equation (7). Set $P_E = AB$. Weighted adjointness also gives $\Pi_E A = B^\top\Pi$, so P_E is self-adjoint and positive semidefinite in the edge inner product. Its Perron vector is $\psi_E = A\psi > 0$, and P_E and P share their nonzero eigenvalues. The energy equation 7 therefore has the same pairwise and spectral representations as the node energy, with edge weights and ψ_E in place of node weights and ψ.

Stagewise contraction. We show that, under the assumptions of Proposition 4.3,

$$\mathcal{E}_E(AX) \leq \lambda_2\mathcal{E}_P(X), \qquad \mathcal{E}_P(BZ) \leq \lambda_2\mathcal{E}_E(Z). \tag{10}$$

Expand $X = \sum_k w_k c_k^\top$ in a π-orthonormal eigenbasis of P. Then

$$\mathcal{E}_E(AX) = \langle X, P(I - P/\lambda_1)X\rangle_\pi = \sum_{k\geq 2}\lambda_k(1 - \lambda_k/\lambda_1)\|c_k\|_2^2 \leq \lambda_2\mathcal{E}_P(X).$$

The second inequality follows in the same way from $\mathcal{E}_P(BZ) = \langle Z, P_E(I - P_E/\lambda_1)Z\rangle_{\pi_E}$: every nonleading eigenvalue of P_E is at most λ_2, and any additional zero eigenvalues contribute zero to the left-hand side.

Channel maps and activations. For either energy, right multiplication by a channel map obeys $\mathcal{E}(U\Theta) \leq \|\Theta\|_2^2\mathcal{E}(U)$. A shared 1-Lipschitz row map is nonexpansive in the pairwise energy when the corresponding collapse vector is constant. For a nonconstant collapse vector, positive homogeneity gives

$$\frac{\phi(u_i)}{\psi_i} = \phi\left(\frac{u_i}{\psi_i}\right), \qquad \left\|\frac{\phi(u_i)}{\psi_i} - \frac{\phi(u_j)}{\psi_j}\right\|_2 \leq \left\|\frac{u_i}{\psi_i} - \frac{u_j}{\psi_j}\right\|_2.$$

The edge argument uses ψ_E instead of ψ.

Conclusion. Applying these estimates at each stage of equation 1 gives

$$\mathcal{E}_E(Z^{(l)}) \leq \lambda_2\|\Theta_1^{(l)}\|_2^2\mathcal{E}_P(X^{(l)}),$$
$$\mathcal{E}_P(X^{(l+1)}) \leq \lambda_2\|\Theta_2^{(l)}\|_2^2\mathcal{E}_E(Z^{(l)}).$$

Combining and iterating proves the stated one-layer bound and the exponential decay condition. □

Proof of Proposition 4.5. Write $N_P = \sqrt{\mathcal{E}_P}$ and $N_E = \sqrt{\mathcal{E}_E}$. The node activation's lower Lipschitz bound, together with positive homogeneity or a constant collapse direction, gives

$N_P(\phi_V(U)) \geq c_V N_P(U)$ by the pairwise energy representation. Right multiplication gives $N_P(UM_{V,l}) \geq m_{V,l} N_P(U)$. The reverse triangle inequality and equation 10 therefore yield

$$\begin{aligned} N_P(X^{(l+1)}) &\geq c_V m_{V,l} \left(\alpha N_P(X^{(0)}) - (1-\alpha) N_P(BZ^{(l)}) \right)_+ \\ &\geq c_V m_{V,l} \left(\alpha N_P(X^{(0)}) - (1-\alpha)\sqrt{\lambda_2} N_E(Z^{(l)}) \right)_+ . \end{aligned}$$

For the hyperedge stage, the upper activation bound and equation 10 give

$$N_E(Z^{(l)}) \leq N_E(AX^{(l)} M_{E,l}) \leq \sqrt{\lambda_2}\, s_{E,l} N_P(X^{(l)}).$$

