Fixed-Point Probing for GNN Depth Diagnostics: A Geometry-Consistent Protocol with a Patent-Citation Case Study

Anonymous authors
Paper under double-blind review

Abstract

Deep graph neural networks (GNNs) often degrade with depth, but endpoint metrics and any single probe do not reveal whether late-depth behavior reflects benign stabilization, classical oversmoothing, or a geometry-specific failure mode. Here, we read depth as a sequence of learned representations, not just as a model-size hyperparameter. We introduce fixed-point probing, a post-training protocol that keeps the probe subset fixed and the measurements geometry-consistent, so familiar signals can be read together across depths and embedding geometries. Applied to depth sweeps up to 32 layers on a patent-citation stress test, the protocol reveals geometry-dependent late-depth regimes. Euclidean models exhibit gradual class-structure degradation consistent with classical oversmoothing, while hyperbolic models enter a late-depth regime in which representation drift and graph-local roughness increase as embeddings approach the boundary. A tuned hyperbolic control matches Euclidean performance at shallow depth yet exhibits the same qualitative late-depth pattern, indicating that this effect is not explained by a trivially weak baseline. Taken together, the results point to a boundary-coupled late-depth regime in hyperbolic GNNs that is hard to isolate from endpoint metrics or from any single probe alone, but becomes visible when the probes are read jointly under a shared protocol. The protocol is the main contribution; the patent citation graph is used as a stress-test case study, not as evidence for dataset-universal claims.

1 Introduction

Deep graph neural networks (GNNs) often degrade with depth, but endpoint metrics alone do not reveal whether late-depth behavior reflects benign stabilization, classical oversmoothing, or a geometry-specific failure mode. After training, a depth-L message-passing network induces a sequence of layerwise representations $\{h^{(0)}, \ldots, h^{(L)}\}$. We study this sequence directly and treat depth as an *observed representation trajectory* rather than merely as a model-size hyperparameter.

The harder problem is not finding probes, but making their readings comparable across depths, models, and embedding geometries. Existing analyses typically rely on single-signal summaries such as interlayer drift, graph smoothness, or class separability. Each signal can be informative, but each captures only part of the depth dynamics. Moreover, they are often evaluated on different node subsets, edge samples, or distance conventions, making them difficult to compare across depths, models, and embedding geometries. Without a fixed and geometry-consistent evaluation scheme, qualitatively different regimes can appear deceptively similar: benign stabilization may look like collapse, while geometry-specific amplification may be mistaken for ordinary oversmoothing.

Instead of proposing another scalar depth metric, we introduce *fixed-point probing*, a fixed-subset, geometry-consistent post-training protocol for analyzing depthwise representation dynamics. The protocol reuses the

A large language model was used for limited writing assistance (English copyediting and phrasing suggestions) during manuscript preparation. The scientific ideas, technical claims, experimental design, analyses, and final verification were performed by the authors.

same probe node set and evaluation edges across depths, models, and ablations, and evaluates familiar signals in metrics consistent with the underlying representation geometry. In practice, it jointly tracks interlayer drift, metric-aware Dirichlet energy, label separability, and, for hyperbolic representations, boundary-pressure statistics. Throughout, *fixed point* is used operationally to denote a near-stationary plateau in the observed trajectory, not an exact fixed point of a dynamical system. The protocol is therefore *plateau-aware*: it is designed to detect both near-stationary behavior and departures from it.

Protocol in one sentence. Fixed-point probing fixes a probe-node set $\mathcal{S}$ and an evaluation-edge set $\mathcal{E}_{\text{eval}}$ once, computes geometry-consistent drift, energy, and separability (and boundary summaries in hyperbolic space) on the same observed representations across depth, and assigns regimes only from the joint coupling of these signals.

Geometry makes this distinction particularly important. Hyperbolic embeddings in the Poincaré ball are attractive for graphs with hierarchical structure (Nickel & Kiela, 2017; Chami et al., 2019; Bachmann et al., 2020), but the local metric expands near the boundary (Nickel & Kiela, 2017; Ganea et al., 2018). Consequently, small coordinate updates can correspond to large geodesic displacements. Apparent coordinate-level stability—or even smoothly varying endpoint performance—can therefore coexist with rising geodesic drift and graph-local roughness. If probes are not interpreted in a geometry-consistent way, such effects can be conflated with benign stabilization or with conventional oversmoothing. For this reason, our protocol explicitly distinguishes coordinate-level updates from geodesic drift and monitors boundary proximity in the hyperbolic case.

Our aim is diagnostic rather than remedial. A substantial literature on deeper GNNs develops architectural and training mechanisms to counter oversmoothing and stabilize optimization, including residual and identity mappings, multiscale layer aggregation, diffusion-style propagation, stochastic edge dropping, normalization-based corrections, and very-deep/reversible designs (Chen et al., 2020b; Xu et al., 2018; Gasteiger et al., 2019; Rong et al., 2020; Zhao & Akoglu, 2020; Li et al., 2021). That mitigation-oriented line of work is valuable, but it is typically assessed through endpoint performance, smoothness proxies, or gains from the proposed remedy. Here we ask a different question: once a model has been trained, what *failure shape* is visible in its internal depth trajectory, and does that shape depend on the embedding geometry? Framed this way, endpoint metrics alone are not sufficient, and no single familiar probe is reliably decisive.

We present a protocol contribution evaluated through a bounded case study. The protocol is the primary contribution. The patent citation graph serves as the main stress-test environment because it is large, many-class, temporally split, and plausibly compatible with a hyperbolic representation hypothesis. GCN results on the same patent subgraph and public-benchmark runs reported in the appendix are included as scope checks rather than as the basis for dataset-universal claims. Appendix E.1 makes this boundary explicit by listing what the patent stress test, the same-data GCN backbone check, the public Cora/Texas checks, and the OGBN-Arxiv reference checks support—and what they do not support.

Applying fixed-point probing to depth sweeps up to 32 layers on the patent graph (479k nodes; 316 classes), we observe geometry-dependent late-depth regimes. Euclidean models are best characterized by gradual degradation of class structure consistent with classical oversmoothing (Oono & Suzuki, 2020; Rusch et al., 2023). In the evaluated hyperbolic models, by contrast, a late-depth regime emerges in which geodesic drift and graph-local roughness increase as representations approach the boundary. A tuned hyperbolic control (HypT) matches Euclidean shallow-depth performance yet exhibits the same qualitative late-depth pattern, reducing the plausibility of a trivially weak-baseline explanation. A radius-penalty perturbation then provides diagnostic evidence consistent with a boundary contribution to the observed amplification in the evaluated setting; because its response is timing-sensitive and nonmonotonic, we do not use it as evidence for a clean radius-control law. We do not claim a dataset-universal law of hyperbolic depth failure; rather, we show that a fixed-subset, geometry-consistent joint view can reveal a boundary-coupled regime that endpoint metrics or any individual probe do not isolate on their own.

Overview. The main-text evidence is organized in three layers. Figures 1–3 provide the core patent-graph narrative: layerwise dynamics, the boundary-coupling view, and an intervention test. Figures 4 and 7 provide secondary summaries relating probe behavior to class structure and endpoint performance. Figure 8 then

brings a Euclidean GCN backbone check on the same patent subgraph into the main text to show that the protocol is not tied to a single GraphSAGE configuration, while public-transfer checks, curvature sweeps, and detailed intervention audits remain in the appendix.

Contributions.

- **Protocol contribution.** We introduce fixed-point probing: a post-training protocol that fixes a probe-node set $\mathcal{S}$, a fixed evaluation-edge set $\mathcal{E}_{\text{eval}}$, geometry-consistent measurements on observed representations, and a joint interpretation rule across drift, energy, separability, and boundary pressure.
- **Case-study finding.** On a large patent citation graph, the protocol reveals a late-depth boundary-coupled regime in the evaluated hyperbolic models that is not isolated by endpoint metrics or by any single probe in isolation.
- **Controlled analysis.** Tuned baselines, boundary-focused interventions, and robustness checks help bound this interpretation and reduce alternative explanations, without claiming dataset-universal thresholds.

Fixed-point probing does not modify optimization and is not presented as a training-time remedy. Its role is diagnostic: the same probe nodes and evaluation edges are reused across depths, geometries, and ablations, hyperbolic signals are interpreted with geometry-consistent metrics and explicit boundary statistics, and regime labels are assigned only from a joint reading of the probe family.

Fixed-point probing: protocol at a glance. For a trained depth-L model, the protocol (i) fixes a probe node set $\mathcal{S}$ and an evaluation edge set $\mathcal{E}_{\text{eval}}$ once and reuses them across depths, geometries, and ablations; (ii) converts layer activations into observed representations $\{\mathbf{z}^{(\ell)}\}_{\ell=0}^{L}$ in a geometry-consistent domain; (iii) computes the same probe family at each layer—interlayer drift, metric-aware Dirichlet energy, label separability, and hyperbolic boundary-pressure summaries when applicable; and (iv) interprets these signals jointly, so that drift, roughness, class structure, and boundary proximity define a single diagnostic signature rather than isolated scalar checks.

2 Related Work

Depth pathologies in deep GNNs. A large literature studies why message-passing GNNs degrade with depth, most prominently under the notion of oversmoothing, where repeated propagation drives node representations toward low-frequency or weakly distinguishable states (Li et al., 2018; Oono & Suzuki, 2020; Rusch et al., 2023; Wu et al., 2023). Prior analyses characterize this effect through spectral viewpoints, contraction or mixing behavior of graph operators, and asymptotic limits of message passing. A related but distinct line of work studies oversquashing and long-range information bottlenecks in graph propagation (Alon & Yahav, 2021; Topping et al., 2021). In parallel, a broad range of architectural and training modifications—including residual or identity mappings, normalization schemes, multiscale aggregation, diffusion-style propagation, stochastic edge dropping, and very-deep/reversible constructions—have been proposed to stabilize deeper GNNs (Chen et al., 2020b; Xu et al., 2018; Gasteiger et al., 2019; Rong et al., 2020; Zhao & Akoglu, 2020; Li et al., 2021).

These studies provide important theoretical and empirical insights, but they are commonly assessed through endpoint performance, asymptotic behavior, or a single diagnostic viewed in isolation. Our focus is complementary: we study how internal representations evolve across intermediate layers after training.

Geometric representations and hyperbolic GNNs. To better capture hierarchical and non-Euclidean structure, graph representation learning has increasingly explored curved latent spaces, especially hyperbolic geometry (Nickel & Kiela, 2017; Ganea et al., 2018; Chami et al., 2019; Bachmann et al., 2020; Yang et al., 2022). Hyperbolic embeddings can represent tree-like or hierarchical relations efficiently, and hyperbolic GNNs

extend message passing to such spaces through manifold-aware operations, tangent-space approximations, or curvature-aware transformations. Recent work has also emphasized that conclusions in hyperbolic graph learning can be sensitive to baseline parity and evaluation protocol choices (Katsman & Gilbert, 2025), and that geometry-task alignment may matter separately from graph hyperbolicity alone (Naddeo et al., 2026).

Much of this literature evaluates geometric models primarily through downstream performance, distortion, or final-layer embedding quality. Comparisons between Euclidean and hyperbolic models are therefore often reported at the level of endpoint metrics, leaving open the question of whether their internal representation dynamics differ qualitatively with depth.

Post-hoc representation diagnostics. A separate line of work analyzes learned graph representations using post-hoc diagnostics such as interlayer drift, smoothness or Dirichlet-style quantities, class separability, and auxiliary probing models (Chen et al., 2020a; Zhou et al., 2021; Rusch et al., 2023; Guan & Shi, 2025). These signals are often informative, but they are usually applied one at a time and under paper-specific sampling choices, node subsets, or distance conventions. Recent work has also questioned whether Dirichlet-like measures alone reliably track oversmoothing-related performance loss across realistic depth ranges, motivating alternative rank-based views in some settings (Zhang et al., 2026). As a result, the resulting measurements are often difficult to compare directly across depths, models, or embedding spaces.

Our work does not introduce new primitives for each of these views. Instead, it coordinates familiar diagnostics under a fixed probe subset, shared evaluation edges, and geometry-consistent metrics so that they become jointly reproducible and jointly interpretable.

Position of this work. Our work is closest in spirit to post-hoc depth analysis, but differs in emphasis. We do not propose a new GNN architecture, a new hyperbolic layer, or a new scalar oversmoothing index. Instead, we present a reproducible post-training protocol that treats depth as an observed representation trajectory and analyzes it using a coordinated set of geometry-consistent probes.

We evaluate this protocol through a bounded patent-citation case study, in which the joint view reveals a late-depth boundary-coupled regime in the evaluated hyperbolic models. Euclidean backbone checks on the same patent subgraph and public-benchmark runs in the appendix bound the scope of this interpretation rather than support dataset-universal claims.

3 Background and problem setting

3.1 A representation-centric view of depth

Oversmoothing is commonly described as a depth-induced phenomenon in which node representations become increasingly indistinguishable. Many studies diagnose this through downstream accuracy loss, implicitly using performance decay as a proxy for representational failure. However, accuracy degradation can be delayed or even absent depending on the task and supervision, whereas representational collapse is a property of the learned depth dynamics themselves (Li et al., 2018; Oono & Suzuki, 2020; Wu et al., 2023). We therefore adopt a representation-centric view: after training, a depth-L message-passing network induces a discrete sequence of layerwise representations, and depth corresponds to the composition of the learned update maps. In general, these maps are layer-specific, so the resulting dynamics are non-autonomous.

Oversmoothing is distinct from *oversquashing*, in which information from distant nodes is compressed through graph bottlenecks (Alon & Yahav, 2021; Topping et al., 2021). Our focus is on depthwise representation dynamics and on geometry-aware diagnostics for oversmoothing-like degradation and boundary-associated metric amplification.

3.2 Why a joint protocol is needed

Graph smoothness measures, such as (metric-aware) Dirichlet energy, quantify neighborhood-level variation and have been widely used to diagnose or mitigate oversmoothing (Chen et al., 2020a; Zhou et al., 2021; Rusch et al., 2023), with recent refinements beyond first-order Dirichlet energy (Guan & Shi, 2025). Drift-based

and separability-based views are likewise informative, but no single signal is uniquely diagnostic. The same low-energy or low-drift pattern can reflect either useful convergence or degenerate homogenization, and in non-Euclidean settings it can coexist with geometry-driven effects that are invisible to a single scalar.

These observations motivate a joint protocol that explicitly tracks the co-evolution of interlayer drift, graph-local roughness, class structure, and, in hyperbolic space, boundary pressure across depth. Fixing probe nodes and evaluation edges is essential here: without fixed subsets and geometry-consistent metrics, depthwise trends are confounded by resampling variation or by incomparable distance conventions across models and embedding spaces.

