Analysis of Dirichlet Energies as Over-smoothing Measures

Anonymous authors
Paper under double-blind review

Abstract

We analyze the distinctions between two metrics often used to measure over-smoothing: the Dirichlet energies induced by the unnormalized graph Laplacian and the normalized graph Laplacian. We demonstrate that the latter fails to satisfy the axiomatic definition of a node-similarity measure proposed by Rusch *et al.* By formalizing fundamental spectral properties of these two definitions, we highlight critical distinctions necessary to select the metric that is spectrally compatible with the GNN architecture, thereby resolving ambiguities in monitoring the dynamics.

1 Introduction

One of the most analyzed problems in Graph Neural Networks (GNNs) is over-smoothing, that is usually described as the exponential convergence of node embeddings to a common vector through the GNN layers. One of the more frequently used metrics to analyze both theoretically and empirically over-smoothing is the Dirichlet energy, that is induced by the graph Laplacian, with different possibilities as analyzed in the next section. A formal axiomatic definition of over-smoothing, based on the definition of a "total over-smoothing" state where all node embeddings are identical, has been proposed in Rusch et al. (2023). A key axiom of the proposal is that a smoothness measure should be zero if and only if this state is reached. In this paper, we point out that the widely-used Dirichlet energy induced by the *normalized* graph Laplacian does not satisfy this axiom. Recently, some other issues in adopting Dirichlet energies in order to measure over-smoothing were pointed out in Zhang et al. (2025), where it is explained that Dirichlet energy induced by the normalized Laplacian vanishes when node embeddings converge to the subspace spanned by $\boldsymbol{v}$ s.t. $v_i = \sqrt{d_i + 1}$, with d_i the degree of the i-th node, which corresponds to the kernel of the normalized Laplacian augmented with self-loops, and carries information dependent only on node degrees. The difference between our work and Zhang et al. (2025) is that while they also use this convergence to argue that the Dirichlet metric is too strict as a definition (motivating their rank-based proposal), they do not focus on the fact that this represents a formal violation of the Rusch et al. (2023) axiom. We, instead, start our analysis from this axiom violation. Our primary objective is to demonstrate that the diagnostic validity of Dirichlet energy is not absolute, but strictly conditional on its consistency with the operator governing the GNN dynamics. We use the distinction between the standard Graph Laplacian, Δ, and its normalized counterpart, Δ_{norm}, as a paradigmatic example of this broader principle. While often treated as interchangeable, we show that measuring normalized dynamics with unnormalized metrics leads to a breakdown in detection. In this incompatible regime, the standard energy fails to capture genuine over-smoothing (convergence to the normalized kernel); conversely, observing a vanishing energy becomes a misleading indicator: since the dynamics do not drive the signal to the metric's kernel, such decay strictly implies the collapse of embedding norms (over-shrinking), rather than geometric alignment. A pertinent example is found in recent work Arroyo et al. (2025), where the analysis of normalized models using the standard energy leads to the conclusion that over-smoothing is driven by contractive layers. As we argue, this is an artifact of the metric mismatch. While Johnson and Zhang Xu et al. (2019) previously highlighted the bias introduced by normalization in standard semi-supervised learning, our work investigates the spectral consequences of this discrepancy specifically within the GNN framework. We analyze how the non-commutativity between the filter and the target metric generates spectral dispersion, thereby invalidating standard approaches to monitoring over-smoothing.

In light of these results, the main contributions of this paper are summarized as follows:

- **Theoretical requirement for compatible metrics:** we formalize the spectral incompatibility (i.e., non-commutativity) between standard and normalized Laplacians. We establish that in linear propagation frameworks (e.g., SGC), using an energy metric consistent with the GNN filter is strictly necessary to correctly diagnose geometric collapse and the subsequent loss of linear separability, thus avoiding the conflation of over-smoothing with over-shrinking.
- **Empirical limitations and task-aware diagnosis:** we demonstrate the fragility of purely structural proxy metrics. Minor architectural additions, such as learnable biases, shift the asymptotic convergence and entangle structural metrics with optimization artifacts. We argue that in the general framework of non-linear deep GNN architectures, isolating topological over-smoothing via a single structural scalar may be fundamentally ill-posed. Consequently, we consider as a future direction of analysis, the diagnosis of over-smoothing through a task-aware lens, prioritizing the observable degradation of downstream predictive performance over purely structural proxies.

2 Preliminaries and Notation

We denote scalars by lowercase letters (e.g., x), vectors by bold lowercase letters (e.g., $\mathbf{x}$), and matrices by bold uppercase letters (e.g., $\mathbf{X}$). The i-th entry of a vector $\mathbf{x}$ is denoted by x_i, and the entry at row i and column j of a matrix $\mathbf{X}$ is denoted by X_{ij}. We denote the vector of all ones as $\mathbf{1} \in \mathbb{R}^N$ and the Kronecker delta as δ_{ij}. The standard inner product between two vectors $\mathbf{u}, \mathbf{v}$ is denoted by $\langle \mathbf{u}, \mathbf{v} \rangle$ or $\mathbf{u}^\top \mathbf{v}$.

We consider an undirected, connected[1] graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V} = \{1, \ldots, N\}$ and edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, where $N = |\mathcal{V}|$. The graph is associated with a node feature matrix $\mathbf{X} \in \mathbb{R}^{N \times m}$, where the i-th row, denoted by $\mathbf{x}_i \in \mathbb{R}^m$, represents the feature vector of node i.

Let $\mathbf{A} \in \{0,1\}^{N \times N}$ be the adjacency matrix, where $A_{ij} = 1$ if $(i,j) \in \mathcal{E}$ and 0 otherwise. Let $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_N)$ be the degree matrix, where $d_i = \sum_j A_{ij}$ corresponds to the degree of node i, and let $\mathcal{N}_i$ denote the set of neighbors of node i. Finally, we define the unnormalized Laplacian as $\Delta = \mathbf{D} - \mathbf{A}$, the normalized Laplacian as $\Delta_{\text{norm}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$; the normalized Laplacian augmented with self-loops as $\tilde{\Delta}_{\text{norm}} = \mathbf{I} - \widetilde{\mathbf{D}}^{-1/2}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-1/2} = \widetilde{\mathbf{D}}^{-1/2}\Delta\widetilde{\mathbf{D}}^{-1/2}$, where $\widetilde{\mathbf{D}} = \boldsymbol{D} + \boldsymbol{I}$, $\widetilde{\mathbf{A}} = \boldsymbol{A} + \boldsymbol{I}$, and $\tilde{\mathbf{A}}_{\text{norm}} = \widetilde{\mathbf{D}}^{-1/2}\tilde{A}\widetilde{\mathbf{D}}^{-1/2}$.

3 Axiomatic Formulation

Consider the definition of node-similarity given in Rusch et al. (2023):

Definition 1. Over-smoothing Let $\mathcal{G}$ be an undirected, connected graph and let $\boldsymbol{X}^{(n)} \in \mathbb{R}^{N \times m}$ denote the n-th layer hidden features of an N-layer GNN defined on $\mathcal{G}$. Moreover, we call $\mu : \mathbb{R}^{N \times m} \to \mathbb{R}_{\geq 0}$ a node-similarity measure if it satisfies the following axioms:

1. $\exists\, \boldsymbol{c} \in \mathbb{R}^m$ with $\boldsymbol{x}_i = \boldsymbol{c}$ for all nodes $i \in \mathcal{V} \Leftrightarrow \mu(\boldsymbol{X}) = 0$, for $\boldsymbol{X} \in \mathbb{R}^{N \times m}$
2. $\mu(\boldsymbol{X} + \boldsymbol{Y}) \leq \mu(\boldsymbol{X}) + \mu(\boldsymbol{Y})$, for all $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{N \times m}$

Over-smoothing with respect to μ is then defined as the layer-wise exponential convergence of the node-similarity measure μ to zero, i.e.,

3. $\mu(\boldsymbol{X}^{(n)}) \leq C_1 e^{-C_2 n}$, for $n = 0, \ldots, N$ with some constants $C_1, C_2 > 0$.

One example of the most common over-smoothing measure used in literature is the Dirichlet energy. It can be defined as a function of the Laplacian, and depending on the definition of the Laplacian chosen, it can happen that it could not induce a node-similarity measure with respect to the previous definition.

Definition 2. Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a graph and $\boldsymbol{X} \in \mathbb{R}^{N \times m}$ be the matrix of the node embeddings associated to $\mathcal{G}$. The following definitions of Dirichlet energy are considered:

[1] This assumption simplifies the spectral analysis by ensuring that the eigenvalue $\lambda_0 = 0$ of the Laplacian has multiplicity 1.

- from Rusch et al. (2023)[2], induced by Δ:

$$\mathcal{E}_{\Delta}(\boldsymbol{X}) = \frac{1}{N}\mathrm{tr}(\boldsymbol{X}^\top \Delta \boldsymbol{X}) = \frac{1}{N}\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{N}_i} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\,;$$

- induced by Δ_{norm}:

$$\mathcal{E}_{\Delta_{\mathrm{norm}}}(\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{X}^\top \Delta_{\mathrm{norm}} \boldsymbol{X}) = \sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{N}_i} \left\| \frac{\boldsymbol{x}_i}{\sqrt{d_i}} - \frac{\boldsymbol{x}_j}{\sqrt{d_j}} \right\|_2^2;$$

- induced by $\tilde{\Delta}_{\mathrm{norm}} := \boldsymbol{I} - (\boldsymbol{D}+\boldsymbol{I})^{-1/2}(\boldsymbol{A}+\boldsymbol{I})(\boldsymbol{D}+\boldsymbol{I})^{-1/2}$, version with self-loops considered in Cai & Wang (2020):

$$\mathcal{E}_{\tilde{\Delta}_{\mathrm{norm}}}(\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{X}^\top \tilde{\Delta}_{\mathrm{norm}} \boldsymbol{X}) = \sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{N}_i} \left\| \frac{\boldsymbol{x}_i}{\sqrt{1+d_i}} - \frac{\boldsymbol{x}_j}{\sqrt{1+d_j}} \right\|_2^2.$$

As reported in Rusch et al. (2023), $\sqrt{\mathcal{E}_{\Delta}}$ satisfies the definition of a node-similarity measure. Nevertheless, there are conditions of exponential convergence of the node embeddings to a subspace of dimension 1 (e.g., the kernel of Δ_{norm}) that should be considered conditions of over-smoothing, but which Definition 1 does not take into account.