Substitution and squaring prove equation 9. The minimum singular value can be bounded by $m_{V,l} \geq (1 - \beta_l \|\Theta_V^{(l)} - I\|_2)_+$, which makes explicit how identity mapping can prevent the node channel map from becoming singular. □

Proof of Theorem 4.6. Write $r = 1-\alpha$ and $\varepsilon_{s,l} = \|M_{s,l} - I\|_2 = \beta_l \|\Theta_s^{(l)} - I\|_2$ for $s \in \{E, V\}$. Define $\|Z\|_{\pi_E}^2 = \operatorname{tr}(Z^\top \Pi_E Z)$. Weighted adjointness, $\Pi B = A^\top \Pi_E$, implies

$$\|AX\|_{\pi_E}^2 = \langle X, PX \rangle_\pi \leq \lambda_1 \|X\|_\pi^2, \qquad \|BZ\|_\pi \leq \sqrt{\lambda_1} \|Z\|_{\pi_E}.$$

Define the linear reference trajectory

$$Y^{(0)} = X^{(0)}, \qquad Y^{(l+1)} = rPY^{(l)} + \alpha X^{(0)}.$$

Since $\lambda_1 \leq 1$, induction gives $\|Y^{(l)}\|_\pi \leq \|X^{(0)}\|_\pi$. If $X^{(0)} \geq 0$, then $Y^{(l)} \geq 0$ and $AY^{(l)} \geq 0$ because $A, P \geq 0$. Either activation assumption in the theorem therefore ensures $\phi_E(AY^{(l)}) = AY^{(l)}$ and $\phi_V(Y^{(l+1)}) = Y^{(l+1)}$.

Set $e_l = \|X^{(l)} - Y^{(l)}\|_\pi$ and $N_0 = \|X^{(0)}\|_\pi$. Since ϕ_E and ϕ_V apply the same 1-Lipschitz map to every row and both weighted norms are weighted sums of squared row norms, $\|\phi_V(U) - \phi_V(V)\|_\pi \leq \|U - V\|_\pi$, $\|\phi_E(U) - \phi_E(V)\|_{\pi_E} \leq \|U - V\|_{\pi_E}$ for all U, V. Using $\phi_E(AY^{(l)}) = AY^{(l)}$, the hyperedge error satisfies

$$\begin{aligned} \|Z^{(l)} - AY^{(l)}\|_{\pi_E} &\leq \|A(X^{(l)} - Y^{(l)}) M_{E,l} + AY^{(l)}(M_{E,l} - I)\|_{\pi_E} \\ &\leq \sqrt{\lambda_1}\big((1 + \varepsilon_{E,l}) e_l + \varepsilon_{E,l} N_0\big). \end{aligned}$$

The node error satisfies

$$\begin{aligned} e_{l+1} &\leq \|rB(Z^{(l)} - AY^{(l)}) M_{V,l} + Y^{(l+1)}(M_{V,l} - I)\|_\pi \\ &\leq r\sqrt{\lambda_1}(1 + \varepsilon_{V,l}) \|Z^{(l)} - AY^{(l)}\|_{\pi_E} + \varepsilon_{V,l} N_0 \\ &\leq q_l e_l + \eta_l N_0, \end{aligned}$$

where

$$q_l = r\lambda_1 (1 + \varepsilon_{E,l})(1 + \varepsilon_{V,l}), \qquad \eta_l = r\lambda_1 \varepsilon_{E,l}(1 + \varepsilon_{V,l}) + \varepsilon_{V,l}.$$

This yields $e_l \leq d_l N_0$, where

$$d_0 = 0, \qquad d_{l+1} = q_l d_l + \eta_l. \tag{11}$$

Uniform boundedness of both learned channel sequences and $\beta_l \to 0$ imply $\varepsilon_{E,l}, \varepsilon_{V,l} \to 0$. Hence $q_l \to r\lambda_1 < 1$ and $\eta_l \to 0$. For some L_0 and $\bar{q} < 1$, $q_l \leq \bar{q}$ for all $l \geq L_0$, giving

$$d_L \leq \bar{q}^{L - L_0} d_{L_0} + \sum_{j=L_0}^{L-1} \bar{q}^{L-1-j} \eta_j \longrightarrow 0.$$

The limit follows by splitting the sum into a finite prefix and a tail on which η_j is arbitrarily small. The linear reference converges to $X^\star = \alpha(I - rP)^{-1} X^{(0)}$, hence so does $X^{(L)}$.