3.3 Hyperbolic geometry and boundary amplification

Hyperbolic GNNs embed representations in a Poincaré ball, where distances can grow rapidly near the boundary. This geometry is well-suited to hierarchical structure (Nickel & Kiela, 2017; Chami et al., 2019; Bachmann et al., 2020; Yang et al., 2022), but it changes how depth diagnostics should be interpreted. In particular, low coordinate drift or low local variation need not imply harmful collapse: hyperbolic representations can remain stable while preserving angular, class-relevant structure. At larger depths, however, representations may saturate toward the boundary ($\|x\| \to 1$), where small directional updates can induce large geodesic displacements, leading to boundary-associated metric amplification (Ganea et al., 2018). Concretely, for a sufficiently small coordinate update Δx in the Poincaré ball,

$$d_{\mathrm{Hyp}}(x, x + \Delta x) \approx \lambda(x)\|\Delta x\|_2, \qquad \lambda(x) = \frac{2}{1 - \|x\|_2^2},$$

so identical coordinate-scale updates become more visible in measured geodesic drift near the boundary, and squared-distance quantities such as metric-aware Dirichlet energy are locally amplified by approximately $\lambda(x)^2$. This is why explicit boundary monitoring is incorporated into our probing framework.

4 Method: Fixed-point (plateau) probing on fixed subsets

Throughout this section, *fixed-point(-like)* refers to near-stationary plateaus in the observed layerwise trajectory in the operational sense formalized in Appendix D.1. We first formalize the representation geometry and the observed representations, then introduce three complementary probes—drift, Dirichlet energy, and separability—followed by boundary monitoring for hyperbolic models.

Intuitively, *drift* measures how far representations move along the layerwise trajectory, *Dirichlet energy* measures graph-local roughness, *separability* serves as a proxy for retained task-relevant class structure, and boundary pressure tracks hyperbolic-specific metric sensitivity. We interpret these signals jointly because they form a practical complementary probe set. We do not claim universal completeness for this set; rather, in the settings studied here it is sufficient to separate benign stabilization, oversmoothing-like degradation, and boundary-coupled amplification without relying on endpoint accuracy alone.

Let $G = (V, \mathcal{E})$ be a graph with node features $\{x_i\}_{i \in V}$. A trained message-passing GNN of depth L produces layerwise *internal activations* $\{\tilde{\mathbf{h}}_i^{(\ell)}\}_{i \in V, \ell=0}^{L}$, where $\tilde{\mathbf{h}}_i^{(0)}$ is the input projection of x_i. We write the depth-indexed layer update as

$$\tilde{H}^{(\ell+1)} = F_{\theta^{(\ell)}}\left(\tilde{H}^{(\ell)}; G\right), \quad \ell = 0, \ldots, L-1, \tag{1}$$

where $\tilde{H}^{(\ell)} = \{\tilde{\mathbf{h}}_i^{(\ell)}\}_{i \in V}$ and $F_{\theta^{(\ell)}}$ denotes the trained message-passing layer at depth ℓ. For notational simplicity, we occasionally write F_θ when layer-specific parameters are not essential.

Representation geometry. We consider two representation manifolds: (i) Euclidean space $\mathbb{R}^d$ with $d_{\mathrm{Euc}}(u, v) = \|u - v\|_2$; and (ii) a d-dimensional Poincaré ball with curvature parameter $c > 0$,

$$\mathbb{B}_c^d = \{u \in \mathbb{R}^d : c\,\|u\|_2^2 < 1\},$$

equipped with the geodesic distance $d_{\mathbb{B}_c}(u, v)$ (Appendix D.4). Throughout this paper, we write $d_{\mathrm{Hyp}}(u, v) := d_{\mathbb{B}_c}(u, v)$. In our experiments, HYPA uses $c = 1.0$ and HYPT uses $c = 3.0$ (see Table A.3). For readability, some expressions are written in unit-ball form ($c = 1$); all probe computations use the model-specific curvature, and the full c-dependent formulas are provided in Appendix D.4. All probes below are defined for a generic metric $d(\cdot, \cdot)$.

Observed representations for probing. Intermediate activations may be parameterized in different coordinate systems, depending on the model. We therefore distinguish between the *internal activations* $\tilde{\mathbf{h}}_i^{(\ell)}$ and the *observed representations* $\mathbf{z}_i^{(\ell)}$ used by the probes. For Euclidean models, we set $\mathbf{z}_i^{(\ell)} := \tilde{\mathbf{h}}_i^{(\ell)} \in \mathbb{R}^d$. For hyperbolic models, the internal activations are mapped to the Poincaré ball via the exponential map at the origin:

$$\mathbf{z}_i^{(\ell)} := \exp_0\left(\tilde{\mathbf{h}}_i^{(\ell)}\right) \in \mathbb{B}_c^d, \tag{2}$$

where $\exp_0(\cdot)$ is defined in Appendix D.4. All subsequent probes are evaluated on the observed representations $\{\mathbf{z}_i^{(\ell)}\}$ under the metric induced by the target geometry.

This probing setup follows a protocol-centric design. Observed representations provide a geometry-consistent coordinate domain, while fixed probe nodes and evaluation edges ensure that diagnostic signals (drift, energy, separability, and boundary statistics) are comparable across depths, geometries, and ablation settings.

Fixed probe sets. All diagnostics are computed on fixed subsets to avoid resampling artifacts: a probe node set $\mathcal{S} \subset V$ (for drift, separability, and boundary monitoring), and an evaluation edge set $\mathcal{E}_{\mathrm{eval}} \subset \mathcal{E}$ (for Dirichlet energy). Unless otherwise stated, $\mathcal{S}$ and $\mathcal{E}_{\mathrm{eval}}$ are sampled once from the training-time induced subgraph and reused across depths, geometries, and activation variants (Appendix A.1). This design ensures that depthwise trends reflect representation dynamics rather than subset variation. When a node-level probe is reported on a specific split, we intersect the fixed probe set with the corresponding split mask after subset construction; unless a caption states otherwise, edge-based summaries use the same fixed evaluation-edge set throughout. For reproducibility, we release the corresponding IDs and seeds as part of the diagnostic bundle (Appendix A.1).

Practical budget choice. The main experiments use $|\mathcal{S}| = 50{,}000$ probe nodes and $|\mathcal{E}_{\mathrm{eval}}| = 200{,}000$ evaluation edges as conservative fixed budgets. Appendix E.3 shows that smaller pilot budgets already preserve the qualitative regime ordering in the tested settings. In practice, we recommend increasing $|\mathcal{S}|$ and $|\mathcal{E}_{\mathrm{eval}}|$ until the regime label and the rank/order of depthwise summaries are stable under a few resampling seeds, then fixing and reporting the selected subsets for all comparisons.

4.1 Drift: Inter-layer displacement

For node i and transition $\ell \to \ell + 1$, we define the per-node drift as

$$\delta_i^{(\ell)} = d\left(\mathbf{z}_i^{(\ell+1)}, \mathbf{z}_i^{(\ell)}\right), \qquad \ell = 0, \ldots, L-1. \tag{3}$$

We summarize the drift over the probe set $\mathcal{S}$ using

$$D_{\mathrm{mean}}(\ell) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \delta_i^{(\ell)}, \tag{4}$$

$$D_{\mathrm{p50}}(\ell) = \mathrm{median}_{i \in \mathcal{S}}\, \delta_i^{(\ell)}. \tag{5}$$

Reporting convention. Unless otherwise stated, we report drift using the median summary $D_{\mathrm{p50}}(\ell)$, denoted in figures/tables as "drift (p50)." When a single depth-level drift value is required (e.g., depth-sweep plots), we use the final transition

$$D_{\mathrm{final}}(L) := D_{\mathrm{p50}}(\ell = L - 1). \tag{6}$$

For upper-tail diagnostics reported in the appendix, we analogously use

$$D_{\mathrm{p90}}(\ell) := \mathrm{p90}_{i \in \mathcal{S}}\, \delta_i^{(\ell)}. \tag{7}$$

Across multiple random seeds, reported drift values are summarized as mean±std or explicitly identified per-seed summaries in the corresponding caption.

For hyperbolic models, d in Eq. (3) is instantiated as the Poincaré-ball geodesic distance d_{Hyp}; thus, drift measures the geodesic inter-layer displacement on $\mathbb{B}_c^d$. A small drift over consecutive layers indicates a *near-stationary (fixed-point-like) plateau* in the operational sense adopted in this study (Appendix D.1). Concretely, the audit rule in Appendix D.1 treats a depth region as plateau-like when $D_{\mathrm{p50}}(\ell)$ remains below a predeclared threshold for k=3 consecutive layers. This descriptive label does *not* imply the existence of a true dynamical fixed point or attractor, and by itself it does not distinguish contraction-driven collapse from geometry-induced confinement.

Metric scale and cross-geometry comparisons. Because d_{Euc} and d_{Hyp} live on different scales, and d_{Hyp} can be strongly amplified near the boundary, we do *not* compare absolute drift magnitudes across geometries. Our cross-geometry claims are therefore qualitative rather than quantitative: we compare which regimes arise and how probe couplings (e.g., drift–energy–boundary interactions) evolve within each geometry across layers and trained depths. Accordingly, hyperbolic drift is interpreted jointly with boundary proxies (Section 4.4) and coordinate-step diagnostics (Appendix D.5).

Why boundary proximity can amplify measured dynamics. In the Poincaré ball, the Riemannian metric is conformal: $g_x = \lambda(x)^2 I$ with $\lambda(x) = 2/(1 - \|x\|_2^2)$. For a sufficiently small coordinate update Δx, this yields $d_{\mathrm{Hyp}}(x, x + \Delta x) \approx \lambda(x)\|\Delta x\|_2$, and the edgewise squared distances entering the Dirichlet energy are locally amplified by approximately $\lambda(x)^2$. As $\|x\|_2 \to 1$, identical coordinate-scale updates can therefore induce much larger geodesic drift and metric-aware roughness near the boundary, which is why hyperbolic probe signals must be interpreted jointly with boundary statistics. Equivalently, near the boundary the same coordinate-scale update can induce disproportionately large geodesic displacement, while squared-distance functionals such as metric-aware Dirichlet energy inherit approximately quadratic sensitivity in $\lambda(x)$; this is the minimal mechanism behind the boundary-coupled amplification pattern that we observe empirically.

Together with Dirichlet energy and label separability, interlayer drift serves as a complementary diagnostic signal: drift captures representation dynamics across layers, energy characterizes graph-local variation, and separability reflects task-relevant class structure.

4.2 Dirichlet energy: Graph-local roughness

To quantify local variation of node representations over the graph, we measure a metric-aware Dirichlet energy on a fixed evaluation edge set. For directed citation graphs, this means the symmetrized set $\mathcal{E}_{\mathrm{eval}}^{\mathrm{sym}}$ defined in Eq. (9); for undirected graphs, $\mathcal{E}_{\mathrm{eval}}^{\mathrm{sym}} = \mathcal{E}_{\mathrm{eval}}$.

$$E(\ell) = \frac{1}{|\mathcal{E}_{\mathrm{eval}}^{\mathrm{sym}}|} \sum_{(u,v)\in\mathcal{E}_{\mathrm{eval}}^{\mathrm{sym}}} d\Big(\mathbf{z}_u^{(\ell)},\, \mathbf{z}_v^{(\ell)}\Big)^2. \tag{8}$$

Here $d(\cdot,\cdot)$ denotes the distance induced by the representation geometry (e.g., Euclidean or hyperbolic). Intuitively, Dirichlet energy measures how rapidly neighboring node representations vary across edges and therefore provides a geometry-aware notion of graph-local roughness.

For directed citation graphs, we define

$$\mathcal{E}_{\mathrm{eval}}^{\mathrm{sym}} := \{(u,v) \mid (u,v) \in \mathcal{E}_{\mathrm{eval}} \text{ or } (v,u) \in \mathcal{E}_{\mathrm{eval}}\}. \tag{9}$$

This symmetrization ensures that the symmetric squared-distance form used by the energy probe is well-defined and comparable across conditions. It is used only for the energy probe; message passing itself remains directed as described in Section 5.

The evaluation edge subset is sampled once (excluding self-loops) and then reused across depths, geometries, and activation variants in order to eliminate resampling variance in the diagnostic measurements. Under this definition, a decreasing $E(\ell)$ indicates increasing similarity among neighboring node representations,

whereas an increase in $E(\ell)$ indicates growing neighborhood-level variation (or representational roughness). For upper-tail summaries reported in the appendix, we also use

$$E_{\mathrm{p90}}(\ell) := \mathrm{p90}_{(u,v)\in\mathcal{E}_{\mathrm{eval}}^{\mathrm{sym}}}\, d\Big(\mathbf{z}_u^{(\ell)},\, \mathbf{z}_v^{(\ell)}\Big)^2. \tag{10}$$

By fixing the evaluation edges, changes in $E(\ell)$ reflect representation dynamics rather than variance introduced by edge resampling.

4.3 Separability: Class structure at the output layer

Drift and energy do not directly measure whether representations remain discriminative. We therefore track label separability as a ratio of between-class to within-class dispersion, where larger values indicate stronger class structure.

Let $\mathcal{C}$ denote the label set, and let $\mathcal{V}_c = \{i \in V : y_i = c\}$ denote the node set for class c. Because separability is evaluated on the fixed probe set $\mathcal{S}$, we define the represented class set

$$\mathcal{C}_{\mathcal{S}} := \{c \in \mathcal{C} : |\mathcal{V}_c \cap \mathcal{S}| > 0\}.$$

Using only the probe nodes $\mathcal{V}_c \cap \mathcal{S}$ for $c \in \mathcal{C}_{\mathcal{S}}$, we define the class prototypes

$$\mu_c^{(\ell)} = \begin{cases} \dfrac{1}{|\mathcal{V}_c \cap \mathcal{S}|} \displaystyle\sum_{i\in\mathcal{V}_c\cap\mathcal{S}} \mathbf{z}_i^{(\ell)}, & \text{(Euclidean)}, \\ \exp_0\Bigg(\dfrac{1}{|\mathcal{V}_c \cap \mathcal{S}|} \displaystyle\sum_{i\in\mathcal{V}_c\cap\mathcal{S}} \log_0\Big(\mathbf{z}_i^{(\ell)}\Big)\Bigg), & \text{(hyperbolic)}, \end{cases} \tag{11}$$

where $\log_0$ and $\exp_0$ denote the Riemannian logarithmic and exponential maps at the origin of the Poincaré ball, respectively (Appendix D.4).