Therefore, it seems worth pointing out that $\mathcal{E}_{\Delta_{\mathrm{norm}}}$ does not satisfy the first axiom of node-similarity measure. Indeed, when a graph is **not regular**[3], the required implication does not hold:

$$\forall\, i \in \mathcal{V},\ \boldsymbol{x}_i = \boldsymbol{c} \Rightarrow \mathcal{E}_{\Delta_{\mathrm{norm}}}(\boldsymbol{X}) = \sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{N}_i} \|\boldsymbol{c}\|_2^2 \left(\frac{1}{\sqrt{d_i}} - \frac{1}{\sqrt{d_j}}\right)^2 > 0 \quad \text{if } \exists\, \{i,j\} \in \mathcal{E} \mid d_i \neq d_j.$$

Hence, Dirichlet energy induced by the normalized Laplacian does not constitute a node-similarity measure. An analogous result holds for $\mathcal{E}_{\tilde{\Delta}_{\mathrm{norm}}}$.

A formalization of the above discussion is given in the following. In particular, all the obtained formal results are consequences of the following straightforward mathematical fact:

Theorem 1. *In general, Δ_{norm} and Δ do not commute, and therefore they are not simultaneously diagonalizable.*

Proof. See Appendix D. □

The non-commutativity between the two Laplacians arises from the non-commutativity between $\boldsymbol{D}$ and $\boldsymbol{A}$. A straightforward example of $[\boldsymbol{A}, \boldsymbol{D}] = \boldsymbol{AD} - \boldsymbol{DA} \neq \boldsymbol{0}$ is given by a four-node graph consisting of a triangle and an additional edge:

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \qquad \boldsymbol{D} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \implies \boldsymbol{AD} \neq \boldsymbol{DA}.$$

One effect of the previous theorem is the following:

[2]The scaling factor $1/N$ is included solely to align with the definition in Rusch et al. (2023); it will be omitted in our subsequent analysis as it does not impact the results for the intended applications and is typically introduced to stabilize the value ranges of Dirichlet energy, allowing for the comparison of energies across graphs with varying node counts.

[3]If a graph is regular, both $\mathcal{E}_{\Delta_{\mathrm{norm}}}$ and $\mathcal{E}_{\tilde{\Delta}_{\mathrm{norm}}}$ are proportional to $\mathcal{E}_{\Delta}$, hence they satisfy the first axiom.

Proposition 1. *Let $\mathcal{G}$ be an undirected, connected and* ***non-regular*** *graph. Let $\boldsymbol{M}_\mu \in \mathbb{C}^{N\times N}$ be any operator that induces a node-similarity measure via $\mu(\boldsymbol{X}) := tr(\boldsymbol{X}^\top \boldsymbol{M}_\mu \boldsymbol{X}) \geq 0$. Then, the normalized operator*

$$\boldsymbol{D}^{-1/2}\boldsymbol{M}_\mu \boldsymbol{D}^{-1/2}$$

cannot induce a node-similarity measure.

Proof. See Appendix D. □

In particular:

Corollary 1. *The normalized Laplacian Δ_{norm} does not induce a node-similarity measure.*

Remark 1. All the previous results hold also for the normalized Laplacian augmented with self-loops, $\widetilde{\Delta}_{\text{norm}}$. This follows from the identity $\widetilde{\Delta}_{\text{norm}} = (\boldsymbol{D}+\boldsymbol{I})^{-1/2}\Delta(\boldsymbol{D}+\boldsymbol{I})^{-1/2}$, which is algebraically equivalent to applying the substitution $d_i \mapsto d_i + 1$ to the degree matrix $\boldsymbol{D}$ within the definition of Δ_{norm}.

Hence, analysis of over-smoothing of GCNs augmented with self-loops should be based on the Dirichlet energy definition induced by $\tilde{\Delta}_{\text{norm}}$, since it measures the smoothness with respect to the Laplacian that governs the evolution of graph signals. The simplest definition of a layer of a GCN is $\sigma(\tilde{\boldsymbol{A}}_{\text{norm}}\boldsymbol{X}\boldsymbol{W}) = \sigma\big((\boldsymbol{I} - \tilde{\Delta}_{\text{norm}})\boldsymbol{X}\boldsymbol{W}\big)$, which applies an element-wise activation to the output of a low pass filter. Asserting that a GNN implements a filter implies that it attenuates frequencies consistently w.r.t. a fixed Laplacian. However, since Δ_{norm} and Δ do not commute, a filter defined on one implies spectral dispersion on the other.

4 Indeterminacy

Spectral analysis is straightforward when GNN layers act as consistent spectral filters (e.g., polynomials) within a single Laplacian eigenvector basis. However, as established in the previous section, a consistent Laplacian choice is prerequisite for unbiased over-smoothing analysis. Here, we analyze the evolution of a graph signal through a GNN, assuming the composition of its first k layers acts as a spectral filter $f_k(\Delta_{\text{norm}})$. Formally, applying a spectral function f_k to a diagonalizable operator strictly preserves its original eigenvectors (and thus its eigenspaces) while applying the scalar function $f_k(\cdot)$ element-wise to its eigenvalues. Therefore, given the spectral decomposition $\Delta_{\text{norm}} = U\Lambda U^\top$, the operator mapping the signal at layer k is explicitly defined as $f_k(\Delta_{\text{norm}}) = U f_k(\Lambda) U^\top$. For instance, the standard propagation operator $\boldsymbol{A}_{\text{norm}} = \boldsymbol{I} - \Delta_{\text{norm}}$ can be straightforwardly viewed as a spectral function of Δ_{norm} defined by $f(\lambda) = 1 - \lambda$.

We demonstrate how the spectral inconsistency between the driving operator (Δ_{norm}) and the measuring operator (Δ) leads to misleading analytical results. Specifically, we evaluate the standard Dirichlet energy $\mathcal{E}_\Delta$ of a signal evolving under Δ_{norm} dynamics.

To formalize this spectral mismatch, let us define two orthonormal bases of $\mathbb{R}^{|\mathcal{V}|}$:

- $\mathcal{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$: the eigenvectors of the unnormalized Laplacian Δ, with associated eigenvalues $\sigma_\Delta = \{\lambda_1, \ldots, \lambda_N\}$.
- $\mathcal{V} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\}$: the eigenvectors of the normalized Laplacian Δ_{norm}, with associated eigenvalues $\sigma_{\Delta_{\text{norm}}} = \{\mu_1, \ldots, \mu_N\}$.

The relationship between these two spectral domains is encoded in the **change-of-basis matrix** (or superposition matrix) $\boldsymbol{P} \in \mathbb{R}^{N\times N}$, defined by the inner products of the basis vectors:

$$P_{ij} = \langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle = \boldsymbol{u}_i^\top \boldsymbol{v}_j.$$

If a graph signal $\boldsymbol{X}^{(0)}$ evolves according to the filter $f_k(\Delta_{\text{norm}})$, the filtered signal $\boldsymbol{X}^{(k)}$ remains diagonal in the basis $\mathcal{V}$. However, when observing this signal through the lens of the standard Dirichlet energy $\mathcal{E}_\Delta$, we project these filtered components onto the basis $\mathcal{U}$.

By expanding the trace definition of the Dirichlet energy using these bases, we observe that the energy contribution is no longer localized to single frequencies. Instead, it depends on the off-diagonal elements of $\boldsymbol{P}$. The standard energy of the filtered signal can be expressed as:

$$\mathcal{E}_{\Delta}(\boldsymbol{X}^{(k)}) = \sum_{i=1}^{N} \lambda_i \left\| \sum_{j=1}^{N} P_{ij} f_k(\mu_j)(\boldsymbol{v}_j^\top \boldsymbol{X}^{(0)}) \right\|^2 .$$

For the full derivation of the spectral expansion involving the triple summation of projection coefficients, we refer the reader to Appendix A.

This formulation highlights the phenomenon of **spectral dispersion**. In general, on non-regular graphs, the operators Δ and Δ_{norm} do not commute, implying that $\boldsymbol{P}$ is a dense matrix. Consequently, a signal that is perfectly "smooth" (low-frequency) with respect to Δ_{norm}, such as a vector $\boldsymbol{v}_{\text{ker}}$ that lives in the kernel, does not map to a single low-frequency component in Δ. Instead, it projects onto multiple components of $\mathcal{U}$ (i.e., $P_{i,\text{ker}} \neq 0$ for many i), potentially including high-frequency modes.

Thus, $\mathcal{E}_{\Delta}$ effectively integrates "noise" from parts of the spectrum that are irrelevant to the actual dynamics. From Theorem 1, it follows that if a GNN implements a spectral function with respect to Δ_{norm}, it generally fails to act as a consistent spectral filter with respect to Δ. Specifically, a low-pass filter in the Δ_{norm} domain may effectively amplify high-frequency components in the Δ domain due to spectral dispersion.