For completeness, expand $X^{(0)} = \sum_k w_k c_k^\top$ in a π-orthonormal eigenbasis of P. The reference gain on mode k is

$$g_k(L) = (r\lambda_k)^L + \alpha \sum_{j=0}^{L-1} (r\lambda_k)^j.$$

Since $g_k(0) = 1$ and $g_k(L+1) = (1-\alpha)\lambda_k g_k(L) + \alpha$ with $0 \le \lambda_k \le 1$, induction gives $\alpha \le g_k(L) \le 1$ for $L \ge 1$. Therefore

$$\alpha^2 \mathcal{E}_P(X^{(0)}) \le \mathcal{E}_P(Y^{(L)}) \le \mathcal{E}_P(X^{(0)}).$$

Continuity proves the claimed bounds at $X^\star$. □

For learned channel maps, the following result controls the finite-depth energy through their deviations from identity.

Theorem A.1. *Under the assumptions of Theorem 4.6, set* $\varepsilon_s = \sup_l \beta_l \|\Theta_s^{(l)} - I\|_2$ *for* $s \in \{E, V\}$, *and define*

$$q = (1-\alpha)\lambda_1(1+\varepsilon_E)(1+\varepsilon_V), \qquad \eta = (1-\alpha)\lambda_1\varepsilon_E(1+\varepsilon_V) + \varepsilon_V.$$

If $q < 1$, *let* $\Delta = \eta\|X^{(0)}\|_\pi/(1-q)$. *Every* $L \ge 1$ *satisfies*

$$\left(\alpha\sqrt{\mathcal{E}_P(X^{(0)})} - \Delta\right)_+^2 \le \mathcal{E}_P(X^{(L)}) \le \left(\sqrt{\mathcal{E}_P(X^{(0)})} + \Delta\right)^2,$$

where $(t)_+ = \max\{t, 0\}$ *and* $\|X\|_\pi^2 = \mathrm{tr}(X^\top \Pi X)$. *In particular, if* $\Delta < \alpha\sqrt{\mathcal{E}_P(X^{(0)})}$, *the lower bound is strictly positive and independent of* L.

Proof of Theorem A.1. Using the reference trajectory and deviation estimate established above, we now bound the energy at each finite depth. The spectrum of $I - P/\lambda_1$ in the weighted inner product lies in $[0, 1]$. Thus $\sqrt{\mathcal{E}_P}$ is a seminorm bounded by $\|\cdot\|_\pi$, and

$$\left|\sqrt{\mathcal{E}_P(X^{(L)})} - \sqrt{\mathcal{E}_P(Y^{(L)})}\right| \le e_L \le d_L N_0.$$

Combining this with the energy bounds for $Y^{(L)}$ gives the finite-depth estimate

$$\left(\alpha\sqrt{\mathcal{E}_P(X^{(0)})} - d_L N_0\right)_+^2 \le \mathcal{E}_P(X^{(L)}) \le \left(\sqrt{\mathcal{E}_P(X^{(0)})} + d_L N_0\right)^2. \tag{12}$$

With q, η as in the theorem, $q_l \le q$ and $\eta_l \le \eta$. If $q < 1$, then $d_L \le \eta(1-q^L)/(1-q) \le \eta/(1-q)$, proving the uniform bound in the theorem. The layer-specific estimate can remain informative even when this uniform sufficient condition fails. □.