Implementation note (hyperbolic prototypes). The hyperbolic prototype in Eq. (11) (and the prototype mean in Eq. (14)) is computed as the tangent-space average at the origin, followed by $\exp_0$. This is a numerically stable and scalable approximation to the Fréchet mean for large probe sets; see Appendix D.2.

We define the within-class dispersion as

$$\mathrm{Disp}_w(\ell) = \sum_{c\in\mathcal{C}_{\mathcal{S}}} \frac{1}{|\mathcal{V}_c \cap \mathcal{S}|} \sum_{i\in\mathcal{V}_c\cap\mathcal{S}} d\Big(\mathbf{z}_i^{(\ell)},\, \mu_c^{(\ell)}\Big)^2, \tag{12}$$

and the between-class dispersion as

$$\mathrm{Disp}_b(\ell) = \frac{1}{|\mathcal{C}_{\mathcal{S}}|} \sum_{c\in\mathcal{C}_{\mathcal{S}}} d\Big(\mu_c^{(\ell)},\, \bar{\mu}^{(\ell)}\Big)^2, \tag{13}$$

where $\bar{\mu}^{(\ell)}$ denotes the mean of the class prototypes:

$$\bar{\mu}^{(\ell)} = \begin{cases} \dfrac{1}{|\mathcal{C}_{\mathcal{S}}|} \displaystyle\sum_{c\in\mathcal{C}_{\mathcal{S}}} \mu_c^{(\ell)}, & \text{(Euclidean)}, \\ \exp_0\Bigg(\dfrac{1}{|\mathcal{C}_{\mathcal{S}}|} \displaystyle\sum_{c\in\mathcal{C}_{\mathcal{S}}} \log_0\Big(\mu_c^{(\ell)}\Big)\Bigg), & \text{(hyperbolic)}. \end{cases} \tag{14}$$

Separability is defined as the global ratio

$$\mathrm{Sep}(\ell) = \frac{\mathrm{Disp}_b(\ell)}{\mathrm{Disp}_w(\ell) + \varepsilon}, \tag{15}$$

where $\varepsilon > 0$ is a fixed stabilizing constant (Table A.3). Separability is a single scalar computed on the probe set; it is not aggregated over nodes by a median or mean. Unless otherwise stated, reported separability values refer to the final layer $\ell = L$ on the test split.

Interpretation at extreme depth. $\mathrm{Sep}(\ell)$ is a ratio statistic. In extreme-depth regimes, both $\mathrm{Disp}_b(\ell)$ and $\mathrm{Disp}_w(\ell)$ may shrink, and the ratio in Eq. (15) can increase if the within-class term collapses faster (or because of the stabilizer ε). We therefore treat separability as a *secondary* class-structure summary and interpret it jointly with complementary evidence (drift/energy and, when needed, separability-independent homogenization proxies; Appendix B.6).

Remark. $\mathrm{Disp}_w(\ell)$ is defined as the sum of per-class dispersions, so each represented class contributes equally regardless of its size. Averaging $\mathrm{Disp}_w(\ell)$ over $|\mathcal{C}_\mathcal{S}|$ would only rescale $\mathrm{Sep}(\ell)$ by a constant factor (up to the stabilizer ε) and does not affect the qualitative depthwise trends. We do not treat separability as a stand-alone diagnostic, but rather as a projection of class structure whose interpretation requires complementary probes, particularly drift- and dispersion-based homogenization proxies.

4.4 Boundary monitoring in hyperbolic models

To monitor boundary effects in hyperbolic models, we define the *curvature-normalized radius*

$$r_i^{(\ell)} := \sqrt{c}\,\|\mathbf{z}_i^{(\ell)}\|_2 \in [0,1),$$

which reduces to the Euclidean ball radius in the unit-curvature case $c = 1$. We then track the mean normalized radius

$$\bar{r}(\ell) = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} r_i^{(\ell)}, \tag{16}$$

which measures average boundary proximity (larger $\bar{r}(\ell)$ means closer to the boundary), and the boundary occupancy ratio

$$B(\ell) = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \mathbb{I}\Big[r_i^{(\ell)} > \tau_r\Big], \tag{17}$$

where τ_r is a fixed threshold close to the boundary (we used $\tau_r = 0.95$; Table A.3).

These quantities are monotone proxies for boundary proximity and should not be confused with geodesic distance from the origin. Accordingly, drift and Dirichlet energy near the boundary should be interpreted jointly with $\bar{r}(\ell)$ and $B(\ell)$ (Appendix D.5).

In some plots (e.g., Figure 3), we also report the upper-tail boundary statistic

$$r_{\mathrm{p90}}(\ell) = \mathrm{p90}_{i\in\mathcal{S}}\, r_i^{(\ell)},$$

which serves as a sensitive indicator of boundary saturation. Appendix-only median boundary summaries use

$$r_{\mathrm{p50}}(\ell) := \mathrm{p50}_{i\in\mathcal{S}}\, r_i^{(\ell)}.$$

When helpful, we additionally report the corresponding boundary deficit

$$\Delta_i^{(\ell)} := 1 - r_i^{(\ell)},$$

whose values near zero indicate near-boundary saturation. In the appendix intervention study on OGBN-Arxiv, $r_{\mathrm{p90}}^{\mathrm{late}}$ denotes the mean of $r_{\mathrm{p90}}(\ell)$ over the designated late-layer window for the corresponding experiment.

4.5 Reproducibility and reporting choices

Fixed-point probing requires explicit choices that can affect the diagnostics. We fix and report the following: (i) how are $\mathcal{S}$ and $\mathcal{E}_{\mathrm{eval}}$ sampled (including random seeds); (ii) directed-edge handling for message passing and for$\mathcal{E}_{\mathrm{eval}}$; (iii) metrics and constants for d_{Hyp} (curvature); and (iv) numerical stabilizers such as ε and thresholds such as τ_r.

For visualization, we plotted the probe signals on logarithmic scales (and used log1p transforms, where indicated in figure captions) to accommodate a large dynamic range. All analyses used raw values.

5 Experimental setup

All probe nodes and evaluation edges are fixed across runs to eliminate resampling variance in diagnostic measurements. All diagnostics follow a fixed measurement protocol with released evaluation artifacts, enabling exact reproduction of the reported depthwise dynamics. Unless otherwise stated, endpoint metrics are reported as mean±std over three training seeds; fixed probe subsets and fixed evaluation edges remove resampling variance from the diagnostics, and the released per-seed CSVs show consistent qualitative depth trends. Main-text captions explicitly state when a plot shows single-checkpoint layerwise traces rather than across-seed aggregates and point to the seed-aggregated counterparts in Appendix E.2 where applicable.

5.1 Dataset: Large-scale patent citation graph

We constructed a directed patent citation graph from patent grants issued by the United States Patent and Trademark Office (USPTO). Nodes correspond to patents and directed edges represent citations (citing → cited).

Edge direction and temporal leakage. Raw citations are stored as directed edges (citing → cited). For message passing, we reverse the propagation direction (cited → citing) so that information flows from older patents to newer ones. This choice follows the knowledge-flow interpretation of citation graphs and, under our year-based temporal split, prevents training/validation nodes from receiving messages from future test nodes at any hop. Unless explicitly stated, all reported GNN results use this directed propagation graph; the original citation direction is used only for dataset semantics, and the Dirichlet-energy probe uses a separate symmetrized evaluation edge set (Section 4.2).

Each node was assigned a single CPC subclass label derived from the Cooperative Patent Classification (CPC) system, which we used as a representative technological category for the patent. Each patent was also associated with its textual abstract.

The textual abstracts were embedded using a text encoder and used as node features. For reproducibility, we document and release the embedding recipe (text model/version, embedding dimensionality, and preprocessing) together with the GNN code and CSV artifacts.

To reflect realistic deployment settings in which future patents must be inferred from past structures, we used a temporal split by grant year to ensure that training, validation, and test sets are disjoint in time.

The induced subgraph used in our experiments contains 479,533 patents and 1,864,213 citation edges. Node features are 384-dimensional abstract embeddings, and the label space consists of 316 CPC subclasses (see Appendix A.4 for full details).

Citation edges pointing from newer patents to older patents are preserved in the dataset but cannot introduce temporal leakage because message passing is performed in the cited → citing direction.

Why this dataset is the main stress test. We use the patent citation graph as the primary mechanism-revealing setting because it combines the properties most relevant to our question: large scale, directed citation structure under a temporal split, and a many-class label space with substantial hierarchy and imbalance. These conditions make depth-induced representational regime changes particularly likely to surface and make hyperbolic geometry a plausible representation choice. We therefore treat the patent graph as the main stress test for whether single-signal diagnostics miss boundary-coupled regimes, while the public benchmarks in Appendix C serve as transfer/reference checks rather than as a search for dataset-universal quantitative thresholds.

Embedding details. Node features were obtained from patent abstracts using a *fixed transformer-based sentence encoder*, which outputs d_{in}=384-dimensional vectors. We used the standard single-vector pooling of the encoder (mean pooling over token embeddings) and applied L2 normalization of resulting sentence vectors. The abstracts were tokenized using the subword tokenizer of the encoder and truncated to a fixed maximum sequence length. All embeddings were computed once prior to GNN training and reused across depths, geometries, and activation variants (no embedding fine-tuning). To preserve a double-blind review,

we have not mentioned the exact model name/version here, but we will release the model identifier, version hash, and full preprocessing recipe alongside code and CSV diagnostics. These details are publicly available in the camera-ready version.

Reproducibility and release plan (camera-ready). To reduce ambiguity about what "will be released" under double-blind review, the camera-ready submission will provide a *public reproducibility archive* organized into five named packages:

- **`diagnostic_csv_core`:** deterministic patent-subgraph diagnostics used in the main analysis (probe summaries, energy summaries, per-seed analysis CSVs, manifests, and the fixed-subset metadata required to interpret them).
- **`paper_figure_csv_small`:** compact figure-linked artifacts used to regenerate the paper figures, including the geometry/regime-view exports for Figure 2 and the GCN backbone-check summaries on the same patent subgraph for Figure 8, together with pre-rendered PDF/SVG figure files.
- **`public_fp_csv`:** public-benchmark fixed-point summary CSVs and the scripts used to regenerate the corresponding reference plots.
- **`patent_subgraph_bundle`:** a self-contained training-time induced-subgraph bundle containing node features/labels, edge index, temporal split masks, fixed probe/evaluation subsets, node-index mapping, and metadata.
- **`scripts_bundle`:** curated training/analysis/figure-generation scripts, released with both the original working filenames and clean canonical aliases plus a rename map.

Because review-time supplementary uploads are size-limited, the anonymized submission supplement will contain a lightweight *reviewer bundle* consisting of figure-linked artifacts, public-benchmark summaries, canonical scripts, and a compact diagnostic excerpt; the full five-package archive (including the patent subgraph bundle) will be posted immediately upon acceptance. Items withheld here for double-blind review (primarily the exact sentence-encoder identifier/version and any public hosting location) will be disclosed at camera-ready. Appendix A provides the authoritative package-level specifications and the figure-to-artifact mapping.

5.2 Models: Euclidean vs. hyperbolic GraphSAGE

We used GraphSAGE (Hamilton et al., 2017) as the common message-passing backbone. The Euclidean variant embeds nodes in $\mathbb{R}^d$. The hyperbolic variant embeds in the Poincaré ball $\mathbb{B}_c^d$ and uses standard hyperbolic operations (tangent-space aggregation and Riemannian maps), similar to those in prior hyperbolic neural network implementations (Ganea et al., 2018; Chami et al., 2019). Both variants share the same neighborhood sampling and aggregation structure; the key difference is the representation geometry and metrics used by the probes.

Hyperbolic baselines and tuning parity. To avoid conflating the hyperbolic geometry effects with underoptimization, we report two hyperbolic variants throughout the paper. This emphasis on tuning parity is deliberate: recent audit work shows that conclusions in hyperbolic graph learning can be highly sensitive to Euclidean baseline strength and evaluation protocol design (Katsman & Gilbert, 2025). HypA is a representative off-the-shelf hyperbolic instantiation under the shared training protocol. HypT is a minimally tuned hyperbolic control (tangent-space classifier head, input scaling α_{in}, curvature c, and a learning-rate split between the encoder and the head; Table A.3) introduced to ensure competitive shallow-depth performance under a comparable tuning budget to the Euclidean baseline. In particular, HypT matches the Euclidean baseline at shallow depth (e.g., L=2; Table 2 and Appendix B.1) while preserving the same qualitative late-depth signature. Thus, the hyperbolic regime reported here is not explained by trivial shallow-depth underoptimization or by a uniformly weak hyperbolic baseline. HypT isolates curvature-aware aggregation without explicit radial control (i.e., without the radius regularization used in Section 6.3).

To address concerns about curvature choice, we additionally report a curvature robustness sweep for HYPT over $c \in \{1, 2, 3, 5\}$ in Appendix B.7, showing that varying c changes the absolute probe scales but does not remove the qualitative very-deep failure regime.

We sweep the depth $L \in \{2, 4, 8, 16, 32\}$ unless otherwise stated (with additional plots and tables in the Appendix). Unless otherwise stated, ReLU activation was used.

5.3 Training protocol and pointwise nonlinearity ablation

The models were trained using minibatch neighbor sampling. To reduce undertraining confounds in deep models, we allocated a larger epoch budget for $L \geq 16$ (800 vs. 200 epochs for $L \leq 8$; Table A.3), while keeping the optimizer and other hyperparameters fixed across depths. To isolate the role of *pointwise* nonlinearities in depth-induced dynamics, we compared standard ReLU networks to their ablated counterparts obtained by replacing ReLU with an identity map at the same depth (ReLU→identity). We emphasize that this ablation removes only pointwise activation, whereas other non-Euclidean operations (e.g., Riemannian maps in hyperbolic models) remain nonlinear. Task performance is reported using *test macro-F1* as a reference; however, our analysis focused on layerwise probe signals. Unless otherwise stated, each run is evaluated at the checkpoint selected by the validation metric used in the training pipeline, and the reported test macro-F1 is taken at that validation-selected checkpoint. Because the task involved many classes (316 CPC subclasses) with a strong imbalance under a temporal split, macro-F1 values can be small in absolute terms; therefore, we treated them as supplementary endpoint references rather than as optimization targets.

5.4 Diagnostics computation

All probes were computed post-training for each trained model. We fixed $\mathcal{S}$ (for drift, separability, and boundary monitoring) and $\mathcal{E}_{\text{eval}}^{\text{sym}}$ (for Dirichlet energy) and reused them across depths, geometries, and activation variants. Training uses stochastic minibatch neighbor sampling, but the post-training probes are computed on fixed $\mathcal{S}$ and fixed $\mathcal{E}_{\text{eval}}^{\text{sym}}$ with a fixed evaluation-time sampling seed (`seed_eval`=0), so the released diagnostics are deterministic given a trained checkpoint.