The Special Case of Regular Graphs and Interpretation

The only exception arises when the graph is d-regular. In this case, $\Delta_{\text{norm}} = \frac{1}{d}\Delta$, meaning the operators share the same eigenspaces. The change-of-basis matrix $\boldsymbol{P}$ becomes the identity (up to permutation), and the dispersion vanishes ($P_{ij} = \delta_{ij}$). The expression then simplifies to the standard consistent form:

$$\mathcal{E}_{\Delta}(\boldsymbol{X}^{(k)}) = \sum_{i=1}^{N} \lambda_i \left| f_k\left(\frac{\lambda_i}{d}\right) \right|^2 \|\boldsymbol{u}_i^\top \boldsymbol{X}^{(0)}\|^2.$$

This simplification implies that the GNN behaves like a spectral filter f_k with respect to Δ frequencies *only* when processing regular graphs.

Consequently, the frequency band of Δ that passes through the network depends heavily on the degree distribution of the nodes. For regular graphs, the GNN applies the same filter up to a rescaling; for general graphs, the resulting effective filter depends on the decomposition of Δ_{norm} frequencies into Δ frequencies, varying across different input topologies.

Asserting that a GNN *implements* a filter implies that, *for all* processed graphs, it must attenuate or amplify specific frequencies of the signal consistently with respect to a fixed Laplacian definition (i.e., regardless of the degree distribution $\boldsymbol{D}$). However, if the filter function f is uniquely defined on the spectrum of one Laplacian, generally it cannot preserve its spectral behavior with respect to the other, as frequencies are transformed depending on the graph structure.

Consider, for instance, the kernel of Δ_{norm}, which is spanned by the vector $\boldsymbol{D}^{1/2}\mathbf{1}$. In a graph exhibiting high degree variability (containing both high-degree and low-degree nodes), the frequency associated with this vector is zero with respect to Δ_{norm}. Conversely, this same vector $\boldsymbol{D}^{1/2}\mathbf{1}$ is generally not in the kernel of the unnormalized Laplacian Δ (which is spanned by $\mathbf{1}$). Instead, it appears as a superposition of various non-zero frequencies with respect to Δ. In this sense, an eigen-signal of one Laplacian, associated with a specific frequency, exhibits spectral dispersion when analyzed with respect to the other: instead of being frequency-localized, it appears scattered across multiple components of the measuring spectrum.

4.1 Over-shrinking and Over-smoothing

Overlooking the formal distinction between the smoothing metrics associated with different Laplacians (e.g., $\mathcal{E}_{\Delta}$ vs. $\mathcal{E}_{\tilde{\Delta}_{\text{norm}}}$) can lead to significant interpretative pitfalls. Consider, for instance, the analysis of GNN

dynamics governed by $\tilde{A}_{\text{norm}} = \boldsymbol{I} - \tilde{\Delta}_{\text{norm}}$. A consistent analysis would require employing the associated energy $\mathcal{E}_{\tilde{\Delta}_{\text{norm}}}$, and relying on the standard Dirichlet energy $\mathcal{E}_{\Delta}$ as an indicator for over-smoothing introduces a fundamental methodological mismatch. As established in the previous sections, $\mathcal{E}_{\Delta}$ measures convergence specifically to the topology-independent subspace span($\mathbf{1}$). Since in general the dynamics of a GCN does not drive the signal towards this state, but rather towards the topology-dependent kernel of $\tilde{\Delta}_{\text{norm}}$, the standard energy $\mathcal{E}_{\Delta}$ is not guaranteed to vanish. Indeed, due to the spectral misalignment, $\mathcal{E}_{\Delta}$ suffers from a leakage on the kernel of $\tilde{\Delta}_{\text{norm}}$, meaning it remains non-zero even as the signal smooths optimally relative to the dynamics. Consequently, observing $\mathcal{E}_{\Delta} \to 0$ in this regime implies a different phenomenon entirely: it requires the collapse of the embedding norms themselves to zero. This effectively conflates over-smoothing with over-shrinking. If one relies on $\mathcal{E}_{\Delta}$ without accounting for this spectral mismatch, the decay of the energy might be misidentified as a consequence of the contractive nature of the network layers, rather than correctly attributing the asymptotic behavior to the alignment with the Laplacian kernel. This distinction underscores the importance of consistency: to avoid misinterpreting spectral behavior, the chosen metric must commute with the operator governing the dynamics.

A critical ambiguity arises in the literature when over-smoothing is analyzed through a single scalar metric. For instance, Arroyo et al. (2025) attributes over-smoothing to the contractive nature of GNNs, observing that vanishing gradients drive the energy $\mathcal{E}_{\Delta}$ to zero. However, this perspective conflates two fundamentally different phenomena. To resolve this, we propose a strict decoupling based on systemic universality. We refer to spectral over-smoothing as the regime where, for arbitrary input graphs and generic signals, the output aligns with the low-frequency kernel of the underlying Laplacian while maintaining a strictly positive norm. Conversely, over-shrinking denotes the unconditional collapse of the embedding magnitudes to zero, irrespective of the graph topology or signal initialization. By monitoring only the energy without these universal constraints, one risks misinterpreting the trivial null-solution as spectral convergence.

4.2 End-to-End Topological Over-smoothing

While the previous sections establish that evaluating the standard Dirichlet energy through normalized dynamics introduces spectral leakage and could misidentify over-shrinking (Arroyo et al., 2025), an inversion occurs when shifting focus to the final, discrete predictions. Current literature predominantly formalizes over-smoothing as a layer-wise property of latent representations $H^{(l)}$ (Rusch et al., 2023). Under that paradigm, representations in a Simple GCN (Wu et al., 2019) asymptotically collapse to $\ker(\tilde{\Delta}_{\text{norm}})$, which preserves degree-dependent topological structure and prevents the standard energy $\mathcal{E}_{\Delta}(H^{(l)})$ from vanishing.

However, this perspective overlooks the role of the non-differentiable classification head $\Phi(\cdot)$ (e.g., `argmax` applied to the final linear projection). Let $\hat{Y}^{(K)} \in \{0,1\}^{|\mathcal{V}| \times C}$ be the discrete one-hot predictions generated by an Simple GCN architecture that performs node classification with C classes trained with K propagation steps and a depth-specific linear classifier $W^{(K)}$. As $K \to \infty$, the latent representations take the asymptotic form $\mathbf{h}_i^{(\infty)} \propto \sqrt{d_i}\mathbf{c}$, for a certain $\mathbf{c} \in \mathbb{R}^m$. Crucially, when $\Phi(\cdot)$ is applied to the resulting logits, the positive node-dependent scalar $\sqrt{d_i}$ is factored out by the `argmax` operator, as it does not alter the relative ordering of the class components.

Consequently, the non-differentiable head strips away the degree-normalization artifact, forcing all nodes to abruptly collapse to a single majority class c^*, yielding the uniform discrete output $\hat{Y}^{(\infty)} = \mathbf{1}\mathbf{e}_{c^*}^{\top}$. Because the all-ones vector $\mathbf{1}$ resides in the kernel of the combinatorial Laplacian ($\Delta\mathbf{1} = 0$), the *standard* Dirichlet energy of this discrete output necessarily vanishes:

$$\operatorname{Tr}\left(\hat{Y}^{(K),\top}\Delta\hat{Y}^{(K)}\right) \to 0 \quad \text{as } K \to \infty. \tag{1}$$

This reveals an analytical paradox: while tracking the standard energy through normalized dynamics is misleading due to spectral mismatch, its decay to zero at the discrete prediction level is an inevitable consequence of the topological shift induced by the classification head.

Therefore, we argue that a vanishing energy prior to classification is neither a necessary nor a well-posed condition for diagnosing performance degradation. Instead, diagnosing pathological over-smoothing requires a task-aware, end-to-end approach: it must jointly evaluate the layer-wise evolution of a dynamics-compatible metric and the simultaneous degradation of downstream predictive performance. This formulation decouples

the mathematically valid dynamics of graph diffusion from the empirical pathology of task failure. A formal mathematical proof of this discrete topological collapse is provided in Appendix C.

5 Numerical Examples

In this section, we present a series of empirical experiments designed to illustrate the theoretical points discussed in the previous sections. We aim to verify that over-smoothing conditions with respect to the normalized Laplacian that nothing have to do with over-shrinking, remain invisible when measured through standard proxies such as the standard Dirichlet energy or the Frobenius norm of the embeddings. Providing concrete evidence that these states are only identified by the compatible metric allows to disentangle over-smoothing from simple signal decay.[4]

5.1 Non Over-shrinking GCNs Over-smooth

We present empirical evidence showing that the information loss resulting from spectral collapse is distinct from the decay measured by the standard Dirichlet energy. Specifically, we isolate regimes where the dynamics are incompatible with Δ and asymptotically non-contractive: in these cases, the signal converges to a smooth state with non-zero embedding norms, a phenomenon that standard metrics fail to capture.