B Supplementary properties of fixed propagation

This section provides properties of the propagation that support the discussion in Sections 3 and 4 but are not needed for the proofs in Appendix A. We first give a spectral representation of the operator-relative energy and show that it is equivalent to the weighted distance to the collapse subspace (Proposition B.3). We then rewrite the energy as the Dirichlet form of a reversible Markov chain (Proposition B.4) and show that no single seminorm can serve every shared-incidence operator (Theorem B.5 and Corollary B.6). Finally, we record the single-stage analogue of Theorem 4.4 (Proposition B.7) and sharper limit bounds for initial-residual propagation (Proposition B.8). Throughout this section, $\mathcal{C}_P = \{\psi c^\top : c \in \mathbb{R}^h\}$ denotes the collapse subspace of P.

Spectral representation and distance to collapse. We use the following two notions.

Definition B.1. A propagation operator P is *admissible* if it is nonnegative and irreducible, self-adjoint in the weighted inner product $\langle\cdot,\cdot\rangle_\pi$ induced by its energy weights Π, and has a strictly dominant eigenvalue $\lambda_1 > |\lambda_k|$ for every $k \ge 2$.

Definition B.2. A seminorm N on $\mathbb{R}^{n\times h}$ is *valid* for P if $N(X) = 0$ holds exactly when $X \in \mathcal{C}_P$.

Every operator with a shared-incidence factorization on a connected hypergraph is admissible by Proposition 4.3. For admissible P, the energy has an explicit spectral form and is equivalent to the squared distance to $\mathcal{C}_P$.

Proposition B.3. *Let P be admissible with energy weights Π. In a π-orthonormal eigenbasis $\{w_k\}$ of P with $w_1 = \psi/\|\psi\|_\pi$, write $X = \sum_k w_k c_k^\top$. Then*

$$\mathcal{E}_P(X) = \sum_{k\geq 2} \left(1 - \frac{\lambda_k}{\lambda_1}\right) \|c_k\|_2^2. \tag{13}$$

In particular, $\sqrt{\mathcal{E}_P}$ is valid for P. Moreover, if $d_\pi(X, \mathcal{C}_P)$ denotes the π-weighted Frobenius distance from X to $\mathcal{C}_P$, then

$$\left(1 - \frac{\lambda_2}{\lambda_1}\right) d_\pi(X, \mathcal{C}_P)^2 \leq \mathcal{E}_P(X) \leq \left(1 - \frac{\lambda_n}{\lambda_1}\right) d_\pi(X, \mathcal{C}_P)^2.$$

Proof. Self-adjointness gives $\langle w_j, P w_k\rangle_\pi = \lambda_k \delta_{jk}$, so inserting the expansion into equation 3 yields equation 13. The coefficient of $k = 1$ vanishes and every other coefficient is positive because $\lambda_1 > |\lambda_k|$. Hence $\mathcal{E}_P(X) = 0$ if and only if $c_k = 0$ for all $k \geq 2$, that is, $X \in \mathcal{C}_P$. The π-orthogonal projection of X onto $\mathcal{C}_P$ is $w_1 c_1^\top$, so $d_\pi(X, \mathcal{C}_P)^2 = \sum_{k\geq 2} \|c_k\|_2^2$. Bounding each coefficient in equation 13 by its extreme values over $k \geq 2$ gives both inequalities. □

A stochastic representation. The energy can be read as the Dirichlet form of a reversible Markov chain after rescaling the features by the Perron vector.

Proposition B.4. *Let P be admissible, $\Psi = \mathrm{diag}(\psi)$, and define*

$$Q = \lambda_1^{-1} \Psi^{-1} P \Psi, \qquad \mu_i = \pi_i \psi_i^2 = \ell_i \psi_i.$$

Then $Q \geq 0$ is row-stochastic and reversible with respect to μ, that is, $\mu_i Q_{ij} = \mu_j Q_{ji}$. For $F = \Psi^{-1} X$ with rows f_i,

$$\mathcal{E}_P(X) = \mathrm{tr}\big(F^\top \mathrm{diag}(\mu)(I - Q)F\big) = \frac{1}{2} \sum_{i,j} \mu_i Q_{ij} \|f_i - f_j\|_2^2.$$