Computational overhead. Fixed-point probing is computed after training and requires only forward-pass representations of the fixed subsets. With a probe node set $\mathcal{S}$ and evaluation edges $\mathcal{E}_{\text{eval}}^{\text{sym}}$, costs scale linearly with depth: drift, separability, and boundary summaries are $O(L|\mathcal{S}|)$ distance evaluations (plus class-wise aggregation), whereas the Dirichlet energy is $O(L|\mathcal{E}_{\text{eval}}^{\text{sym}}|)$ distance evaluations. No backpropagation is required, and the probes can be computed in a streaming fashion over layers without storing full-graph embeddings, making the overhead comparable to a few evaluation-time forward passes on the same subsets.

6 Results

To distinguish boundary-associated metric amplification from unusually large coordinate updates, we explicitly separate ambient coordinate-step magnitudes from geodesic drift (Appendix D.5; Figure D.3) and provide layerwise coupling plots that tie boundary proximity directly to drift and Dirichlet energy.

An important scope note is that the very deepest endpoints are poor for all three model families in this patent stress test. Our claim is therefore not that the probes uncover a hidden useful deep regime, but that they separate different *pathways into failure*. Under a shared fixed-subset protocol, Euclidean and hyperbolic models fail differently enough that endpoint accuracy or any single familiar probe obscures the distinction.

Table 1 summarizes the characteristic probe signatures observed across depth-induced regimes, highlighting that low drift or low energy is not a universal indicator of health or collapse.

Reading guide for Section 6. The main-text evidence is organized in three layers. Figures 1–3 carry the core mechanism narrative on the patent graph. Figures 4, 5, 6, and 7 provide secondary output-layer and probe–endpoint summaries. Figure 8 then provides a Euclidean GCN backbone check on the same patent subgraph. Public-benchmark reference runs and additional robustness studies remain in the appendix.

Table 1: Qualitative regime summary of depth-dependent representation dynamics. The first row summarizes the common early-layer phase; the geometry-specific rows summarize the late-depth regime after that phase. Arrows indicate qualitative trend direction only and do not imply strict monotonicity or universal behavior across runs. The mean radius $\bar{r}$ is defined only for hyperbolic models.

Regime	Drift D	Energy E	Separability Sep	$\bar{r}$ (hyperbolic only)
Both: early stabilization	$\downarrow$	$\downarrow$	$\uparrow$ / $\approx$	$\uparrow$ (mild)
Euclidean: late-depth degeneration	$\uparrow$ (mild)	$\uparrow$ (mild)	$\downarrow$	–
Hyperbolic: boundary-associated amplification (late)	$\uparrow$ (sharp)	$\uparrow$ (sharp)	$\approx$ / $\downarrow$	$\uparrow\rightarrow 1$

6.1 Layerwise dynamics at fixed depth

To diagnose whether late-depth degradation reflects contraction-driven collapse or geometry-induced metric amplification, we examined layerwise drift and metric-aware Dirichlet energy at a fixed depth ($L = 16$). Figure 1 reveals the qualitative difference in the late-layer behavior between the Euclidean and hyperbolic geometries. Although both models exhibit an early regime in which the interlayer drift and Dirichlet energy decreased or remained low, the late-depth dynamics diverged. In Euclidean space, the drift and energy increase only slightly in the later layers. By contrast, the hyperbolic model enters a boundary-associated regime in which both quantities are sharply amplified as the representations approach Poincaré boundary, as indicated by a concurrent increase in $\bar{r}(\ell)$.

This pattern is more consistent with geometry-induced metric amplification than with a purely contraction-driven oversmoothing account. The distances and energies were measured using the geometry-specific metric d (Euclidean for EUCLID and Poincaré for the hyperbolic models). Because these metrics differ, we interpreted the *trends* using the layer indices and their coupling to the boundary proximity, rather than the absolute cross-geometry magnitudes. The shared axes in Figure 1(a,b) are therefore used to show within-geometry regime shape, not absolute cross-geometry magnitude comparability. Figure A.5 illustrates the drift and energy relative to the mean radius $\bar{r}(\ell)$. Appendix D.5 contrasts coordinate steps and geodesic drift near the boundary, and public benchmark sweeps are reported in Appendix C (Figures C.9 and C.10).

6.2 Representation geometry and boundary pressure across depth

Figure 2 is the main mechanism figure in the patent-citation case study. The right-hand panels carry the claim directly: only in the hyperbolic model does late-depth boundary proximity co-amplify geodesic drift and metric-aware graph-local roughness. The left-hand panels provide deep-state context at the same trained depth, showing that the Euclidean state remains compact whereas the deep hyperbolic state is boundary-adjacent in tangent space. We do not claim the absence of oversmoothing in the hyperbolic models; rather, in the evaluated patent-graph runs the late-depth diagnostic signature is dominated by boundary-coupled amplification rather than by a smooth Euclidean-style contraction pattern.

6.3 Boundary intervention ablation at fixed depth

Although the layerwise analyses show a strong association between boundary proximity and late-depth amplification in hyperbolic models, correlation alone does not establish mechanism. Within this scope, the radius-penalty experiment (Figure 3) provides timing-sensitive diagnostic evidence consistent with a boundary contribution to the late-depth amplification observed here.

Concretely, we augment the supervised objective with a hinge-squared penalty on the (normalized) Poincaré ball radius at the final layer:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda\, \mathcal{L}_{\text{rad}}, \qquad \mathcal{L}_{\text{rad}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[\max\left(0,\ \sqrt{c}\, \|\mathbf{z}_i^{(L)}\|_2 - \rho_0\right) \right]^2, \tag{18}$$

where $\mathcal{B}$ denotes the minibatch seed nodes used for supervision, c is the hyperbolic curvature, and $\rho_0 \in (0, 1)$ is the target radius cap, and $\lambda \geq 0$ controls the intervention strength. We used ρ_0=0.90 for the controlled

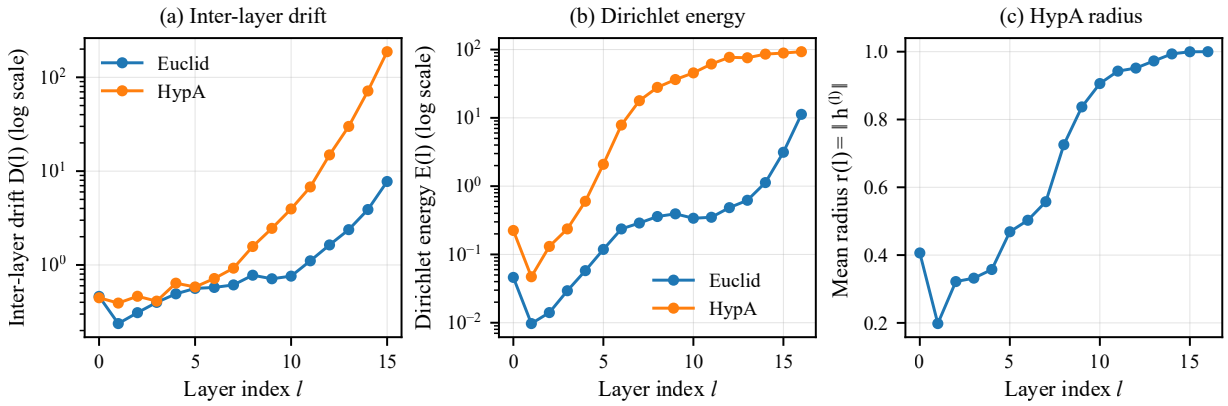

Figure 1: Layerwise representation dynamics at fixed depth ($L = 16$) on the patent citation graph. Panels show single-checkpoint layerwise summaries (no across-seed averaging), computed on the same fixed subset pair: the fixed probe nodes $\mathcal{S}$ for drift and the fixed symmetrized evaluation edges $\mathcal{E}_{\text{eval}}^{\text{sym}}$ for energy. (a) Interlayer drift $D(\ell)$, defined as the median geometry-consistent distance between representations at consecutive layers over the fixed probe nodes. (b) Metric-aware Dirichlet energy $E(\ell)$ computed on the fixed symmetrized evaluation edge set. (c) Mean normalized radius $\bar{r}(\ell)$ for the hyperbolic model, where $r_i^{(\ell)} = \sqrt{c}\,\|\mathbf{z}_i^{(\ell)}\|_2$; values closer to 1 indicate stronger boundary saturation. Distances and energies are evaluated under the geometry-specific metric. The shared axes in panels (a,b) are used to show regime shape only; absolute magnitudes are interpreted within each geometry, not across geometries. Boundary saturation in hyperbolic models should not be interpreted as benign stabilization; rather, it marks a boundary-associated amplification regime in which local metric distortion dominates despite continued coordinate-level updates. Seed-aggregated mean±std counterparts for the core diagnostics are provided in Appendix E.2; probe-subset resampling robustness is reported in Appendix A.6 and Appendix E.3.

runs shown in Figure 3. The *late-only* setting applied the same penalty only during the final stage of training (the last 25% of the epochs in our implementation), testing whether the delayed boundary control is sufficient.

Unless explicitly stated (i.e., in Figure 3), all the other hyperbolic results in this study used a standard baseline without radius regularization ($\lambda = 0$). By varying the strength and timing of this constraint, we distinguish no control, delayed control, and global control from the start of training. Figure 3 shows the resulting layerwise behavior at a fixed depth $L = 32$ for the hyperbolic model (HypA). The purpose of this intervention is not to preserve task performance or to propose a radius regularizer, but to perturb the boundary-associated trajectory. The response is timing-sensitive and nonmonotonic: weak late-only control remains boundary-saturated, while stronger always-on control changes the saturated trajectory and reduces the high-energy portion of the curve in the pre-terminal window (Appendix E.5). This asymmetric response is expected from the geometry in Section 4: local geodesic drift scales to first order with the conformal factor $\lambda(x)$, whereas metric-aware Dirichlet energy is a squared-distance quantity and is therefore approximately $\lambda(x)^2$-sensitive. Thus radius perturbations can reduce the energy portion more visibly than median drift, while drift remains entangled with layerwise direction changes and schedule timing. We therefore interpret the intervention as evidence consistent with a boundary contribution, not as a clean monotonic control law.

These results support an interpretation consistent with a boundary contribution in this setting: delayed control does not move the boundary-saturated trajectory, while stronger global control reduces the high-energy portion of that trajectory without providing a clean monotonic attenuation of all diagnostics. This is also consistent with the first-order versus squared-distance sensitivity of the probes: energy is expected to be more directly affected by radius changes than drift, which also depends on the direction and timing of the layerwise trajectory. Appendix E.5 reports the corresponding pre-terminal window summary and explains why terminal-window averages are not used as the primary numeric evidence.

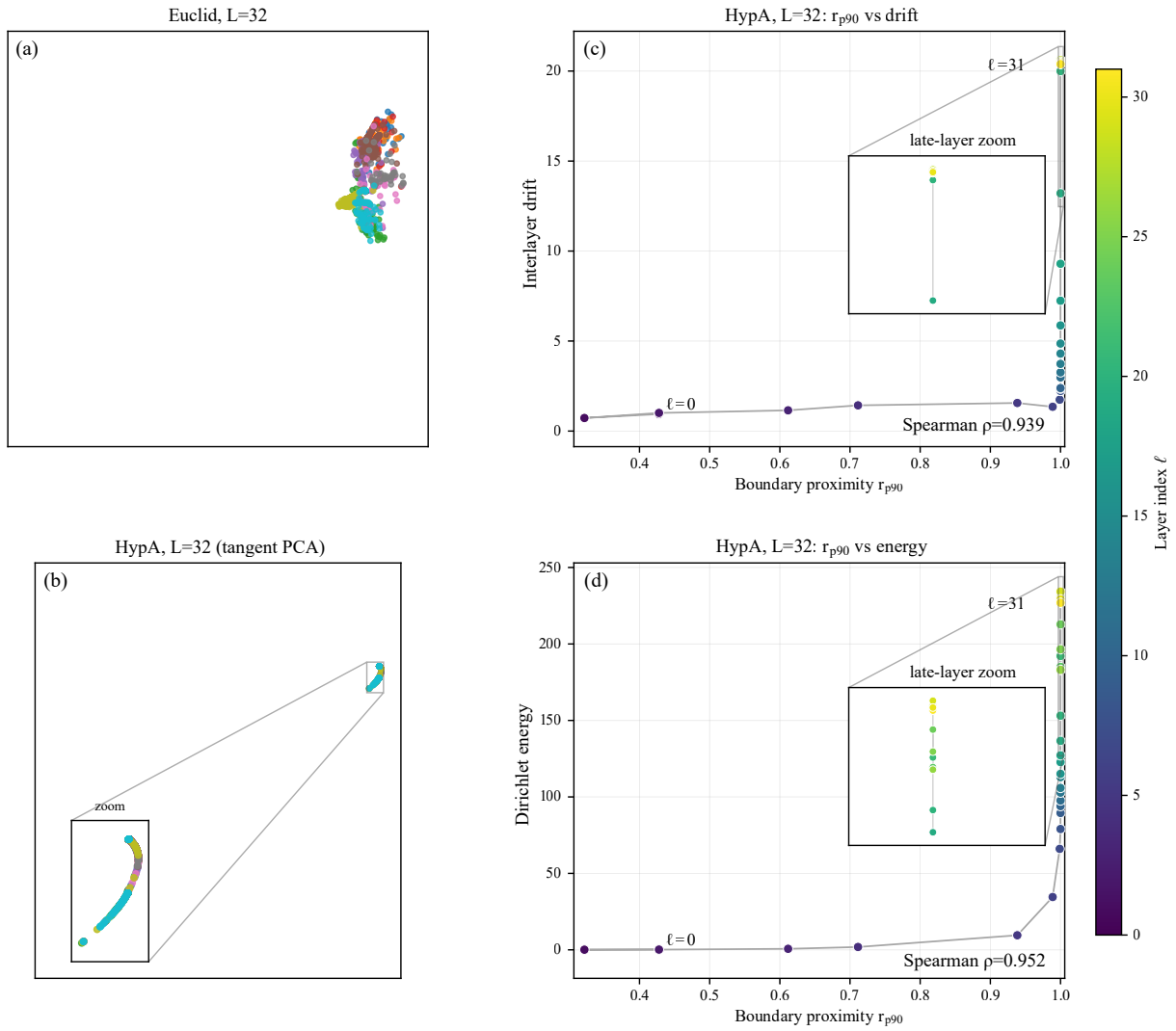

Figure 2: **Boundary-coupled amplification as the main late-depth signature in the patent-citation stress test.** Panels (a,b) provide deep-state context at $L = 32$: Euclidean representations remain compact in a common Euclidean PCA view, whereas deep HypA occupies a boundary-adjacent configuration in tangent-space PCA. Panels (c,d) are the primary evidence and show single-checkpoint layerwise couplings for the same trained HypA model (no across-seed averaging). For HypA at $L = 32$, as upper-tail boundary proximity $r_{\mathrm{p90}}(\ell)$ approaches 1, both interlayer drift and metric-aware Dirichlet energy rise sharply. Points are colored by layer index, and the insets zoom the late-layer regime. All four panels are built from the same deterministic subset drawn from the fixed probe set $\mathcal{S}$; the left-hand geometry panels show the deep $L = 32$ checkpoints only, so that the figure prioritizes late-depth state and mechanism rather than shallow-versus-deep comparison. Euclidean representations are displayed in a common Euclidean PCA view, whereas HypA is shown by tangent-space PCA via $\log_0$. These tangent-space PCA panels are visualization-only flattenings and should not be read as preserving geodesic structure; all quantitative diagnostics in this paper are computed on the observed representations under the geometry-consistent metric defined in Section 4. For completeness, the original same-subset boundary-versus-depth summaries are retained in Figure D.2. In this stress-test setting, these coupled trends are consistent with a boundary-coupled amplification regime rather than a smooth Euclidean-style contraction. Seed-aggregated mean±std counterparts are provided in Appendix E.2; probe-subset resampling robustness is reported in Appendix A.6 and Appendix E.3.