To observe the isolated effect of the GCN propagation operator, we focus at first purely on signal smoothing. This experiment eliminates all learnable parameters (i.e., weight matrices $\boldsymbol{W}$).
We select two distinct graphs from the ENZYMES dataset: one graph exhibiting a regular structure and one non-regular graph. We then simulate the pre-processing of a Simple GCN Wu et al. (2019) by iteratively applying only the standard propagation operator for $K = 50$ layers. The state at layer k is thus defined by:

$$\boldsymbol{X}^{(k)} = \boldsymbol{A}_{\text{norm}}\boldsymbol{X}^{(k-1)} = \boldsymbol{A}_{\text{norm}}^{k}\boldsymbol{X}^{(0)},$$

where $\boldsymbol{A}_{\text{norm}} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$.
At each layer $k \in [0, K]$, we track the same three metrics: the Frobenius Norm ($\|\boldsymbol{X}^{(k)}\|_F$), the standard Dirichlet energy ($\mathcal{E}_\Delta(\boldsymbol{X}^{(k)}) = \frac{1}{N}\text{tr}\big(\boldsymbol{X}^{(k)\top}\Delta\boldsymbol{X}^{(k)}\big)$), and the normalized Dirichlet energy ($\mathcal{E}_{\Delta_{\text{norm}}}(\boldsymbol{X}^{(k)}) = \frac{1}{N}\text{tr}\big(\boldsymbol{X}^{(k)\top}\Delta_{\text{norm}}\boldsymbol{X}^{(k)}\big)$). The objective is to directly compare the convergence behavior and smoothing properties of the $\boldsymbol{A}_{\text{norm}}$ operator, contrasting its effect on graphs with regular versus non-regular topologies.

Results, computed in double precision, are shown in Figure 1: for the non-regular graph, $\|\boldsymbol{X}^{(k)}\|_F$ converges to a non-zero value around 10, while $\mathcal{E}_{\Delta_{\text{norm}}}$ decreases faster than $\mathcal{E}_\Delta$, in line with theoretical expectations predicting convergence on each dimension to a graph signal $\boldsymbol{x}$ such that $x_i = c\sqrt{d_i}$ (for some $c \in \mathbb{R}$).
The plot reflects the low-pass filtering nature of a simple GCN without activations: it acts as a spectral function of the normalized Laplacian $\boldsymbol{A}_{\text{norm}} = \boldsymbol{I} - \Delta_{\text{norm}}$, which has eigenvalues in $[-1, 1]$ and thus attenuates signal components over multiple iterations while components residing in the kernel of Δ_{norm} (coinciding with the eigenspace of $\boldsymbol{A}_{\text{norm}}$ associated with eigenvalue 1) are left invariant.

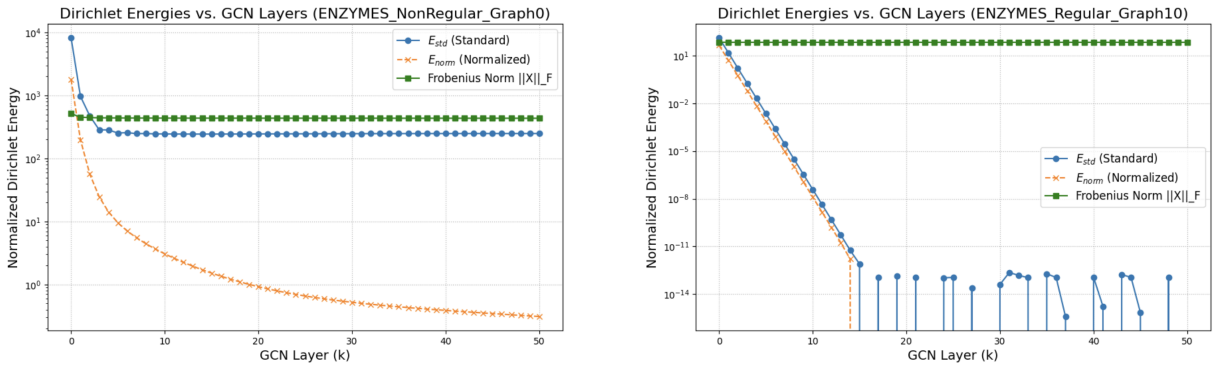
In the regular graph case (Figure 1(b)), $\mathcal{E}_{\Delta_{\text{norm}}}$ theoretically tends asymptotically to zero, indistinguishably from $\mathcal{E}_\Delta$. Empirically, we observe that the values decrease so rapidly that, after fewer than 20 layers, the trend is dominated by numerical instabilities (machine precision artifacts). This comparison highlights a regime where over-smoothing occurs without over-shrinking (i.e., without feature norms vanishing), demonstrating how misleading it can be to conflate the two phenomena.

5.1.1 Note on Weight Matrices

If a linear GCN (without intermediate non-linearities) performs the operation $\boldsymbol{X}^{(k+1)} = \boldsymbol{A}_{\text{norm}}\boldsymbol{X}^{(k)}\boldsymbol{W}^{(k+1)}$ at each layer, the final representation is generally $\boldsymbol{X}^{(K)} = \boldsymbol{A}_{\text{norm}}^{K}\boldsymbol{X}^{(0)}\boldsymbol{W}^{(1)}\dots\boldsymbol{W}^{(K)}$. By the associativity of matrix multiplication, the cascade of layer-specific weight matrices collapses into a single equivalent transformation matrix $\boldsymbol{W} = \boldsymbol{W}^{(1)}\dots\boldsymbol{W}^{(K)}$. Therefore, the operation can be exactly rewritten as:

$$\boldsymbol{X}^{(K)} = (\boldsymbol{A}_{\text{norm}}^{K}\boldsymbol{X}^{(0)})\boldsymbol{W}.$$

[4]The source code to reproduce the experiments is provided as supplementary material.



(a) Non-regular graph (ENZYMES, Graph 1)

(b) Regular graph (ENZYMES, Graph 10)

Figure 1: Comparison of Dirichlet energy evolution and Frobenius norm. (a) On a non-regular graph, we observe spectral dispersion. (b) On a regular graph, the behavior is consistent. Note the logarithmic scale on the y-axis.

This formulation precisely mirrors the Simplified Graph Convolutional Network (SGC) Wu et al. (2019), which explicitly decouples feature propagation from feature transformation. The model first applies the graph shift operator $\boldsymbol{A}_{\text{norm}}$ for K steps directly to the initial input, and only subsequently applies the learned linear transformation $\boldsymbol{W}$. Because $\boldsymbol{A}_{\text{norm}}$ inherently acts as a low-pass filter, the diffused features $\boldsymbol{A}_{\text{norm}}^K \boldsymbol{X}^{(0)}$ inevitably collapse toward the kernel of Δ_{norm} as $K \to \infty$, regardless of the initial input $\boldsymbol{X}^{(0)}$. Consequently, the final multiplication by $\boldsymbol{W}$ merely computes linear combinations of already over-smoothed representations, meaning the learned weights are mathematically powerless to avert the topological collapse.

5.2 Analysis within a Learning Framework

While the previous sections isolated the propagation operator, in this section we verify our theoretical expectations within an end-to-end learning framework. The relation with Section 4.2 is that we experimentally verify that the classification accuracy collapses after a sufficiently large number of application of the diffusion operator $\boldsymbol{A}_{\text{norm}}$ while energy and norm metrics maintain the relative values that we shown to be expected. Specifically, we construct a Simplified Graph Convolution (SGC) model Wu et al. (2019) coupled with a linear classifier. This setting allows us to investigate whether the over-smoothing phenomenon, and its general independence from over-shrinking, persists when structural node features are projected and model weights are optimized via gradient descent.

Specifically, we precompute the diffused features up to K propagation steps: $\boldsymbol{X}^{(k)} = \boldsymbol{A}_{\text{norm}}^k \boldsymbol{X}^{(0)}$. For each $k \in \{0, \ldots, K\}$, a linear classifier devoid of sbias terms is trained to predict node labels directly from the diffused features $\boldsymbol{X}^{(k)}$, fitting a weight matrix $\boldsymbol{W}^{(k)} \in \mathbb{R}^{m \times C}$, such that the logits are given by $\boldsymbol{Z}^{(k)} = \boldsymbol{X}^{(k)} \boldsymbol{W}^{(k)}$.

To ensure the statistical robustness of our observations, the optimization is evaluated across 10 random topological train/validation/test splits. We track the test performance alongside multiple structural and optimization metrics, including the relative and standard Dirichlet energies computed on the network logits $\boldsymbol{Z}^{(k)}$, and the Frobenius norms of the diffused features ($\|\boldsymbol{X}^{(k)}\|_F$), the learned weights ($\|\boldsymbol{W}^{(k)}\|_F$), and the output logits ($\|\boldsymbol{Z}^{(k)}\|_F$).

Experimental Setup and Reproducibility Details. The linear classifier module is trained for a fixed budget of 200 epochs using the Adam optimizer, with an initial learning rate set to 0.1 and a weight decay coefficient configured at 5×10^{-4}. To guarantee statistical robustness and mitigate sampling anomalies, all performance and structural metrics report the mean and standard deviation accumulated across 10 random topological splits of the Largest Connected Component (LCC) of the Cora dataset. Each macro-split enforces a partition composed of 500 nodes for validation and 1000 nodes for test evaluation. Mathematical

reproducibility regarding both the random node partitioning and weight initialization is rigorously anchored by pinning a sequential deterministic seed (ranging from 0 to 9) to each individual run.

To prevent latent numerical instabilities during the longitudinal tracking of Dirichlet energies across the *symlog* scale, the entire codebase infrastructure, spanning from the sparse pre-computation of diffused features to gradient backpropagation, is executed using 64-bit double precision.

We provide experimental results on the LCC of the Cora dataset. The selection of this specific dataset is theoretically motivated by its spectral properties, which directly govern the relaxation rate of the over-smoothing dynamics.

Let the eigenvalues of the normalized Laplacian Δ_{norm} be $0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_N$. The propagation operator $\boldsymbol{A}_{\text{norm}} = \boldsymbol{I} - \Delta_{\text{norm}}$ consequently has eigenvalues $1 - \lambda_i$. While the dynamics asymptotically converge to the eigenspace associated with 1 (the kernel of the Laplacian), the number of layers K required to suppress the orthogonal components is dictated by the spectral gap, corresponding to the second largest eigenvalue of the propagation operator, which we denote as $\mu_2 = 1 - \lambda_2$.