Proof. Nonnegativity of Q follows from $P \geq 0$ and $\psi > 0$, and $Q\mathbf{1} = \lambda_1^{-1}\Psi^{-1} P \psi = \mathbf{1}$. Reversibility follows from $\mu_i Q_{ij} = \lambda_1^{-1} \pi_i \psi_i \psi_j P_{ij}$ and the symmetry of ΠP. Substituting $X = \Psi F$ into equation 3 gives $\mathcal{E}_P(X) = \mathrm{tr}(F^\top \Psi \Pi \Psi (I - Q)F) = \mathrm{tr}(F^\top \mathrm{diag}(\mu)(I - Q)F)$, and the pairwise form follows from $Q\mathbf{1} = \mathbf{1}$ and reversibility as in the proof of equation 4. □

In the rescaled coordinates F the collapse direction is constant, which explains the relation between random-walk and symmetric normalizations. The change of coordinates does not alter the zero set in the original feature coordinates.

Obstruction to a universal seminorm. The axiomatic perspective of Rusch et al. (2023) uses a nonnegative, subadditive measure that vanishes exactly on constant node signals. The motivating example in Section 3 shows why that zero set must change for HGNN. The following result shows more generally that a seminorm decays under a family of operators exactly when it vanishes on all of their collapse subspaces. We normalize every operator to spectral radius one so that decay cannot arise from global scale contraction alone.

Theorem B.5. *Let $\mathcal{F}$ be a family of admissible operators with spectral radius one, and let N be a seminorm on $\mathbb{R}^{n\times h}$. Then $N(P^l X) \to 0$ for every $X \in \mathbb{R}^{n\times h}$ and every $P \in \mathcal{F}$ if and only if $N(\psi_P c^\top) = 0$ for every $P \in \mathcal{F}$ and every $c \in \mathbb{R}^h$.*

Proof. Fix $P \in \mathcal{F}$. If $X = \psi_P c^\top \in \mathcal{C}_P$, then $P^l X = X$ for every l because $P\psi_P = \psi_P$, so $N(P^l X) \to 0$ forces $N(X) = 0$. Conversely, suppose N vanishes on $\mathcal{C}_P$. Expanding $X = \sum_k w_k c_k^\top$ in a π-orthonormal eigenbasis gives $P^l X = w_1 c_1^\top + \sum_{k\geq 2} \lambda_k^l w_k c_k^\top$, and $|\lambda_k| < 1$ for $k \geq 2$ implies $P^l X \to w_1 c_1^\top \in \mathcal{C}_P$. A seminorm on a finite-dimensional space is continuous, hence $N(P^l X) \to N(w_1 c_1^\top) = 0$. □

Corollary B.6. *Let $\mathcal{P}$ contain all propagation operators on $\mathcal{H}$ that admit a shared-incidence factorization and have spectral radius one. If a seminorm N satisfies $N(P^l X) \to 0$ for every X and every $P \in \mathcal{P}$, then $N \equiv 0$.*

Proof. Let $R = D_V^{-1} H W D_E^{-1} H^\top$ be the random-walk operator on $\mathcal{H}$ and fix any $\psi \in \mathbb{R}^n_{>0}$. The operator $\Psi R \Psi^{-1}$ has the shared-incidence factorization equation 5 with $b_v = \psi_v / d_v$, $f_e = w_e$, $a_e = 1/|e|$, and $c_v = 1/\psi_v$, so it belongs to $\mathcal{P}$ and is admissible by Proposition 4.3. Its spectral radius is one and its collapse direction is ψ, since $\Psi R \Psi^{-1} \psi = \Psi R \mathbf{1} = \psi$. By Theorem B.5, $N(\psi c^\top) = 0$ for every $\psi > 0$ and every c. Every $u \in \mathbb{R}^n$ is a difference of two positive vectors, so subadditivity gives $N(u c^\top) = 0$ for every rank-one matrix, and subadditivity again extends this to every matrix. □

This obstruction motivates a family of measures indexed by the propagation operator, rather than a single seminorm for all collapse directions.