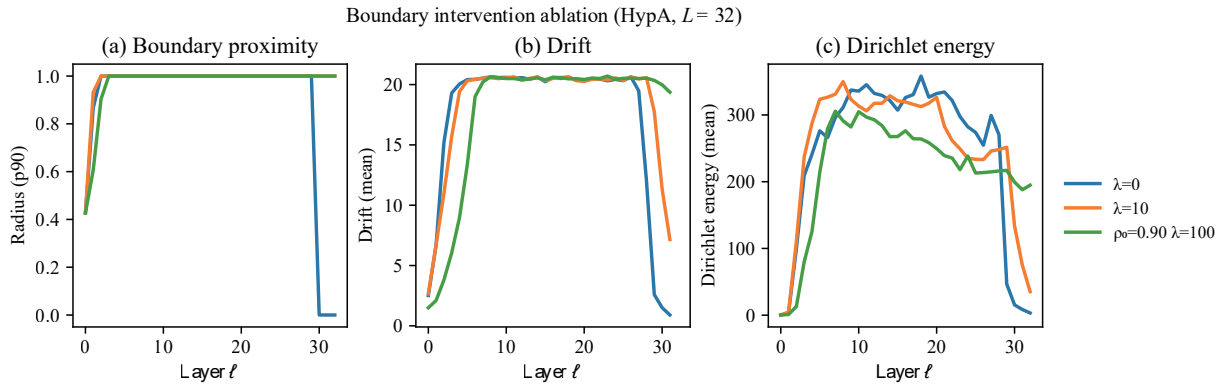

Figure 3: Boundary intervention ablation at fixed depth ($L = 32$, HypA). Curves are per-checkpoint layerwise summaries for the displayed intervention settings (no across-seed averaging). (a) Layerwise boundary proximity measured by the 90th-percentile normalized radius $r_{\mathrm{p90}}(\ell)$. (b) Layerwise mean interlayer drift. (c) Layerwise mean metric-aware Dirichlet energy. Curves correspond to three radius-penalty schedules in Eq. (18): no control ($\lambda = 0$), weak late-only control ($\rho_0 = 0.90$, $\lambda = 10$, active only in the final training stage), and stronger global control ($\rho_0 = 0.90$, $\lambda = 100$, active from epoch 0). The late-only condition remains boundary-saturated and does not materially change the amplified trajectory. The global condition also remains near the boundary in the pre-terminal saturated window, but reduces the high-energy portion of the curve. The drift response is timing-sensitive and nonmonotonic, so this intervention is used as a diagnostic perturbation consistent with a boundary contribution, not as evidence for a clean monotonic radius-control law or a performance-preserving regularizer. Appendix E.5 reports the pre-terminal numeric summary; Appendix E.2 reports seed-aggregated uncertainty summaries for the core diagnostics.

6.4 Depth sweep: Separability at the output layer

As a compact, *secondary* output-layer summary of class structure (Section 4.3), we report the final-layer separability as a function of model depth for the Euclidean baseline and both hyperbolic variants. This depth-sweep view is included only as a coarse endpoint complement to the layerwise probes in Figures 1 and 3 and to the geometry/boundary visualization in Figure 2; the regime interpretation throughout the paper continues to rely on joint probe couplings rather than on Sep alone.

Figure 4 shows that the endpoint class-structure summary is not monotonic in depth. In Euclidean space, separability improves up to moderate depth (L=8), drops again at L=16, and shows an apparent rise at L=32. We do not interpret that late Euclidean increase as recovered class structure: as discussed in Section 4.3, Sep is a ratio statistic and can become misleading when both between-class and within-class dispersion collapse. For this reason, the Euclidean extreme-depth endpoint is read jointly with the drift/energy probes in Table 2 and with the separability-independent homogenization proxy in Figure 5.

The two hyperbolic variants retain nontrivial endpoint separability over a broader depth range, but in different ways. HypA remains comparatively high from shallow to mid depth and then declines gradually toward L=32, whereas HypT peaks at intermediate depth and softens again at larger depth. Including both hyperbolic variants makes clear that this endpoint summary is not tied to a single hyperbolic parameterization. At the same time, these endpoint curves do not replace the layerwise diagnosis: in hyperbolic space, nontrivial final-layer separability can coexist with late-layer drift amplification and boundary-coupled sensitivity, as shown by the layerwise probes and intervention views.

Public benchmark depth sweeps are reported in Appendix C as supportive reference checks for probe–performance decoupling rather than as the main mechanism evidence of this paper.

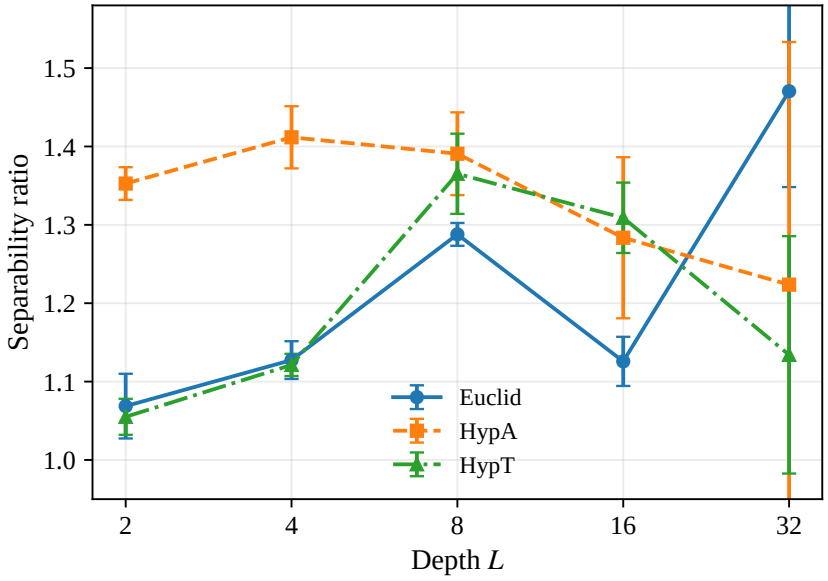

Figure 4: **Class separability vs. depth (secondary output-layer summary).** Final-layer separability (Eq. (15)) across the standard depth sweep $L \in \{2, 4, 8, 16, 32\}$ for the Euclidean baseline, HypA, and HypT. Points and error bars show mean±std over three training seeds. This is a compact endpoint summary and should be interpreted jointly with the layerwise drift/energy probes (and, in hyperbolic space, boundary proxies), rather than in isolation. Euclidean separability is nonmonotonic, and the increase at L=32 is not taken as evidence of recovered class structure; it is read jointly with the separability-independent homogenization proxy in Figure 5 and Appendix B.6. The two hyperbolic variants retain nontrivial endpoint separability over a broader depth range, despite the late-layer amplification documented by the layerwise probes.

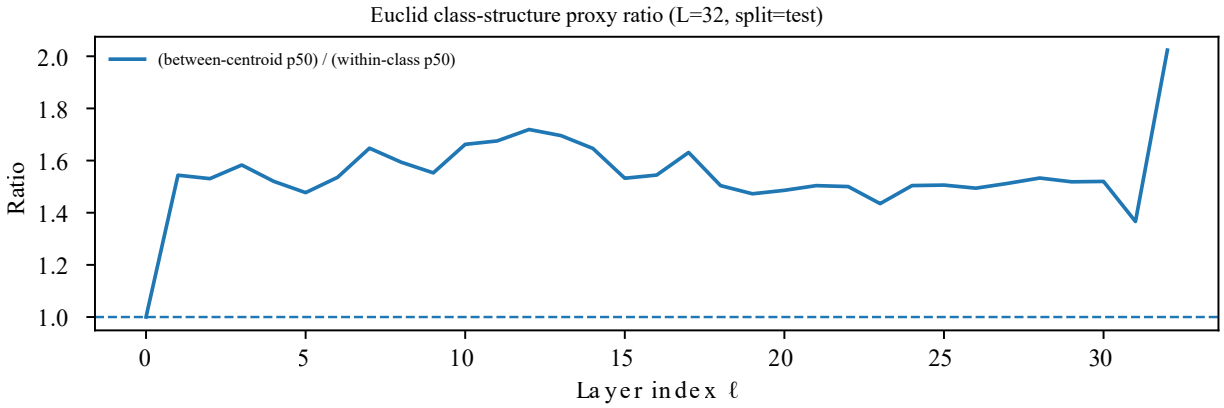

Figure 5: **Separability-independent Euclidean homogenization proxy (patent subgraph, L=32).** We report the ratio between a between-class centroid distance and a within-class dispersion proxy (both medians over sampled pairs), normalized by their layer-0 values. The ratio approaching unity indicates class-wise homogenization consistent with oversmoothing.

6.5 Pointwise nonlinearity ablation (ReLU→Identity)

To distinguish the geometric effects from those induced by pointwise nonlinearities, we performed controlled ablation in which the ReLU activation was replaced by the identity map at fixed depth. Figure 6 compares the

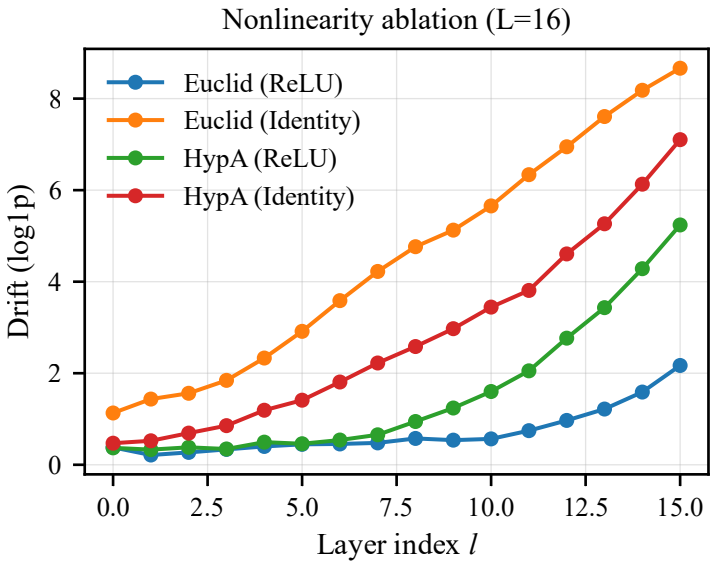

Figure 6: Effect of pointwise nonlinearity ablation at fixed depth ($L = 16$). Each curve is a single-checkpoint layerwise summary (no across-seed averaging), shown for Euclidean and hyperbolic GraphSAGE models with ReLU activations and with ReLU replaced by the identity map. All other architectural components and training settings are held fixed; drift is plotted on a log1p scale. This figure is a representative checkpoint-level trace; seed-aggregated counterparts for the core diagnostics are provided in Appendix E.2.

standard ReLU networks with their ReLU-ablated counterparts (ReLU→Identity) under otherwise identical architectural and training settings. This ablation is conceptually related to linearized GNN variants, such as SGC (Wu et al., 2019), but is applied here as a diagnostic tool to probe depthwise representation dynamics rather than modeling choice.

Figure 6 shows that replacing ReLU with an identity map substantially increases the interlayer drift in both Euclidean and hyperbolic geometries, indicating that pointwise nonlinearities play a stabilizing role in the learned depth trajectory. However, even in the absence of pointwise nonlinearities, hyperbolic models retain a pronounced late-layer drift amplification. This persistence indicates that the observed amplification cannot be attributed solely to activation effects, but is consistent with a geometry-driven mechanism associated with proximity to the Poincaré boundary. Degree-resolved heatmaps for the ReLU-ablated setting (Figure A.2) further show that this amplification can become extreme at greater depths, reinforcing the interpretation that the boundary effects dominate late-layer behavior in hyperbolic models.

6.6 Additional decompositions: Depth × degree and class

Figure A.1 shows the degree-resolved drift heatmaps for ReLU networks at L=8 and L=32, and Figure A.2 reports the same for ReLU-ablated (ReLU→Identity) networks. These decompositions show how the interaction between the depth and graph heterogeneity can yield localized but severe late-stage amplifications, particularly in hyperbolic spaces at large depth. Figure A.4 shows the class-resolved view at L=16.