For the Cora dataset, the spectral gap is extremely narrow. Without self-loops, $\mu_2 \approx 0.995$; the introduction of self-loops via the standard renormalization trick Kipf & Welling (2017) further shrinks the spectrum, making the random walk "lazier" and pushing this value to $\mu_2 \approx 0.996$. Consequently, the signal experiences an exceptionally slow relaxation, retaining a noticeable orthogonal residue ($\mu_2^{1000} \approx 2.6 \times 10^{-2}$) even after $K = 1000$ propagation steps.

The accuracy is computed on the masked test nodes. Conversely, the structural metrics, namely the Dirichlet energies and the Frobenius norms ($\|\boldsymbol{Z}^{(k)}\|_F$, $\|\boldsymbol{X}^{(k)}\|_F$), are calculated over the entire unmasked graph considering the total number of nodes N in the connected component. This dual-evaluation regime is theoretically motivated: over-smoothing acts as a global phenomenon that homogenizes the entire system, and the consequent degradation of the local test-set performance is merely a symptom of this global topological collapse.

As illustrated in Figure 2, the progression of the propagation depth K invariably triggers a severe deterioration in accuracy, visibly signifying the assimilation of node embeddings. This drop in accuracy represents the empirical counterpart of the discrete uniform collapse formalized in Appendix C, where the non-differentiable classification head abruptly forces the final predictions into the kernel of the standard Laplacian.

Despite the degradation of structural information, the metrics strictly associated with the magnitude of the signal and of the network parameters, namely $\|\boldsymbol{X}^{(k)}\|_F$, $\|\boldsymbol{W}^{(k)}\|_F$, and consequently $\|\boldsymbol{Z}^{(k)}\|_F$, stabilize at distinctly non-zero values. The training dynamics explicitly converge to learned weights that maintain output magnitudes, proving that over-smoothing dictates the homogenization of the representations rather than their spatial shrinking to the origin.

Further investigations regarding the asymptotic limits at $K = 1000$ and the introduction of bias terms are relegated to Appendix E.

6 Conclusion

In this paper, we formalized the spectral incompatibility between the standard Laplacian Δ and the normalized operators Δ_{norm} and $\tilde{\Delta}_{\text{norm}}$. We focused on a basic algebraic property, the non-commutativity of these operators, to demonstrate the necessity of adopting metrics consistent with the GCN dynamics in order to measure over-smoothing.

This elementary distinction clarifies why the Dirichlet energy can yield misleading results when applied to incompatible dynamics. We highlighted instances where failing to account for this incompatibility led to misleading conclusions, such as not clearly distinguishing between over-smoothing and over-shrinking. We note that a measure of over-smoothing should ensure spectral consistency between the propagation operator and the energy metric to be well-posed.

Our analysis of the linear Simplified Graph Convolution (SGC) framework establishes a rigorous baseline: when the dynamics are strictly governed by a linear propagation operator, one must employ a compati-

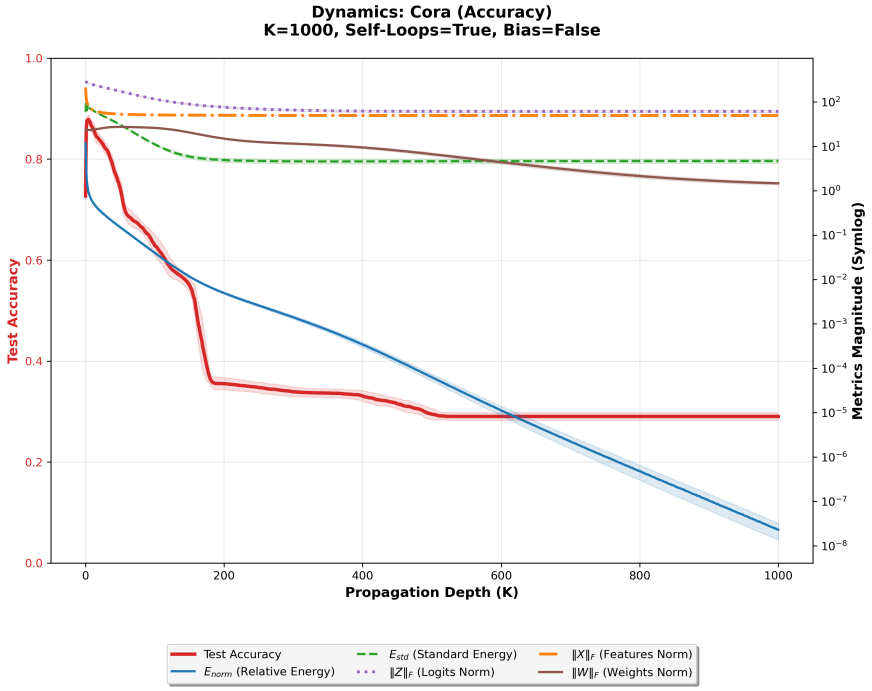


Figure 2: Evolution of test accuracy (red line) alongside energy and norm metrics mapped logarithmically (right axis) over 1000 propagation steps K. The metrics represent the average over 10 random splits, with shaded regions indicating the standard deviation. A bias-free linear classifier is fitted on precomputed features with self loops. Energies are computed on the logits, i.e. before the non-differential classification operator is applied.

ble energy metric (e.g., the normalized Dirichlet energy for the normalized Laplacian) to observe spectral over-smoothing distinct from over-shrinking. Under these pure conditions, the geometric collapse of node representations into the Laplacian kernel prevents the hyperplanes of the linear classifier from shattering the classes, leading to degradation in predictive performance.

Our numerical simulations verify a fundamental fragility in relying on structural proxy metrics within practical machine learning scenarios. As demonstrated, the mere introduction of a learnable bias term shifts the asymptotic convergence, causing structural metrics to capture task-specific optimization artifacts rather than pure topological collapse.

Extrapolating this vulnerability to the general case of deep Graph Neural Networks, where propagation steps are typically interleaved with non-linear activations (e.g., ReLU) and dense feature transformations, the analytical landscape becomes drastically more complex. In such non-linear settings, the dynamics are no longer governed by a single, commuting linear operator. Consequently, the pursuit of a universal, purely structural scalar metric capable of isolating over-smoothing from spatial over-shrinking in highly non-linear architectures may be fundamentally ill-posed.

Ultimately, our analysis suggest a paradigm shift in how over-smoothing is diagnosed in deep learning frameworks. Rather than searching for a mathematically elusive structural proxy, in complex, non-linear GNNs, the most invariant and reliable measure of representational collapse is the phenomenon itself: the unrecoverable deterioration of downstream predictive performance evaluated on the validation set.

Statement on the Use of AI Tools

The authors acknowledge the use of GitHub Copilot and Google Gemini to assist in the preparation of this manuscript. These tools were employed for generating boilerplate text for the experimental section, drafting

and refining the experimental code, and improving the linguistic clarity and flow of the paper. The authors reviewed and edited the generated content and take full responsibility for the accuracy and integrity of the work presented.

References

Alvaro Arroyo, Alessio Gravina, Benjamin Gutteridge, Federico Barbero, Claudio Gallicchio, Xiaowen Dong, Michael Bronstein, and Pierre Vandergheynst. On vanishing gradients, over-smoothing, and over-squashing in gnns: Bridging recurrent and graph learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 74356–74393. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6ba7ebba4d54408b00a2b0275629f625-Paper-Conference.pdf.

Anna Bison and Alessandro Sperduti. On the rank properties of the renormalization trick in gcns. In *ESANN*, 2026. URL https://www.esann.org/sites/default/files/proceedings/2026/ES2026-233.pdf.

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. *ArXiv*, abs/2006.13318, 2020. URL https://api.semanticscholar.org/CorpusID:220042028.

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *CoRR*, abs/2303.10993, 2023. doi: 10.48550/ARXIV.2303.10993. URL https://doi.org/10.48550/arXiv.2303.10993.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6861–6871. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/wu19e.html.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Kaicheng Zhang, Piero Deidda, Desmond Higham, and Francesco Tudisco. Rethinking oversmoothing in graph neural networks: A rank-based perspective, 2025. URL https://arxiv.org/abs/2502.04591.

A Spectral Expansion of the Energy Mismatch

Here we provide the detailed derivation of the Dirichlet energy expansion discussed in Section 4. Let σ_Δ and $\sigma_{\Delta_{\text{norm}}}$ denote the spectra of the two Laplacians. The energy of the signal propagated by $f_k(\Delta_{\text{norm}})$ measured against Δ is:

$$\mathcal{E}_\Delta(\boldsymbol{X}^{(k)}) = \operatorname{tr}\left(\boldsymbol{X}^{(0)^\top}\overline{f_k(\Delta_{\text{norm}})}\Delta f_k(\Delta_{\text{norm}})\boldsymbol{X}^{(0)}\right).$$

Let $\mathcal{B}$ be an arbitrary orthonormal basis of the feature space. Expanding using the resolution of identity for both spectral bases, we obtain the interaction terms dependent on the inner products $\boldsymbol{u}^\top\boldsymbol{v}$.