Single-stage contraction. Theorem 4.4 treats the two-stage layer equation 1. The same argument gives the classical single-stage statement.

Proposition B.7. *Let $P = BA$ admit a shared-incidence factorization on a connected hypergraph, with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, and let $X^{(l+1)} = \sigma(P X^{(l)} \Theta^{(l)})$. Assume σ is a shared 1-Lipschitz row map that is positively homogeneous unless $\psi \propto \mathbf{1}$. If $s = \sup_l \|\Theta^{(l)}\|_2 < \infty$, then*

$$\mathcal{E}_P(X^{(L)}) \leq (\lambda_2 s)^{2L} \, \mathcal{E}_P(X^{(0)}).$$

In particular, the energy decays exponentially when $\lambda_2 s < 1$.

Proof. By equation 13, propagation multiplies the kth coefficient by λ_k, so

$$\mathcal{E}_P(PX) = \sum_{k \geq 2} \lambda_k^2 \left(1 - \frac{\lambda_k}{\lambda_1}\right) \|c_k\|_2^2 \leq \lambda_2^2 \, \mathcal{E}_P(X).$$

The channel map and activation bounds in the proof of Theorem 4.4 give $\mathcal{E}_P(\sigma(U\Theta)) \leq \|\Theta\|_2^2 \mathcal{E}_P(U)$. Combining the two estimates and iterating over L layers proves the claim. □

The contraction condition is sufficient and depends on both the propagation spectrum and the learned transformations. When $\lambda_1 < 1$, absolute energy decay can also reflect shrinking feature magnitudes and should not be read as a scale-invariant rate of collapse.

Linear residual gains and fixed-point bounds. The proof of Theorem 4.6 shows that the linear reference trajectory multiplies the kth eigenmode of $X^{(0)}$ by the gain $g_k(L) \in [\alpha, 1]$. Passing to the limit gives sharper constants than the depth-uniform bounds $\alpha^2 \mathcal{E}_P(X^{(0)}) \leq \mathcal{E}_P(X^\star) \leq \mathcal{E}_P(X^{(0)})$.

Proposition B.8. *Under the assumptions of Theorem 4.6, the limit $X^\star = \alpha(I - (1-\alpha)P)^{-1} X^{(0)}$ satisfies*

$$\left(\frac{\alpha}{1 - (1-\alpha)\lambda_n}\right)^2 \mathcal{E}_P(X^{(0)}) \leq \mathcal{E}_P(X^\star) \leq \left(\frac{\alpha}{1 - (1-\alpha)\lambda_2}\right)^2 \mathcal{E}_P(X^{(0)}).$$

Both constants are attained by nonzero initial features supported on the eigenmode with eigenvalue λ_n or λ_2, respectively.

Proof. With $q_k = (1-\alpha)\lambda_k \in [0, 1)$, the gain from the proof of Theorem 4.6 satisfies $g_k(L) \to \alpha/(1-q_k)$, the personalized PageRank resolvent (Klicpera et al., 2019). Hence $X^\star = \sum_k w_k \, \alpha (1-q_k)^{-1} c_k^\top$, and equation 13 gives

$$\mathcal{E}_P(X^\star) = \sum_{k \geq 2} \left(\frac{\alpha}{1 - (1-\alpha)\lambda_k}\right)^2 \left(1 - \frac{\lambda_k}{\lambda_1}\right) \|c_k\|_2^2.$$

The gain $\alpha/(1-(1-\alpha)\lambda_k)$ is increasing in λ_k, and $\lambda_n \leq \lambda_k \leq \lambda_2$ for $k \geq 2$, which gives the two bounds. Equality holds when only the extremal mode is present. □

Since $X^{(L)} \to X^\star$ under Theorem 4.6, these bounds also describe the limit of the full II trajectory. They do not assert the same bounds at every finite depth for learned channel maps; Theorem A.1 covers that case. For a fixed symmetric operator with spectrum in $[-1, 1]$, replacing it by $(I + P)/2$ supplies a nonnegative spectrum and permits the same argument.