Summary of depth-dependent regimes. Layerwise probing shows that depth-induced behavior is neither uniform nor monotonic but instead emerges through a sequence of regimes. Figure 1 shows an early stabilization regime in both geometries, where the interlayer drift and (metric-aware) Dirichlet energy decrease across the layers. Figure 2 provides a complementary geometry-level and coupling view on the same fixed subset: deep Euclidean representations remain compact, whereas deep hyperbolic representations enter a boundary-saturated regime in which boundary proximity co-amplifies drift and graph-local roughness. High boundary proximity alone is not the defining signature: upper-tail saturation is already visible at shallow

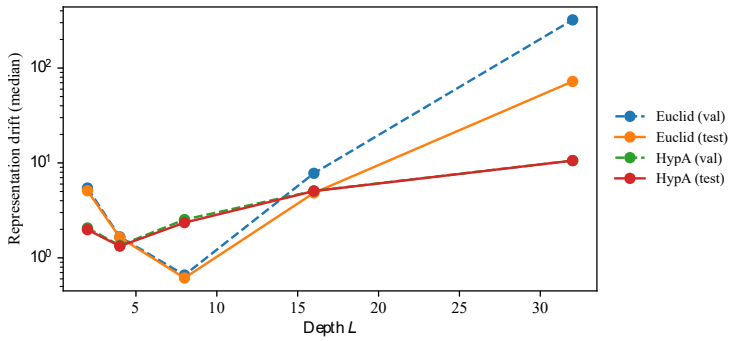

Figure 7: Representation drift versus depth on the patent citation graph. Curves are plotted directly from split-specific diagnostic summaries of the displayed checkpoints (no across-seed averaging in this figure). They show the median interlayer drift (p50 over the validation- or test-restricted probe nodes) at the final transition $\ell = L - 1$ as a function of depth L for Euclidean and hyperbolic models. The y-axis is shown on a logarithmic scale. This figure uses displayed checkpoints directly; seed-aggregated uncertainty summaries for the core diagnostics are provided in Appendix E.2.

depths (Appendix D.5), whereas the distinctive late-depth regime is the sharp co-amplification of geodesic drift and metric-aware roughness together with the coordinate-step and intervention evidence. Figure 4 further shows that final-layer class separability is not monotonic in depth: the Euclidean model improves at moderate depth but becomes unreliable at extreme depth, whereas the two hyperbolic variants retain nontrivial endpoint separability over a broader depth range.

These trends were heterogeneous across the graph subpopulations. Figure A.1 shows that the degree-conditioned drift exhibits a *layer-dependent* pattern. Low-degree nodes stabilize more slowly in the early layers, whereas late-layer amplification (when present) concentrates in high-degree bins and is substantially stronger in the hyperbolic model. Figure A.4 provides a class-resolved view, indicating that the hyperbolic late-layer elevation is distributed across classes rather than driven by a small subset.

Overall, these results support treating depth as a structured dynamic axis with distinct representational regimes and motivate multisignal, geometry-aware diagnostics to identify when additional layers cease to be representationally productive.

Hyperbolic baselines. We consider two hyperbolic variants. *HypA* corresponds to a representative off-the-shelf hyperbolic instantiation using a shared training protocol. *HypT* is a minimally tuned hyperbolic control (tangent-space classifier head, adjusted curvature, and learning-rate split) introduced to make shallow-depth underoptimization an unlikely primary explanation for the boundary-associated drift amplification. HypT used the same training budget and base optimizer hyperparameters as the other models (Table A.3); the only changes were a small set of hyperbolic-specific knobs (tangent-space head, curvature, input scaling, and light learning-rate split; Appendix B.1). Notably, both variants exhibited qualitatively similar late-depth behavior, indicating that the observed amplification was not a simple artifact of a weak baseline.

Figure 7 compares the final-transition drift across depths on the validation and test splits. The close validation/test alignment indicates that the probe trends are not a split-specific artifact, while the strong depth dependence of the drift remains decoupled from endpoint performance, which is reported separately in Table 2 and Appendix B.1. In particular, probe signals can change substantially even when endpoint metrics vary more smoothly with depth.

To verify that our hyperbolic observations are not artifacts of an underoptimized baseline, we evaluated a tuned hyperbolic variant (**HypT**) in the patent subgraph. HypT used a tangent-space classifier head and a minimal set of hyperbolic-specific tuning knobs (input scaling α_{in}, curvature c, and a learning-rate split between the manifold encoder and classifier head; see Appendix B.1). As listed in Table B.1, HypT matches the Euclidean performance at shallow depths (e.g., 0.297 ± 0.002 vs. 0.309 ± 0.004 test macro-F1 at L=2;

Table 2: Representative probe signals and endpoint performance on the patent citation graph. Entries report mean±std over three training seeds. Drift reports the median over probe nodes (p50) at the final transition $\ell = L - 1$. Dirichlet energy is the last-layer mean over 200k sampled evaluation edges. Separability is computed once per run as a global scalar on the probe set (Eq. 15) and then aggregated over seeds. **All entries correspond to the standard baseline without radius regularization** ($\lambda = 0$); the intervention sweep is reported separately in Figure 3. Sep is included only as a *secondary* class-structure summary: at extreme depth it can rise because of ratio shrinkage in the denominator (Section 4.3), so it should be read jointly with drift/energy and with the separability-independent homogenization proxy in Figure 5. At L=32, all three model families are already in an endpoint-collapse regime; in this highly imbalanced 316-class setting, macro-F1 values on the order of 10^{-3} are consistent with near-single-class predictions and should be read only as collapse-level references rather than as meaningful geometry-specific differences. The Euclidean and hyperbolic blocks should therefore be read separately: values are intended for qualitative cross-depth comparison within each geometry, not for absolute cross-geometry magnitude comparison. (For visualization, some plots use log/log1p scales; table entries are raw.)

Geometry	Depth	Macro-F1 (test)	Drift D_{p50}	Dirichlet energy	Separability
Euclid	2	0.309 ± 0.004	5.06 ± 0.02	50.1 ± 3.1	1.07 ± 0.04
Euclid	4	0.239 ± 0.006	1.54 ± 0.03	101.9 ± 6.6	1.13 ± 0.02
Euclid	8	0.120 ± 0.040	0.85 ± 0.25	72.8 ± 22.6	1.29 ± 0.01
Euclid	16	0.013 ± 0.003	3.47 ± 1.21	52.8 ± 8.0	1.13 ± 0.03
Euclid	32	0.001 ± 0.000	21.27 ± 8.16	23.8 ± 7.8	1.47 ± 0.12
HypA	2	0.104 ± 0.002	4.03 ± 0.11	174.5 ± 6.5	1.35 ± 0.02
HypA	4	0.094 ± 0.022	1.08 ± 0.15	192.4 ± 7.3	1.41 ± 0.04
HypA	8	0.085 ± 0.007	3.79 ± 0.90	233.6 ± 11.2	1.39 ± 0.05
HypA	16	0.015 ± 0.001	20.48 ± 0.03	261.5 ± 12.9	1.28 ± 0.10
HypA	32	0.001 ± 0.000	20.44 ± 0.03	245.2 ± 32.5	1.22 ± 0.31
HypT	2	0.297 ± 0.002	4.27 ± 0.06	115.9 ± 1.1	1.06 ± 0.02
HypT	4	0.257 ± 0.006	1.98 ± 0.07	118.4 ± 0.2	1.12 ± 0.01
HypT	8	0.150 ± 0.009	6.98 ± 0.55	115.6 ± 0.2	1.37 ± 0.05
HypT	16	0.031 ± 0.005	11.83 ± 0.00	122.8 ± 3.6	1.31 ± 0.04
HypT	32	0.001 ± 0.000	11.82 ± 0.02	69.4 ± 57.5	1.13 ± 0.15

mean±std over three training seeds), while exhibiting the same qualitative late-depth degradation pattern at large L. This argues against shallow-depth underoptimization as the primary driver of the observed late-depth hyperbolic regime.

This systematic decoupling motivated the use of multisignal, geometry-consistent diagnostics beyond endpoint performance alone.

Table 2 summarizes representative probe values together with endpoint test performance. All depths were trained under the same optimizer settings, with a larger epoch budget for $L \geq 16$ to reduce undertraining confounds (800 vs. 200 epochs; Table A.3). At the deepest setting L=32, the endpoint metrics for all three model families collapse to the 10^{-3} range; under the strongly imbalanced 316-class label space, that scale is consistent with near-single-class prediction and is reported only to mark endpoint collapse rather than to compare geometries at useful accuracy. The key point is not that the probes act as accuracy surrogates—they do not—but that they distinguish geometry-dependent internal regimes even when endpoint behavior alone compresses them into a single "deep models fail" narrative.

6.7 Backbone check on the same patent subgraph (Euclidean GCN)

To bound the architectural scope of the main claim, we repeated the patent-subgraph experiment with a Euclidean GCN backbone on the same patent subgraph, using the same fixed probe-node set $\mathcal{S}$ and fixed evaluation-edge set $\mathcal{E}_{\text{eval}}$. This is not intended as an optimized benchmark comparison: because the GCN variant is evaluated with neighbor-sampling loaders, its degree normalization is approximate on sampled

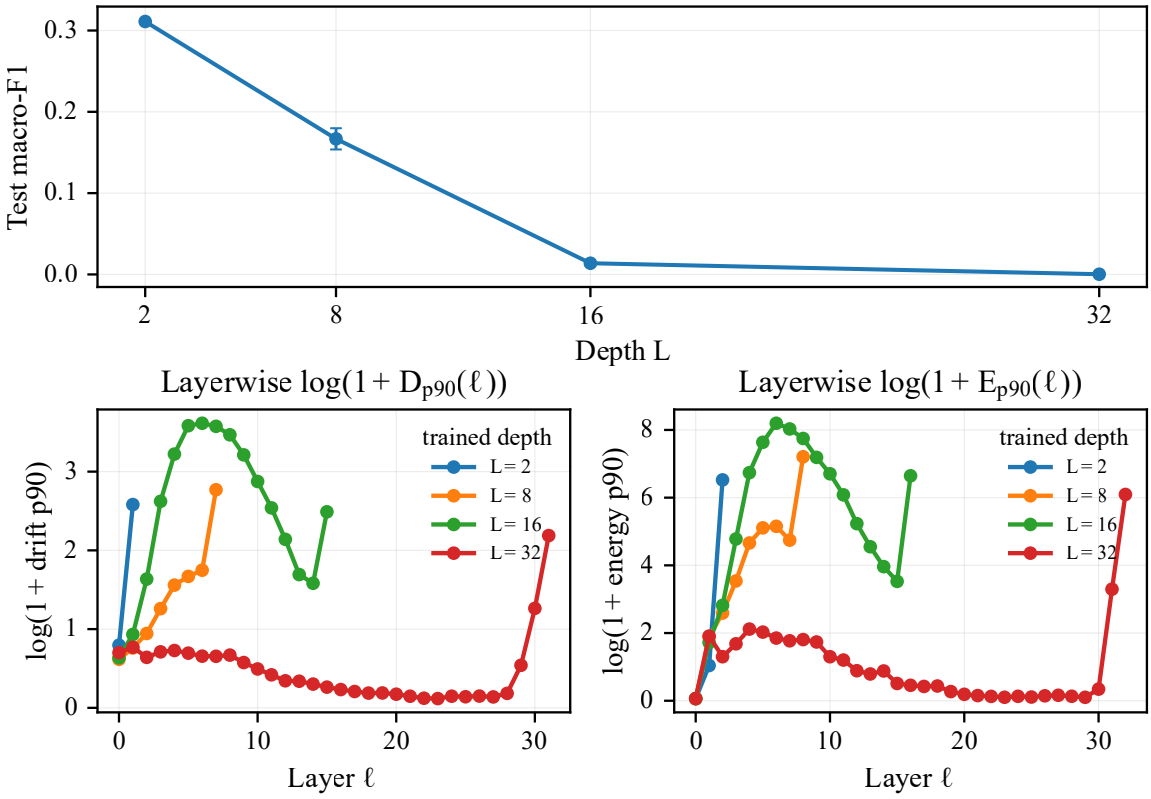

Figure 8: **GCN backbone check on the same patent subgraph (Euclid; three training seeds).** Left: test macro-F1 versus depth (mean±std over seeds). Middle: layerwise $\log(1 + D_{\mathrm{p90}}(\ell))$ for models trained at depths $L \in \{2, 8, 16, 32\}$. Right: layerwise $\log(1 + E_{\mathrm{p90}}(\ell))$ under the same fixed evaluation-edge set. We use upper-tail $p90$ summaries here because this backbone check is intended as a scope-bounding stress probe and the mid-depth GCN spike is most visible in the tail. The qualitative probe–performance decoupling persists beyond GraphSAGE, but the degeneration profile is architecture-dependent: performance collapses with depth, while the L=16 GCN model exhibits a pronounced mid-depth spike in both drift and Dirichlet energy. Seed-aggregated core-diagnostic uncertainty summaries for the main patent GraphSAGE runs are provided in Appendix E.2.

subgraphs. We use it only to ask whether the fixed-point probing protocol remains informative beyond a single GraphSAGE configuration.

Figure 8 shows that the qualitative probe–performance decoupling persists. Test macro-F1 again deteriorates sharply with depth (0.311±0.003 at L=2, 0.167±0.013 at L=8, 0.014±0.003 at L=16, and $2.6\times10^{-4}\pm2.7\times10^{-5}$ at L=32; mean±std over three training seeds), but the internal profile is architecture-dependent: for L=16, both drift and metric-aware Dirichlet energy peak in mid-depth layers rather than only at the final transition. We therefore treat this as a scope-bounding result rather than a generality claim: the protocol is not tied to GraphSAGE, but the onset and layerwise placement of degeneration remain backbone-dependent.

Interpreting Euclid at L=32: low energy with large drift. Table 2 reports that the deepest Euclidean model (L=32) simultaneously exhibits *extremely low* final-layer Dirichlet energy but a *large* final-transition drift. This combination indicates within-layer homogenization with persistent cross-layer translation/rotation, rather than recovered structure. The probes measure different axes: Dirichlet energy summarizes *within-layer* graph-local variation, whereas drift measures *across-layer* per-node displacement. At extreme depths, the representations can therefore become nearly homogeneous *within* each layer (low energy and collapsed class structure) while the remaining low-variance embedding configuration continues to translate or rotate across

layers, yielding large successive-layer displacements. In the same regime, the apparent increase in Sep should not be interpreted as recovered class structure, but as a ratio artifact (Section 4.3; Appendix B.6).

7 Discussion and limitations

What the protocol adds. Our central claim is diagnostic, not algorithmic. Fixed-point probing treats depth as a representation trajectory and shows that endpoint accuracy, smoothness alone, or separability alone can compress distinct late-depth behaviors into the same coarse judgment that "deep models fail." In the patent stress test, the protocol instead distinguishes different pathways into failure: Euclidean models trend toward class-structure loss consistent with oversmoothing, whereas hyperbolic models exhibit a late-depth boundary-coupled amplification signature.

Relation to prior work. The Euclidean behavior aligns with classical oversmoothing accounts that connect repeated message passing to representation homogenization (Oono & Suzuki, 2020; Rusch et al., 2023). The hyperbolic behavior is consistent with well-known metric distortion near the Poincaré boundary, where small coordinate changes can correspond to large geodesic displacements (Nickel & Kiela, 2017; Ganea et al., 2018). Our novelty is therefore not a new scalar, but a protocol that makes the *joint shape* of drift, roughness, class structure, and boundary pressure reproducibly visible on fixed subsets.