Substituting the spectral decompositions:

$$\begin{aligned}\mathcal{E}_\Delta(\boldsymbol{X}^{(k)}) &= \sum_{\lambda\in\sigma_\Delta}\lambda\sum_{\boldsymbol{e}\in\mathcal{B}}\boldsymbol{e}^\top\boldsymbol{X}^{(0),\top}\overline{f_k(\Delta_{\text{norm}})}\boldsymbol{u}_\lambda\boldsymbol{u}_\lambda^\top f_k(\Delta_{\text{norm}})\boldsymbol{X}^{(0)}\boldsymbol{e}\\ &= \sum_{\lambda\in\sigma_\Delta}\sum_{\mu\in\sigma_{\Delta_{\text{norm}}}}\sum_{\mu'\in\sigma_{\Delta_{\text{norm}}}}\lambda\overline{f_k(\mu)}f_k(\mu')\sum_{\boldsymbol{e}\in\mathcal{B}}\boldsymbol{e}^\top\boldsymbol{X}^{(0),\top}\boldsymbol{v}_\mu(\boldsymbol{v}_\mu^\top\boldsymbol{u}_\lambda)(\boldsymbol{u}_\lambda^\top\boldsymbol{v}_{\mu'})\boldsymbol{v}_{\mu'}^\top\boldsymbol{X}^{(0)}\boldsymbol{e}.\end{aligned}$$

Here, the terms $\boldsymbol{v}_\mu^\top \boldsymbol{u}_\lambda$ represent the coefficients of the change-of-basis matrix. Rearranging the terms to group the mixing coefficients:

$$\mathcal{E}_\Delta(\boldsymbol{X}^{(k)}) = \sum_{\lambda \in \sigma_\Delta} \lambda \left\| \sum_{\mu \in \sigma_{\Delta_{\text{norm}}}} (\boldsymbol{u}_\lambda^\top \boldsymbol{v}_\mu) f_k(\mu) (\boldsymbol{v}_\mu^\top \boldsymbol{X}^{(0)}) \right\|^2 .$$

B Proofs of Theorems and Propositions

Proof of Theorem 1

To show that the two operators do not commute, it has to be shown that $[\Delta_{\text{norm}}, \Delta] \neq \mathbf{0}$. Hence, it's sufficient to show that $\exists\, \boldsymbol{v} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$ s.t. $[\Delta_{\text{norm}}, \Delta]\boldsymbol{v} \neq \mathbf{0}$. It is straightforward to see that $\mathbf{1}$ satisfies that condition. Notice that $\Delta\mathbf{1} = \text{diag}(\boldsymbol{A}\mathbf{1})\mathbf{1} - \boldsymbol{A}\mathbf{1} = \mathbf{0}$. Then:

$$[\Delta_{\text{norm}}, \Delta]\mathbf{1} = \Delta_{\text{norm}}\Delta\mathbf{1} - \Delta\Delta_{\text{norm}}\mathbf{1} = -\Delta\Delta_{\text{norm}}\mathbf{1} = \Delta \sum_{i=1}^{N} \left(-1 + \sum_{j \in \mathcal{N}_i} \frac{1}{\sqrt{d_i d_j}} \right) \boldsymbol{e}_i \neq \mathbf{0}$$

since $\Delta\boldsymbol{v} = \mathbf{0} \iff \exists\, c \in \mathbb{R} \mid \boldsymbol{v} = c\mathbf{1}$, and in general, for non-regular topologies, $(-1 + \sum_{j \in \mathcal{N}_i} \frac{1}{\sqrt{d_i d_j}}) \neq (-1 + \sum_{j \in \mathcal{N}_k} \frac{1}{\sqrt{d_k d_j}})$ for $k \in \mathcal{V} \setminus \{i\}$. The four-node graph presented in Section 3 provides an explicit existence proof, as its varying node degrees explicitly prevent the resulting vector from being constant. □

Proof of Proposition 1

Consider a scalar signal $x_i = \frac{c}{\sqrt{d_i}}$, $c \in \mathbb{R}$. Then $\boldsymbol{X}$ can be rewritten as $\boldsymbol{D}^{-1/2}c\mathbf{1}$. By the first axiom of node-similarity measure, μ must be strictly positive in $\boldsymbol{X}$, since x_i is not constant. In particular:

$$\mu(\boldsymbol{X}) = \mu(\boldsymbol{D}^{-1/2}c\mathbf{1}) = \text{tr}(c\mathbf{1}^\top \boldsymbol{D}^{-1/2}\boldsymbol{M}_\mu \boldsymbol{D}^{-1/2}c\mathbf{1}) = \mu'(c\mathbf{1}) > 0.$$

This expression can be interpreted as a different node similarity measure μ', induced by $\boldsymbol{D}^{-1/2}\boldsymbol{M}_\mu\boldsymbol{D}^{-1/2}$, applied to $c\mathbf{1}$ (a constant signal). If $\boldsymbol{D}^{-1/2}\boldsymbol{M}_\mu\boldsymbol{D}^{-1/2}$ induced a valid node-similarity measure, it should satisfy Axiom 1. This would imply that for the null signal ($x_i = 0$), $\text{tr}(\mathbf{0}^\top \boldsymbol{D}^{-1/2}\boldsymbol{M}_\mu\boldsymbol{D}^{-1/2}\mathbf{0}) = 0$. However, the previous derivation shows it assumes positive values for constant signals $c\mathbf{1}$ ($c \neq 0$). Thus, it violates Axiom 1. □

Notice that the proof relies on the fact that $\boldsymbol{M}_\mu$ and $\boldsymbol{M}'_\mu = \boldsymbol{D}^{-1/2}\boldsymbol{M}_\mu\boldsymbol{D}^{-1/2}$ do not commute on non-regular graphs, leading to distinct kernels.

C Formal Proof of End-to-End Topological Collapse

In this section, we derive the functional transition of node representations from $\ker(\Delta_{norm})$ to $\ker(\Delta)$ following the classification step.

Let $H^{(N)} \in \mathbb{R}^{|\mathcal{V}| \times m}$ be the latent features of an N-layer Simple GCN for multi-class single-label classification task with C classes, using the normalized adjacency matrix $\boldsymbol{A}_{\text{norm}}$. As $N \to \infty$, the representations converge to the dominant eigenspace of $\boldsymbol{A}_{\text{norm}}$, corresponding to $\ker(\Delta_{\text{norm}})$. Considering the node representations as column vectors, the representation of node i, denoted as $\mathbf{h}_i^{(N)} \in \mathbb{R}^m$, converges to:

$$\lim_{N \to \infty} \mathbf{h}_i^{(N)} = \sqrt{d_i}\mathbf{c} \tag{2}$$

where d_i is the degree (or augmented degree) and $\mathbf{c} \in \mathbb{R}^m$ is a constant vector shared across all nodes. Applying the final linear transformation $W \in \mathbb{R}^{m \times C}$ yields the logit vector $\mathbf{z}_i \in \mathbb{R}^C$:

$$\mathbf{z}_i^\top = (\mathbf{h}_i^{(N)})^\top W = (\sqrt{d_i}\mathbf{c})^\top W = \sqrt{d_i}\mathbf{v}^\top \tag{3}$$

where $\mathbf{v} = W^\top \mathbf{c} \in \mathbb{R}^C$ is a constant class-space vector. Let $c_i^* \in \{1, \dots, C\}$ be the predicted class index for node i, obtained via the `argmax` operator over the scalar components of the logit vector[5]:

$$c_i^* = \text{argmax}_j([\mathbf{z}_i]_j) = \text{argmax}_j(\sqrt{d_i} v_j) \tag{4}$$

Since $\sqrt{d_i} > 0$ is a node-dependent scalar that does not affect the relative ordering of the components of $\mathbf{v}$, it can be factored out:

$$c_i^* = \text{argmax}_j(v_j) = c^* \tag{5}$$

The predicted class index c^* thus becomes entirely independent of the node index i. Consequently, the discrete one-hot prediction vector for node i, denoted as $\hat{\mathbf{y}}_i \in \{0,1\}^C$, perfectly aligns with the standard basis vector for class c^*:

$$\hat{\mathbf{y}}_i = \mathbf{e}_{c^*} \tag{6}$$

The resulting discrete prediction matrix $\hat{Y} \in \{0,1\}^{|\mathcal{V}| \times C}$, whose i-th row is defined as $\hat{\mathbf{y}}_i^\top$, has identical rows and can be analytically factored as an outer product:

$$\hat{Y} = \mathbf{1}\mathbf{e}_{c^*}^\top \tag{7}$$

where $\mathbf{1} \in \mathbb{R}^{|\mathcal{V}|}$ is the all-ones vector. Evaluating the standard Dirichlet energy $\mathcal{E}_\Delta$ induced by the combinatorial Laplacian Δ:

$$\mathcal{E}_\Delta(\hat{Y}) = \text{Tr}\left(\hat{Y}^\top \Delta \hat{Y}\right) = \text{Tr}\left(\mathbf{e}_{c^*}\mathbf{1}^\top \Delta \mathbf{1}\mathbf{e}_{c^*}^\top\right) \tag{8}$$

Since $\mathbf{1} \in \ker(\Delta)$, we have $\Delta\mathbf{1} = \mathbf{0}$, which strictly implies $\mathcal{E}_\Delta(\hat{Y}) = 0$. This rigorously proves that the non-differentiable classification head forces the system to jump from the normalized kernel to the standard combinatorial kernel, thereby completing the definition of end-to-end topological over-smoothing.

D Additional experiments

LCC of the Cora Dataset

In the following experiment, we apply the same spectral filter applied in Section 5.1 to some graphs of Enzymes, to a different graph: the Largest Connected Component (LCC) of the Cora dataset, composed of 2485 nodes. We expect to observe the same trend of $\mathcal{E}_{\Delta_{\text{norm}}}$ as before, since the GCN implements a low-pass filter w.r.t. Δ_{norm} regardless of the specific input signal.