C Experimental Details

Dataset Preprocessing. The controlled energy experiment uses the largest connected component of each dataset with self-loops added, as detailed in Appendix C.2. Classification experiment uses the full datasets with self-loops added.

C.1 Architecture Formulations and Implementation Details

Table 1 gives the shared-incidence factors used in the fixed-operator analysis. We distinguish those mathematical operators from the implementations underlying the classification results.

HGNN. The propagation is $P = D_V^{-1/2} H W D_E^{-1} H^\top D_V^{-1/2}$, with $d_v = \sum_e w_e H_{ve}$ (Feng et al., 2019). The experiments use unit hyperedge weights and the processed incidence, including singleton completion. The theoretical formulas also allow positive nonunit weights.

HNHN. The factors in Table 1 describe the degree- and cardinality-weighted means of Dong et al. (2020). To avoid confusing these architecture parameters with the residual weight, denote the HNHN exponents by a_{HNHN} and b_{HNHN} here. The controlled energy experiment uses $a_{\mathrm{HNHN}} = -1.5$ and $b_{\mathrm{HNHN}} = -0.5$; classification tunes them independently of the II parameters. The recovered Base configuration enables the intermediate activation, whereas HNHN-II disables it.

UniGCN. The theoretical and energy-experiment operator is UniGCN, with $\bar{d}_e = |e|^{-1} \sum_{v \in e} d_v$ and factors from Table 1 (Huang & Yang, 2021).

HyperGCN. HyperGCN approximates each hyperedge by a sparse graph selected from node features, with mediator edges connecting the selected endpoints to the other nodes (Yadati et al., 2019). The recovered Base and II configurations both enable mediators and the fast option, which constructs the graph once from the input features. Thus these runs do not test a graph rebuilt at every layer. This graph construction is not one of the shared-incidence factorizations in Table 1. The II search additionally considers the lazy propagation option $(I + P)/2$.

AllSetTransformer. AllSetTransformer uses attention-based multiset aggregation for both node-to-edge and edge-to-node updates (Chien et al., 2021). Its learned attention, normalization, and feedforward transformations are outside the fixed-aggregation assumptions. The source identifier `AllSetformer` denotes AllSetTransformer in the tables and figures.

C.2 Operator-relative energy: protocol and full results

Protocol. The six datasets are Cora, Citeseer, 20News100, House-Committee, Pokec, and NTU2012, all using the processed benchmark incidence with singleton hyperedges. To match the connected-hypergraph setting, every operator is restricted to the same largest component: 1,330/2,708 nodes on Cora, 1,019/3,312 on Citeseer, and 64/14,998 on Pokec; the other three datasets are connected. The Pokec measurement therefore describes only that small component. The HNHN exponents are fixed at -1.5 and -0.5. We use identity channel maps and activations, 256 propagation steps, and residual weight $\alpha = 0.1$, with no training or dropout. An initialization control repeats every trajectory on ten Gaussian feature matrices with 64 channels, each rescaled once to $\|X^{(0)}\|_F^2 = 64n$ and paired across operators and recurrences.

Finite-depth behavior. Attenuation varies across operators and datasets. On 20News100, HNHN retains 0.604 of its initial energy at step 256, while HGNN and UniGCN fall below 10^{-24} for R_P. Thus, the decay guarantee does not imply near-complete collapse at the evaluated depth. Under initial-residual propagation, the final energy ratios range from 0.0134 to 0.972, all above $\alpha^2 = 0.01$ (Table 4).