Scope and bounded generality. We intentionally do not claim dataset-universal quantitative thresholds or architecture-universal regime boundaries. The main evidence is the patent-citation stress test. We also do not claim that the patent graph is generically "hyperbolic enough" to justify hyperbolic models a priori; recent work argues that geometry-task alignment matters separately from graph hyperbolicity alone (Naddeo et al., 2026). The tuned hyperbolic control, the radius-penalty intervention, the Euclidean GCN backbone check on the same patent subgraph, the curvature sweep, and the public-benchmark reference runs reported in the appendix are included to bound the interpretation rather than to establish universal laws. Appendix E.1 makes this separation explicit: the same-data GCN and public checks support protocol portability, whereas the detailed boundary-coupled mechanism remains a bounded patent-graph finding. The GCN check is particularly useful here because it shows that the protocol remains informative beyond GraphSAGE, while also showing that the onset and layerwise placement of degeneration can change with the backbone.

Mechanistic interpretation and limitations. The radius-penalty experiment provides timing-sensitive diagnostic evidence consistent with a boundary contribution to the late-depth hyperbolic amplification observed under unconstrained training; the revised interpretation is deliberately cautious. We do not claim the absence of oversmoothing in hyperbolic models; rather, boundary-coupled amplification dominates the late-depth diagnostic signature in the evaluated settings. The intervention is not presented as a universal remedy, and reduced energy under strong control does not imply that the controlled model remains task-useful. In the patent intervention, always-on control reduces the high-energy portion of the saturated trajectory in the pre-terminal window, while the drift response remains timing-sensitive and nonmonotonic (Appendix E.5). This asymmetry is consistent with the probe definitions: geodesic drift is locally first-order in the boundary conformal factor $\lambda(x)$, whereas the Dirichlet energy uses squared distances and is therefore approximately quadratic in $\lambda(x)$. The exploratory OGBN-Arxiv pilots reinforce this caution: late-only controls were effectively inert in the tested L=16 settings, whereas always-on controls could induce nonmonotonic regime switches with substantial endpoint cost. These results support an interpretation consistent with a timing-sensitive boundary contribution, but not a smooth control band on the public benchmark.

Implications and future work. Practically, the results argue for reading depth through multiple geometry-consistent probes rather than through endpoint accuracy or smoothness alone. The protocol is architecture-agnostic at the measurement level and can be reused to audit normalization, residual, boundary-control, or alternative geometry choices under fixed subsets. Important next steps include attention-based and diffusion-style backbones, alternative boundary-avoidance interventions, and using probe-derived signals for training-time model selection or early stopping. Finally, separability should be treated cautiously: because it is a ratio statistic, it can be nonmonotonic when both between- and within-class dispersion shrink, so we recommend interpreting it jointly with the other probes and with the proxy-choice checks in Appendix B.6.

8 Conclusion

We introduced fixed-point probing, a fixed-subset, geometry-consistent post-training protocol for auditing depthwise representation dynamics. The primary contribution of this paper is the protocol itself, while the patent citation graph serves as a bounded stress-test case study in which the protocol reveals geometry-dependent late-depth behavior.

On this stress test, the protocol separates different pathways into very-deep failure. Euclidean GraphSAGE is most consistent with class-structure degradation and homogenization, whereas the evaluated hyperbolic models enter a boundary-coupled regime in which approaching the Poincaré-ball boundary co-amplifies geodesic drift and metric-aware roughness. Tuned baselines, a radius-penalty intervention, a GCN backbone check on the same patent subgraph, and public-benchmark reference runs in the appendix bound this interpretation without supporting dataset-universal thresholds.

More broadly, the paper argues for treating depth not only as a model-size hyperparameter but as an observed representation trajectory. Under that view, familiar probes become substantially more informative when evaluated on fixed subsets and interpreted jointly in a geometry-consistent way.

References

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=i80OPhOCVH2.

Gregor Bachmann, Gary Bécigneul, and Octavian-Eugen Ganea. Constant curvature graph convolutional networks. In *International Conference on Machine Learning*, 2020.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Deli Chen, Yankai Lin, Wei Li, Peng Li, and Jie Zhou. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020a.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1725–1735. PMLR, 2020b. URL https://proceedings.mlr.press/v119/chen20v.html.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gL-2A9Ym.

Weiqi Guan and Zihao Shi. Measuring over-smoothing beyond dirichlet energy, 2025. URL https://arxiv.org/abs/2512.06782.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017.

Isay Katsman and Anna Gilbert. Shedding light on problems with hyperbolic graph learning. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=rKAkp1f3R7.

Guohao Li, Matthias Müller, Bernard Ghanem, and Vladlen Koltun. Training graph neural networks with 1000 layers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6437–6449. PMLR, 2021. URL https://proceedings.mlr.press/v139/li21o.html.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3538–3545, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11604.

Dionisia Naddeo, Jonas Linkerhägner, Nicola Toschi, Geri Skenderi, and Veronica Lachi. Hyperbolic graph neural networks under the microscope: The role of geometry-task alignment, 2026. URL https://arxiv.org/abs/2602.01828.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, 2017.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1ldO2EFPr.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. In *8th International Conference on Learning Representations (ICLR 2020)*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Hkx1qkrKPr.

T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks, 2023. URL https://arxiv.org/abs/2303.10993.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature, 2021.

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

Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. In *Advances in Neural Information Processing Systems*, 2023.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5453–5462. PMLR, 2018. URL https://proceedings.mlr.press/v80/xu18c.html.

Menglin Yang, Min Zhou, Tong Zhang, Jiahong Liu, Zhihao Li, Lujia Pan, Hui Xiong, and Irwin King. Hyperbolic graph neural networks: A review of methods and applications, 2022. URL https://arxiv.org/abs/2202.13852. Last revised 2025.

Kaicheng Zhang, Piero Deidda, Desmond Higham, and Francesco Tudisco. Are we measuring oversmoothing in graph neural networks correctly? In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r3DOISnvHD.

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

Kaixiong Zhou, Xiao Huang, Daochen Zha, Rui Chen, Li Li, Soo-Hyun Choi, and Xia Hu. Dirichlet energy constrained learning for deep graph neural networks. In *Advances in Neural Information Processing Systems*, 2021.

APPENDIX

Supplementary and Reproducibility Details

Reader's map. Appendix A documents the reproducibility and released diagnostics. Appendix B reports additional depth, architecture, and decomposition results. Appendix C reports the public benchmark validation results and exploratory OGBN-Arxiv intervention pilots. Appendix D provides a geometric background and probe interpretation. Appendix E collects reviewer-motivated scope, uncertainty, budget, runtime, boundary-intervention, and hidden-width sensitivity checks added in the revision.

A Experimental and Reproducibility Details

How to read Appendix A. Appendix A documents the experimental protocol and reproducibility details underlying all reported figures. Table A.1 serves as the central index that maps each figure to a minimal set of deterministic intermediate artifact(s) (primarily CSVs, plus coordinate/manifest exports where needed) and the exact script(s) required for regeneration. The following sections specify (i) how fixed probe/evaluation subsets are constructed and reused, (ii) training hyperparameters, (iii) dataset construction and statistics, and (iv) released diagnostics and their seed-dependence.

Reproducibility Statement

All figures in the main text and Appendices were generated from deterministic intermediate artifacts produced by the analysis pipeline. Most figures are CSV-based; the geometry visualization additionally uses deterministic coordinate exports and a selected-node manifest. For each figure, Table A.1 lists the primary artifact file(s) and script(s) used to generate the plots. Unless otherwise stated, all stochastic components, including node sampling for probe evaluation, edge sampling for Dirichlet-energy evaluation, and neighbor sampling during probing, were controlled by fixed random seeds, enabling the exact regeneration of the reported plots and tables.

A.1 Probe Node Set and Evaluation Edge Set

Operational fixed-point notion. Throughout this study, *fixed-point-like* (near-stationary) is used in the operational sense: a layer regime in which the interlayer drift summary (e.g., $D_{\text{mean}}(\ell)$ or $D_{\text{p50}}(\ell)$; Eqs. (4)–(5)) becomes small and remains small over consecutive layers. This does not imply convergence to a single point, the existence of a true attractor, or the availability of a dynamical-systems fixed-point theorem; instead, it identifies a plateau in the depthwise representation dynamics under the metric $d(\cdot, \cdot)$.

Training-time induced-subgraph and fixed subsets. All probe/evaluation subsets ($\mathcal{S}$, $\mathcal{E}_{\text{eval}}$) were sampled once from the *training-time induced-subgraph* (i.e., the nodes and edges available under the temporal split) and reused across depths, geometries, and activations. This design eliminates the variance arising from resampling artifacts and enables layerwise comparisons across conditions.

Table A.1: Mapping between figures and their primary deterministic artifacts. Each entry specifies the minimal set of released artifact file(s) and the exact script(s) required to regenerate the corresponding figure. For visualization-only entries (e.g., Figure 2), the primary artifacts include deterministic coordinate/manifest exports in addition to CSVs.

Figure	Primary artifact file(s)	Script(s)
Fig. 1	`fp_L16_euclid_relu.csv`, `fp_L16_hypA_relu.csv`, `energy_L16_euclid_relu.csv`, `energy_L16_hypA_relu.csv`	`05_probe_fixed_point_nb.py` `09_dirichlet_energy_nb.py`
Fig. 2	`patent_geom_hypA_seed0_coords.csv`, `patent_geom_hypA_seed0_phase_data.csv`, `patent_geom_hypA_seed0_selected_nodes.json`, `patent_geom_hypA_seed0_manifest_v12.json`, `phase_hypA_L32_seedEval0_n1k.merged.csv`	`make_geometry_visualization_v12.py` (also released as `make_fig2_geometry_phase_view.py`)
Fig. 4	`figure4_sep_from_table2_release_aligned.csv`, `Figure4_HypT_timeslike.pdf`, `Figure4_HypT_timeslike.svg`	`make_figure4_hypt_timeslike.py`
Fig. 6	`fp_L16_euclid_relu.csv`, `fp_L16_hypA_relu.csv`, `fp_L16_euclid_identity.csv`, `fp_L16_hypA_identity.csv`	`05_probe_fixed_point_nb.py`
Fig. 8	`depth_summary.csv`, `probe_layer_summary.csv`, `energy_layer_summary.csv`, `final_layer_table.csv`	`aggregate_plot_patent_gcn_backbone_check_v2.py` (also released as `make_figB2_gcn_backbone_check.py`)
Fig. A.1	`drift_groups_L8_*_relu.csv`, `drift_groups_L32_*_relu.csv`	`06_drift_by_group_nb.py`
Fig. A.2	`drift_groups_L8_*_identity.csv`, `drift_groups_L32_*_identity.csv`	`06_drift_by_group_nb.py`
Fig. A.3	`drift_groups_L16_*_relu.csv`	`06_drift_by_group_nb.py`
Fig. A.4	`drift_groups_L16_*_relu.csv`	`06_drift_by_group_nb.py`
Fig. A.5	`fp_L16_hypA_relu.csv`, `energy_L16_hypA_relu.csv`	`05_probe_fixed_point_nb.py` `09_dirichlet_energy_nb.py`
Fig. B.2	`curvature_sweep_summary.csv`	`collect_curvature_sweep_summary.py` `sweep_curvature_train_and_probe.py`
Fig. D.4	`curvature_sweep_summary_with_boundary_deficit.csv`	`plot_normalized_rp90_vs_curvature.py`
Fig. D.5	`curvature_sweep_summary_with_heterogeneity.csv`	`plot_heterogeneity_metrics_vs_curvature.py`

Probe node set $\mathcal{S}$. We defined a fixed probe node set $\mathcal{S}$ to diagnose the depthwise representation dynamics independent of stochastic neighbor sampling. The probe nodes were uniformly sampled at random from the induced subgraph using a fixed random seed. Unless otherwise stated, $|\mathcal{S}| = 50{,}000$. For each trained checkpoint, the drift and separability were computed on $\mathcal{S}$, with the distances instantiated under the geometry-specific metric (Euclidean distance for Euclid; Poincaré geodesic distance for HypA/HypT). We recorded the induced-subgraph construction parameters (`seed`, `hops`, and `seeds_per_split`) and the resulting $|V|, |\mathcal{E}|$ in `meta.json` and treated this file as the source of truth for reproducing the released bundle.

Evaluation edge set $\mathcal{E}_{\text{eval}}$. We computed the metric-aware Dirichlet energy (Eq. (8)) on a fixed evaluation edge set $\mathcal{E}_{\text{eval}}$ restricted to the induced subgraph. Self-loops were excluded from the analysis. For directed graphs, we evaluated the symmetrized set $\mathcal{E}_{\text{eval}}^{\text{sym}}$ (Eq. (9)) by treating the edges as undirected. Unless otherwise stated, we used $|\mathcal{E}_{\text{eval}}| = 200{,}000$ sampled once with a fixed seed and reused across all depths/geometries/activations.

A.2 Reproducibility Commands

This section describes the exact commands used to generate the released deterministic artifacts from the trained model checkpoints. Each command corresponds to a *deterministic analysis*

step in the post-training pipeline and produces the CSV/JSON/manifest artifacts that are subsequently used to generate all figures reported in the main text and in Appendix A.

Notably, these commands operate exclusively on (i) fixed trained checkpoints, (ii) released induced-subgraph bundles (Appendix A.4), and (iii) fixed probe/evaluation subsets (Appendix A.1). No stochastic resampling was performed beyond those explicitly controlled by the listed random seeds. Consequently, the commands listed in Table A.2 have the same inputs, and the seeds regenerate the released artifacts *bit-for-bit* (or, for figure PDFs, from deterministic intermediate artifacts).

Table A.2 lists the minimal set of commands required to produce each class of intermediate artifact. For readability, we show only the key arguments that affect reproducibility; full commands (including paths and environment setup) are available in the accompanying scripts. In the released `scripts_bundle`, the original working-tree filenames are accompanied by clean canonical aliases. Table A.2 lists the working-tree script names and, where relevant, notes the corresponding public alias.

A.3 Hyperparameter and Training Configuration

Table A.3 summarizes the hyperparameters used in all the experiments. All models shared the same optimizer and hyperparameters across depths and geometries. We increased the epoch budget for deeper models ($L \geq 16$) to reduce undertraining concerns; this biases *against* finding depth failures that are merely optimization artifacts.