To empirically isolate the pure smoothing properties of the GCN propagation operator, we apply the diffusion process directly to the raw input features $\boldsymbol{X}^{(0)} \in \mathbb{R}^{N \times m}$ without introducing any learnable parameter, initial feature transformation, or non-linearity. The state at each layer $k = 1, \dots, K$ is thus strictly governed by the repeated application of the normalized adjacency matrix:

$$\boldsymbol{X}^{(k)} = \boldsymbol{A}_{\text{norm}} \boldsymbol{X}^{(k-1)} = \boldsymbol{A}_{\text{norm}}^k \boldsymbol{X}^{(0)}.$$

At each propagation step $k \in [0, K]$, we track the evolution of the Frobenius norm of the embeddings, $\|\boldsymbol{X}^{(k)}\|_F$; the standard Dirichlet energy, $\mathcal{E}_\Delta(\boldsymbol{X}^{(k)})$; and the normalized Dirichlet energy, $\mathcal{E}_{\Delta_{\text{norm}}}(\boldsymbol{X}^{(k)})$. Results are shown in Figure 3.

Indeterminacy of Spectral Filtering

The goal of this subsection is to empirically verify that a GNN filter defined as a spectral function of Δ_{norm} does not behave as a consistent spectral filter with respect to the frequencies of Δ on non-regular graphs. A GCN layer implementing the operator $\boldsymbol{A}_{\text{norm}} = \boldsymbol{I} - \Delta_{\text{norm}}$, i.e., a polynomial of the normalized Laplacian, is applied to both a regular and a non-regular graph. According to the theoretical results of the previous

[5]Standard classification pipelines typically apply a softmax function to the logits to obtain class probabilities. We omit this step in our derivation because the softmax function is strictly monotonic: it preserves the relative ordering of the logit components and therefore does not alter the discrete output of the `argmax` operator.

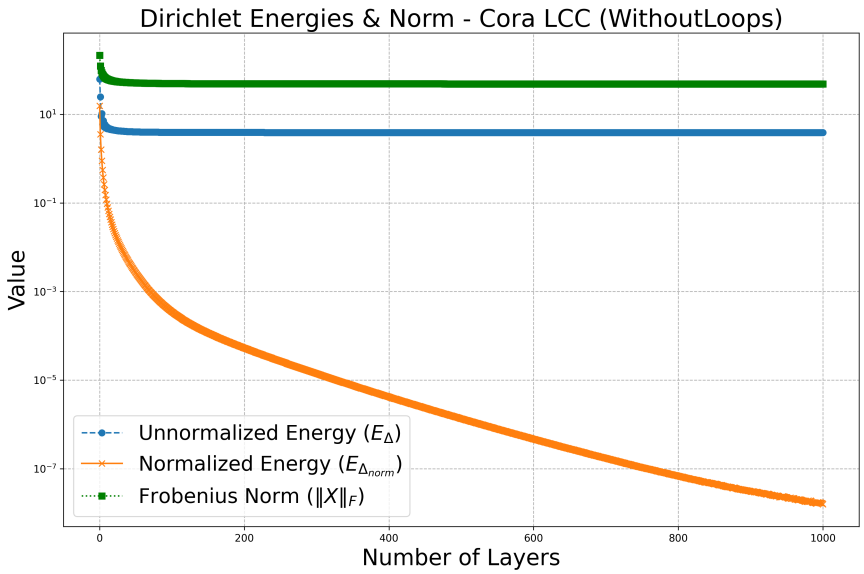


Figure 3: Evolution in logarithmic scale of the standard Dirichlet energy ($\mathcal{E}_\Delta$), normalized Dirichlet energy ($\mathcal{E}_{\Delta_{\text{norm}}}$), and Frobenius norm ($\|\boldsymbol{X}\|_F$) across 1000 propagation steps. The diffusion process $\boldsymbol{X}^{(k)} = \boldsymbol{A}_{\text{norm}}^k \boldsymbol{X}^{(0)}$ is applied directly to the raw input features of the Cora LCC (without self-loops), strictly isolating the pure topological dynamics. The exponential decay of $\mathcal{E}_{\Delta_{\text{norm}}}$ confirms the asymptotic convergence of the signal to the kernel of Δ_{norm}. Simultaneously, the Frobenius norm stabilizes at a strictly positive value, definitively ruling out over-shrinking. The non-zero plateau of $\mathcal{E}_\Delta$ empirically demonstrates the spectral mismatch: the standard energy is structurally blind to this normalized topological collapse.

sections, for a regular graph of degree d, we have $\boldsymbol{D} = d\boldsymbol{I}$. Consequently, the ratio between the two Dirichlet energies (for any $\boldsymbol{X}$ where it is well-defined) is:

$$\frac{\mathcal{E}_\Delta(\boldsymbol{X})}{\mathcal{E}_{\Delta_{\text{norm}}}(\boldsymbol{X})} = \frac{\operatorname{tr}(\boldsymbol{X}^\top(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{X})}{\operatorname{tr}(\boldsymbol{X}^\top d^{-1/2}(\boldsymbol{D}-\boldsymbol{A})d^{-1/2}\boldsymbol{X})} = d.$$

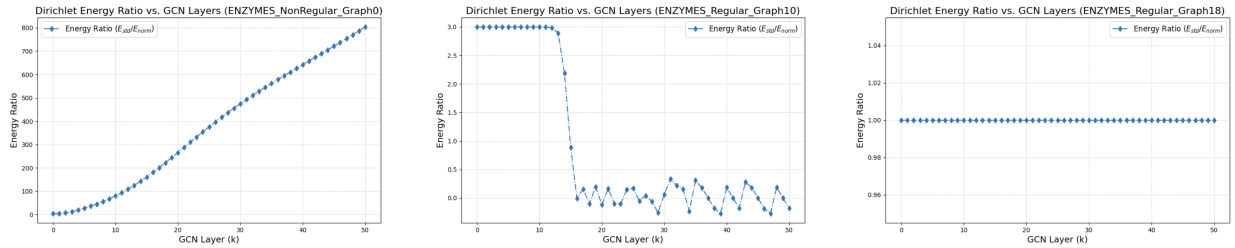
This ratio is a constant independent of $\boldsymbol{X}$, and thus remains constant across all layers of a GCN. Hence, the expected plot of the ratio across layers is a horizontal line at value d.
Conversely, for a non-regular graph, the ratio:

$$\frac{\mathcal{E}_\Delta(\boldsymbol{X})}{\mathcal{E}_{\Delta_{\text{norm}}}(\boldsymbol{X})} = \frac{\operatorname{tr}(\boldsymbol{X}^\top(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{X})}{\operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{D}^{-1/2}(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{D}^{-1/2}\boldsymbol{X})}$$

depends on $\boldsymbol{X}$ and evolves through the layers. This behavior is numerically illustrated in Figure 4: for the regular graph (where $d = 3$) in Figure 4b, the ratio remains perfectly constant for the initial iterations. The instability observed after layer $k \approx 13$ is purely a numerical artifact: as verified in Figure 1b, beyond this point both energy values decay to magnitudes close to machine precision limits. Consequently, their ratio becomes ill-conditioned and dominated by floating-point errors.

E Ablation Studies on Simplified Graph Convolution Dynamics

In Section 4.3, we illustrated the trajectory of the Simplified Graph Convolution (SGC) learning framework under standard, idealized conditions ($K = 1000$, self-loops included, and a bias-free linear layer). In this appendix, we complement those experiments with systematic ablations designed to expose the inherent limitations of purely structural proxy metrics when standard architectural or topological components are introduced. Rather than demonstrating a universal robustness, these experiments reveal that the diagnostic validity of both Dirichlet energies is easily compromised: the simple addition of a trainable bias term introduces optimization artifacts that prevent these metrics from vanishing despite the ongoing representational



(a) 37 nodes, 168 edges, non regular. (b) 4 nodes, 12 edges, regular. (c) 2 nodes, 1 edge, regular.

Figure 4: Evolution of the ratio $\mathcal{E}_{\Delta}(\boldsymbol{X})/\mathcal{E}_{\Delta_{\text{norm}}}(\boldsymbol{X})$ across the layers of a GCN implementing $\boldsymbol{X}^{(k+1)} = \boldsymbol{A}_{\text{norm}}\boldsymbol{X}^{(k)}$ for two graphs from the ENZYMES dataset.

collapse. Ultimately, these ablations underscore why structural scalar metrics fail to provide a reliable, longitudinal indicator of over-smoothing once the idealized propagation rule is altered.

E.1 Influence of the Bias Term

E.1.1 Theoretical Analysis of the Bias Term and Metric Blindness

Let the output logits at layer K be defined as:

$$\mathbf{Z}^{(K)} = \mathbf{A}_{\text{norm}}^{K}\boldsymbol{X}\mathbf{W} + \mathbf{1}_N\mathbf{b}^\top \tag{9}$$

where $\mathbf{W} \in \mathbb{R}^{m\times C}$ is the weight matrix, and $\mathbf{b} \in \mathbb{R}^C$ is the class bias vector.