C.3 Classification protocol and hyperparameter search

Training and model selection The classification results contain 360 configurations: six datasets, ten model identifiers, and depths $L \in \{2, 4, 8, 16, 32, 64\}$. Each configuration is trained for 200 full-

Table 4. Operator-relative energy at step 256 divided by its initial value, using the original benchmark features on the selected component. Each operator is evaluated with base and initial-residual propagation ($\alpha = 0.1$).

	HGNN		HNHN		UniGCN	
Dataset	Base	Restart	Base	Restart	Base	Restart
Cora	6.68×10^{-6}	0.0456	1.04×10^{-4}	0.126	3.15×10^{-10}	0.0371
Citeseer	3.76×10^{-5}	0.0468	3.28×10^{-4}	0.148	1.26×10^{-7}	0.0402
20News100	$< 10^{-24}$	0.0237	0.604	0.972	$< 10^{-24}$	0.0202
House-Committee	$< 10^{-24}$	0.0146	0.00315	0.379	$< 10^{-24}$	0.0134
Pokec	1.90×10^{-5}	0.0375	1.82×10^{-4}	0.093	4.72×10^{-7}	0.0355
NTU2012	2.74×10^{-4}	0.0606	0.00336	0.192	1.35×10^{-6}	0.0581

batch epochs using negative log-likelihood loss. Each seed draws a $50/25/25$ train/validation/test split. Validation accuracy selects the epoch and tuning configuration, and the selected configuration is evaluated with ten seeds. Tuning trials use three seeds.

Each (method, dataset, depth) combination is tuned separately with Optuna (Akiba et al., 2019). The nominal search budget is 30 trials; 60 of the 360 configurations have fewer than 30 completed trials. The recovered counts are retained with each configuration, so the historical sweep should not be interpreted as having an identical completed budget for every cell.

Search spaces and fixed options Tables 5 and 6 list the proposed search values. In particular, HyperGCN uses the fast mediator graph in both variants, and the HNHN intermediate-activation settings differ as described above. The UniGCNII optimizer fixes its learning rate to 0.01 and uses weight decay 0.01 for its regularized parameter group and 0.0005 for the remaining parameters; the suggested `lr` and `wd` do not override these values.

Table 5. Shared search parameters.

Parameter	Values
Searched for all ten architectures	
`lr`	$\{10^{-3}, 5\cdot10^{-3}, 10^{-2}, 5\cdot10^{-2}, 10^{-1}\}$
`wd`	$\{0, 10^{-5}, 10^{-4}, 10^{-3}\}$
`dropout`	$\{0.0, 0.1, \ldots, 0.7\}$
`MLP_hidden`	$\{64, 128, 256, 512\}$ if $L \leq 16$, else $\{64, 128, 256\}$
Searched for the five II variants only	
α (`restart_alpha`)	$\{0.05, 0.10, \ldots, 0.50\}$
λ (`lamda`)	$\{0.5, 1.0, 1.5, 2.0\}$

Table 6. Architecture-specific search parameters. The implementation names follow the code.

Implementation	Additional parameters
HNHN	`HNHN_alpha` $\in \{-1.5, -1, -0.5, 0\}$, `HNHN_beta` $\in \{-1, -0.5, 0\}$
HNHN-II	as HNHN
AllSetTransformer	`heads` $\in \{1, 4, 8\}$, `MLP_num_layers` $\in \{1, 2\}$, `decoder_num_layers` $\in \{1, 2\}$, `normalization` $\in$ {ln, bn}
AllSetTransformer-II	as AllSetTransformer
UniGCN	`input_drop` $\in \{0, 0.2, 0.4, 0.6\}$, `activation` $\in$ {relu, prelu}, `use_norm` $\in$ {true, false}, `first_aggregate` $\in$ {mean, sum}
UniGCNII	as UniGCN

Reported summaries The displayed best depth is selected by test mean from the six evaluated depths, whereas tuning within each depth uses validation accuracy. Best-of-depth comparisons are

therefore descriptive comparisons of the recorded results, not evaluation at a depth chosen on an independent validation set. Per-comparison p-values do not account for multiple comparisons.