E.1.2 Asymptotic Decomposition and Metric Blindness

As demonstrated in previous sections, for $K \to \infty$, the diffused features collapse into the kernel of the normalized Laplacian. The asymptotic diffused feature matrix can be written as $\mathbf{X}^{(\infty)} = \mathbf{D}^{1/2}\mathbf{1}_N\mathbf{c}^\top$, for some vector $\mathbf{c}^\top \in \mathbb{R}^{1\times m}$ that depends on the initial features $\mathbf{X}^{(0)}$. Consequently, the asymptotic logits decompose into two distinct components:

$$\mathbf{Z}^{(\infty)} = \mathbf{X}^{(\infty)}\mathbf{W} + \mathbf{1}_N\mathbf{b}^\top = \mathbf{D}^{1/2}\mathbf{1}_N(\mathbf{c}^\top\mathbf{W}) + \mathbf{1}_N\mathbf{b}^\top$$

If we evaluate the standard Dirichlet energy on these asymptotic logits, $\mathcal{E}_{\Delta}(\mathbf{Z}^{(\infty)})$, the unnormalized Laplacian Δ perfectly annihilates the constant bias term $\mathbf{1}_N\mathbf{b}^\top$. In the case of over-shrinking, the optimization dynamics drive the weight matrix to zero ($\mathbf{W} \to \mathbf{0}$); thus, the first term vanishes and the standard energy drops to zero. However, during genuine over-smoothing, $\mathbf{W} \neq \mathbf{0}$. Since $\mathbf{D}^{1/2}\mathbf{1}_N \notin \text{Ker}(\Delta)$ for non-regular graphs, the standard energy will record a strictly positive plateau, rendering it structurally blind to the underlying topological collapse.

Conversely, evaluating the normalized Dirichlet energy $\mathcal{E}_{\Delta_{\text{norm}}}(\mathbf{Z}^{(\infty)})$ yields the opposite symmetric collision. The operator Δ_{norm} annihilates the topological term $\mathbf{X}^{(\infty)}\mathbf{W}$, but it clashes with the constant bias vector (since $\mathbf{1}_N \notin \text{Ker}(\Delta_{\text{norm}})$), resulting in:

$$\mathcal{E}_{\Delta_{\text{norm}}}(\mathbf{Z}^{(\infty)}) = \left(\frac{1}{N}\mathbf{1}_N^\top\Delta_{\text{norm}}\mathbf{1}_N\right)\|\mathbf{b}\|_2^2 > 0 \tag{10}$$

This demonstrates that, in a standard learning regime where the bias is considered, the normalized energy will not converge to zero but will perfectly track the magnitude of the bias parameterized by the optimizer.

E.1.3 Empirical Validation via Bias Ablation

To empirically prove this asymptotic decomposition, we track the evolution of the SGC across layers on the Cora LCC dataset under two distinct optimization regimes.

Standard Regime (Bias Active): $\mathcal{E}_{\Delta_{\text{norm}}}$ remains strictly positive, mirroring the non-zero magnitude of $\|\mathbf{b}\|_2$, exactly as predicted by our algebraic decomposition (Figure 5).

Strict Topological Collapse (Bias Suppressed): Without the bias, the logits are forced into a state of pure topological collapse ($\mathbf{Z}^{(K)} \to \mathbf{D}^{1/2}\mathbf{1}_N(\mathbf{c}^\top\mathbf{W})$). As shown in Figure 2:

- The predictive performance crashes asymptotically to the random baseline.
- The normalized energy $\mathcal{E}_{\Delta_{\text{norm}}}$ plunges to zero, correctly diagnosing the collapse into the spatial kernel.
- The standard Dirichlet energy $\mathcal{E}_\Delta$ and the Frobenius norms ($\|\mathbf{W}\|_F$, $\|\boldsymbol{X}\|_F$) remain solidly bounded above zero, definitively proving their blindness to spectral over-smoothing.

These results provide conclusive evidence that assessing the smoothing state of a trained GNN requires separating the optimization artifacts from the underlying spatial diffusion. Standard energy metrics applied to logits are fundamentally ill-suited for this diagnostic, since the mere inclusion of a bias term can dictate the way in which these proxies measure the collapse.

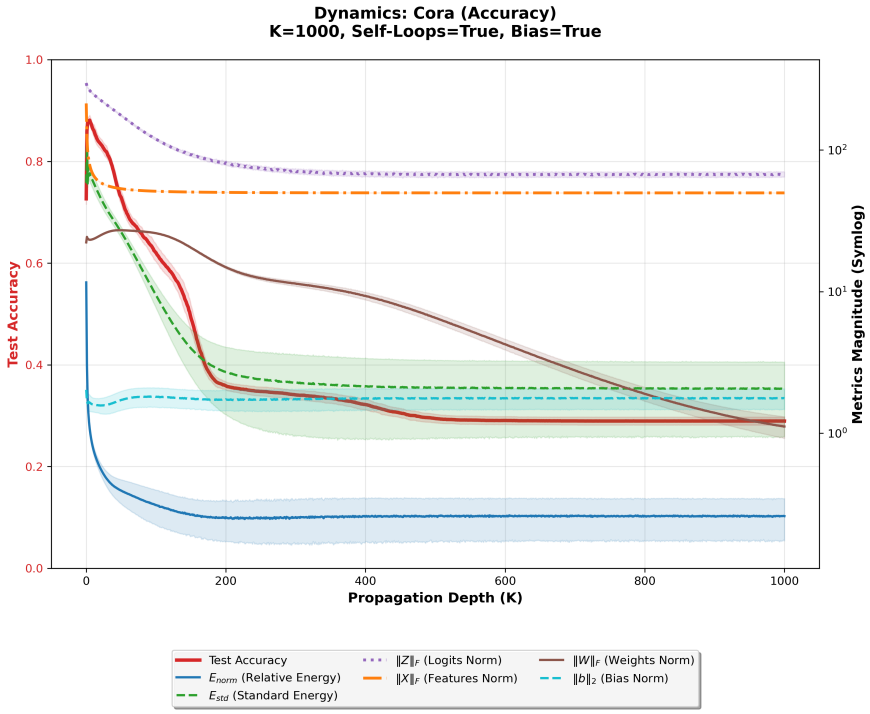


Figure 5: Introduction of the trainable global bias offset on the Cora LCC while maintaining $K = 1000$ bounds and standardized self-loops topological configurations.

E.2 Topological Propagation Without the Renormalization Trick ($SL =$ False)

Finally, we address the absence of graph self-loops ($SL =$ False). In the foundational formulation of the GCN Kipf & Welling (2017), self-loops are added to the adjacency matrix prior to symmetric normalization.

As evidenced in Figure 6, propagating features strictly via the normalized adjacency matrix without self-loops ($\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$) accelerates the onset of over-smoothing, forcing the collapse to occur at a lower depth K compared to the augmented topology. As demonstrated by Wu et al. (2019), the inclusion of self-loops algebraically shrinks the spectrum of the normalized Laplacian toward zero and hence the spectrum of the normalized adjacency matrix toward 1. By narrowing the spectral gap, the self-loops effectively act as a

structural regularizer that mitigates the damping of the signal, thereby delaying the topological collapse. Consequently, for the Cora dataset, the average normalized Dirichlet energy $\mathcal{E}_{\Delta_{\text{norm}}}$ reaches 10^{-5} after fewer than 500 convolutions without self-loops, whereas the "lazier" dynamics induced by the renormalization trick prolong this convergence after 600 convolutions (as previously shown in Figure 2).

However, while self-loops mitigate the rate of over-smoothing, this does not universally guarantee an improvement in downstream predictive performance. Beyond merely altering the eigenvalues, augmenting the graph fundamentally modifies the rank of the propagation operator and reshapes the eigenspace of the Laplacian's kernel. Specifically, as analyzed in Bison & Sperduti (2026), the rank variation induced by self-loops, $\text{rank}(\tilde{\mathbf{A}}) - \text{rank}(\mathbf{A})$, corresponds exactly to $\dim(\text{Ker}(\mathbf{A})) - \dim(E_{-1}(\mathbf{A}))$, where $E_{-1}(\mathbf{A})$ is the eigenspace associated with the eigenvalue -1. Because symmetric normalization via full-rank degree matrices preserves matrix rank, this structural divergence persists directly in the normalized operators. Consequently, the asymptotic eigenspace to which the augmented network relaxes is geometrically distinct from the original kernel.

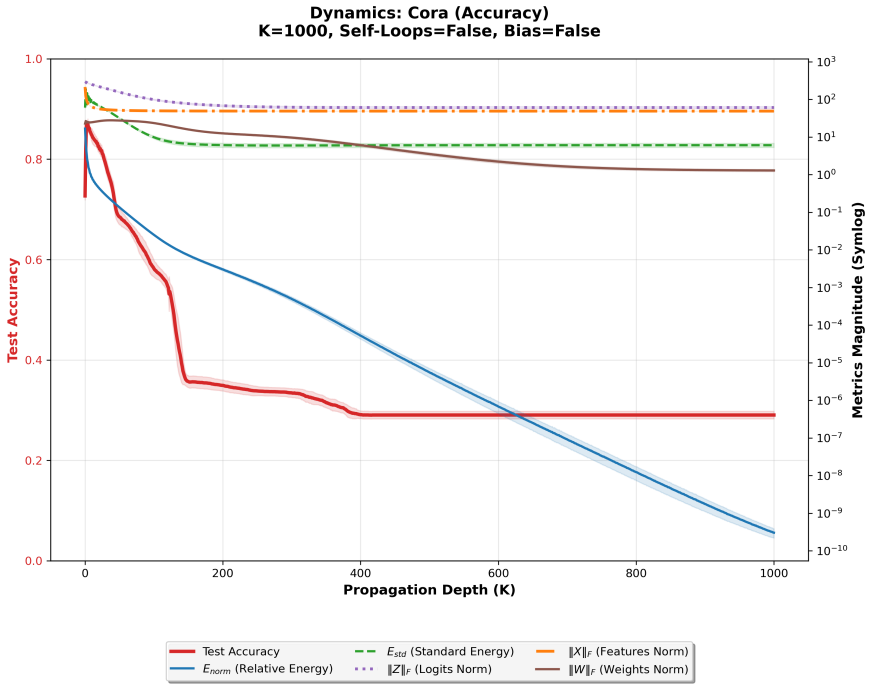


Figure 6: Training metrics reflecting the removal of self loops on the Cora LCC.