A.4 Dataset Construction and Data Card

Source data and graph construction. We constructed a directed citation graph from USPTO utility patents granted between 2010 and 2020. Nodes correspond to patents and directed edges represent citations (citing $\rightarrow$ cited) with respect to temporal ordering. For GNN message passing, we used the reverse propagation direction (cited $\rightarrow$ citing) such that neighborhood aggregation for a newer patent draws only from older cited patents. This prevents information flowing from future nodes back to past nodes under a temporal split. The node features are abstract embeddings, and the labels are CPC subclasses after frequency filtering. We used a temporal split by grant year to reflect realistic deployment settings: training (2010–2016), validation (2017–2018), and test (2019–2020). For context, the underlying raw citation graph prior to temporal induction contains approximately 3.1 M nodes and 11.4 M directed edges (counts may vary slightly with the crawl snapshot and filtering); in this paper, all experiments and released diagnostics operate on the fixed induced subgraph described as follows.

Training-time induced subgraph. All probe/evaluation subsets (node probe set $\mathcal{S}$ and evaluation edge set $\mathcal{E}_{\text{eval}}$) were sampled from *fixed induced subgraph* used in our runs (constructed once prior to training and released as part of the reproducibility bundle). The bundle includes nodes from all temporal splits (training/validation/test), with labels masked by the split; notably, message passing uses the cited$\rightarrow$citing propagation direction, which blocks future-to-past information flow under the temporal split. Therefore, probe construction is compatible with the temporal setting and does not introduce additional leakage beyond the standard transductive observation of features/edges.

Table A.2: **Reproducibility commands.** Commands used to generate the primary deterministic artifacts for all figures (main text and Appendices). Only key arguments are listed for readability; full commands (including paths and environment setup) are available in the accompanying scripts.

Script	Command (key arguments shown)	Output artifact(s)
05_probe_fixed_point_nb.py	-sub_dir SUBGRAPH_BUNDLE -ckpt MODEL -mode euclid|hypA|hypT -depth L -seed_source subset -num_seeds 50000 -seed 0 -batch_size 2048 -fanout_eval 100 -num_workers 0 -hyp_drift_metric euclid|poincare -act relu|identity -seed_eval 0 -out_csv fp_L*_*_*.csv	fp_L*_*_*.csv
06_drift_by_group_nb.py	(same core arguments as 05_probe_fixed_point_nb.py) group binning options (degree / class) -out_csv drift_groups_L*_*_*.csv	drift_groups_L*_*_*.csv
09_dirichlet_energy_nb.py	-sub_dir SUBGRAPH_BUNDLE -ckpt MODEL -mode euclid|hypA|hypT -depth L -act relu -eval_edges eval_edges_200k_sym.npy -out_csv energy_L*_*_relu.csv	energy_L*_*_relu.csv
sweep_curvature_train_and_probe.py	-run_root out/curvature_sweep -curvatures 1.0,2.0,3.0,5.0 -depths 2,4,8,16,32 -seeds 0,1,2 -ckpt_template TEMPLATE -subgraph_dir SUBGRAPH_BUNDLE -subset_ids subset_node_idx.json -eval_edges eval_edges_200k_sym.npy -probe_script 05_probe_fixed_point_nb.py -energy_script 09_dirichlet_energy_nb.py -seed_eval 0 -mode hypT -act relu	fp_c*_L*_hypT_*.csv, energy_c*_L*_hypT_*.csv
collect_curvature_sweep_summary.py	-run_root out/curvature_sweep -out_csv curvature_sweep_summary.csv	curvature_sweep_summary.csv
plot_normalized_rp90_vs_curvature.py	-in_csv curvature_sweep_summary.csv -out_csv curvature_sweep_summary_with_boundary_deficit.csv	curvature_sweep_summary_with_boundary_deficit.csv
plot_heterogeneity_metrics_vs_curvature.py	-in_csv curvature_sweep_summary_with_boundary_deficit.csv -out_csv curvature_sweep_summary_with_heterogeneity.csv	curvature_sweep_summary_with_heterogeneity.csv
make_geometry_visualization_v12.py (make_fig2_geometry_phase_view.py)	-ckpt_dir CKPT_BEST -sub_dir SUBGRAPH_BUNDLE -subset_json subset_node_idx.json -phase_csv phase_hypA_L32_seedEval0_n1k.merged.csv -euclid_depths 2,32 -hyper_depths 2,32 -hyper_mode hypA -seed_eval 0 -topk_classes 10 -per_class_cap 200 -out_prefix patent_geom_hypA_seed0	patent_geom_hypA_seed0_coords.csv, patent_geom_hypA_seed0_phase_data.csv, patent_geom_hypA_seed0_selected_nodes.json, patent_geom_hypA_seed0_manifest_v12.json
aggregate_plot_patent_gcn_backbone_check_v2.py (make_figB2_gcn_backbone_check.py)	-out_base out/patent_gcn_backbone_check -figure_prefix PAPER_FIGURE_CSV_SMALL/figureB2_gcn_backbone_check/fig_patent_gcn_backbone_check	depth_summary.csv, probe_layer_summary.csv, energy_layer_summary.csv, final_layer_table.csv

Released subgraph bundle (data card). For reproducibility, we provide a self-contained *subgraph bundle*, which includes the following: (i) induced-subgraph node features and labels (`x.npy`, `y.npy`); (ii) induced-subgraph edge index (`sub_edge_index.npy`); (iii) temporal split masks (`train_mask.npy`, `val_mask.npy`, `test_mask.npy`); (iv) fixed probe node set identifier file (`subset_node_idx.json` or equivalent); (v) fixed symmetrized evaluation edges for Dirichlet energy (`eval_edges_200k_sym.npy`); (vi) an optional node-index back-map (`x.node_idx.npy`) for relating induced-subgraph rows to the original graph indexing; and

Table A.3: Hyperparameters and training configuration used in all experiments.

Item	Value
Optimizer	Adam
Learning rate	3×10^{-3}
Weight decay	1×10^{-4}
Hidden dimension d	128
Batch size	2048
Neighbor fanout (training)	15
Neighbor fanout (evaluation)	100
Epochs	200 ($L \leq 8$), 800 ($L \geq 16$)
Hyperbolic curvature c	1.0 (HypA), 3.0 (HypT)
Projection $\varepsilon_{\text{proj}}$	1×10^{-5}
Radius threshold τ_r	0.95
Radius-penalty cap ρ_0 (intervention only)	0.90
Radius-penalty weight λ (intervention only)	0 / 10 / 100
Late-only schedule (intervention only)	last 25% of epochs
Stabilizer ε	1×10^{-8}

(vii) a metadata file (`meta.json`) recording snapshot identifiers, preprocessing configurations, checksums, and induced-subgraph statistics used for the reported runs.

Induced-subgraph statistics (exact). In the released bundles used in the reported experiments, the induced subgraph contains $|V| = 479{,}533$ nodes and $|\mathcal{E}| = 1{,}864{,}213$ directed edges, with an input feature dimension $d_{\text{in}} = 384$. We used $|\mathcal{S}| = 50{,}000$ probe nodes and a balanced temporal split with $|\mathcal{V}_{\text{train}}| = |\mathcal{V}_{\text{val}}| = |\mathcal{V}_{\text{test}}| = 50{,}000$ nodes. Labels correspond to 316 CPC subclasses ($|C| = 316$). Class frequencies were highly imbalanced across the induced subgraph, and the minimum/median/maximum per-class counts were 60/497/63,220, respectively (median computed over subclasses with nonzero counts). Given the released bundle, these statistics are deterministic and support exact reproduction.

Fixed evaluation edges. For Dirichlet-energy evaluation, we sampled a fixed set of $|\mathcal{E}_{\text{eval}}| = 200{,}000$ citation edges from the induced subgraph with a fixed seed, excluded self-loops, and symmetrized them for evaluation (Eq. (9)). This edge set was reused across all depths, geometries, and activations to eliminate resampling variance.

A.5 Released reproducibility packages and diagnostics

This section defines the *authoritative specification* of the released reproducibility artifacts. The public camera-ready archive is organized into five packages: `diagnostic_csv_core`, `paper_figure_csv_small`, `public_fp_csv`, `patent_subgraph_bundle`, and `scripts_bundle`. Review-time supplementary material will contain a lightweight reviewer bundle with the figure-linked artifacts, public-benchmark summaries, canonical scripts, and a compact diagnostic excerpt; the full packages will be posted immediately upon acceptance.

Package roles.

- **`diagnostic_csv_core`:** deterministic patent-subgraph diagnostics and per-seed summaries used in the main paper.

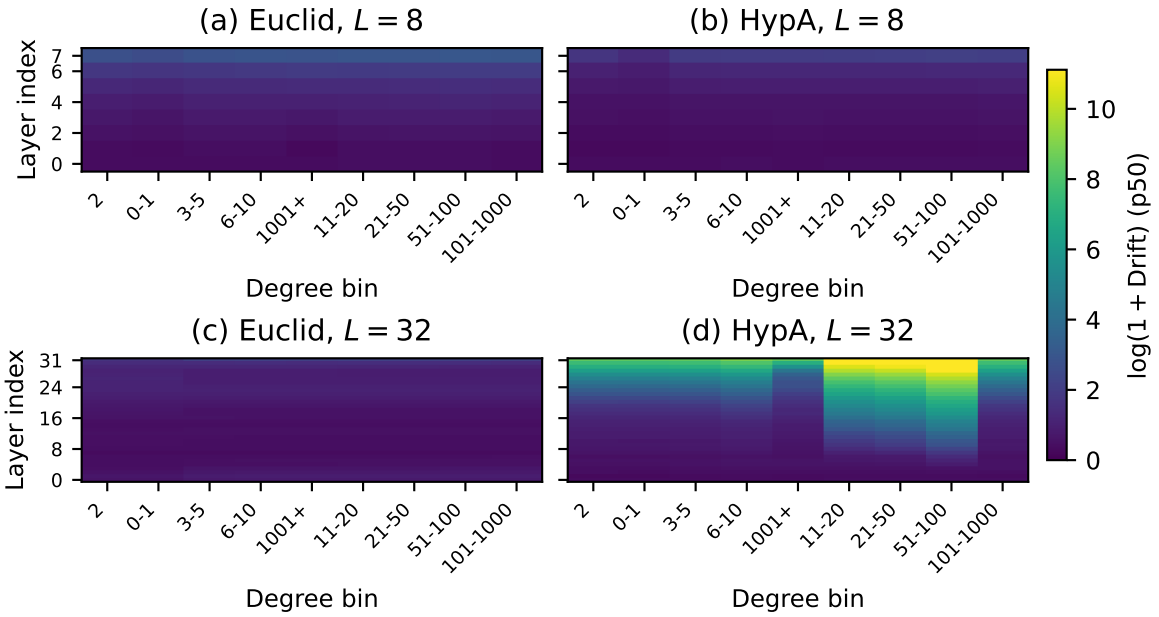

Figure A.1: Degree-resolved drift heatmaps (median, p50) for ReLU models: (a) Euclidean, $L = 8$, (b) Hyperbolic (HypA), $L = 8$, (c) Euclidean, $L = 32$, and (d) Hyperbolic (HypA), $L = 32$. Color scales are fixed within this figure to enable quantitative comparison across panels.

- **`paper_figure_csv_small`:** compact figure-linked artifacts (CSV/JSON/manifest exports and pre-rendered PDF/SVG figures) for the principal paper figures, including Figure 2 and Figure 8.
- **`public_fp_csv`:** public-benchmark summary CSVs and regeneration scripts.
- **`patent_subgraph_bundle`:** the full self-contained induced-subgraph data bundle described in Appendix A.4.
- **`scripts_bundle`:** curated scripts released under both original filenames and clean canonical aliases, accompanied by a rename map.

At the file level, the released diagnostics include (i) layerwise interlayer drift $D(\ell)$, metric-aware Dirichlet energy $E(\ell)$, label separability $\mathrm{Sep}(\ell)$, and hyperbolic boundary statistics at each training checkpoint; (ii) degree- and class-resolved drift summaries underlying Appendix Figures A.1–A.4; (iii) curvature-sweep diagnostics for HypT (per-run fixed-point and energy CSVs, plus derived summary CSVs) underlying Appendix Figures B.2–D.5; and (iv) deterministic coordinate/manifest exports for the geometry and regime-view visualizations. CSV and JSON artifacts include metadata fields (depth L, geometry, activation, training seed, split identifier, and/or selection manifest) sufficient to support exact regeneration of the reported plots. Appendix B exclusively reuses these released artifacts and does not introduce additional specifications.

A.6 Probe-set resampling ablation

To assess whether the qualitative depthwise regime transition could be a probe-set sampling artifact, we reran the fixed-point drift probe while *resampling* the probe node set $\mathcal{S}$

Table A.4: Random seeds and construction parameters influencing each intermediate artifact. All reported results are reproducible by fixing the listed seeds and the induced-subgraph construction parameters recorded in `meta.json`.

Artifact/CSV file	Random seed(s)/fixed params	What is fixed (reviewer notes)
`meta.json`	`seed`=0, `hops`=2, `seeds_per_split`=50,000	Defines the released *training-time induced subgraph* construction and the balanced split seed set size. In our bundle, these yield an induced subgraph with $\lvert V\rvert = 479{,}533$ and $\lvert\mathcal{E}\rvert = 1{,}864{,}213$ (directed).
`subset_node_idx.json` (probe set $\mathcal{S}$)	`seed`=0, `seeds_per_split`=50,000	Fixes the *probe node set* $\mathcal{S}$ (uniformly sampled from the induced subgraph, unless otherwise stated). Reusing $\mathcal{S}$ across depths/geometries/activations eliminates resampling variance in drift/separability probes.
`eval_edges_200k_sym.npy` ($\mathcal{E}_{\text{eval}}$)	`seed_edges`=0, `directed_handling`=`symmetrize`	Fixes the evaluation edge set for Dirichlet energy: sample 200,000 citation edges from the induced subgraph with a fixed seed, exclude self-loops, and symmetrize (Eq. (9)). Reused across all depths/geometries/activations to avoid resampling artifacts.
`fp_L*_*_*.csv`	`seed`=0, `seed_source=subset`; `seed_eval`=0	`seed` selects the fixed probe node IDs via `seed_source=subset`. `seed_eval` controls evaluation-time neighbor sampling (fanout) used when extracting representations for probes. Given fixed checkpoints, this makes drift/separability/radius diagnostics deterministic.
`drift_groups_L*_*_*.csv`	`seed`=0, `seed_source=subset`; `seed_eval`=0	Same sampling configuration as `fp_L*_*_*.csv`. Group assignment (e.g., degree bins) is deterministic given the induced subgraph and the fixed probe nodes.
`energy_L*_*_relu.csv`	`seed_edges`=0; `seed_eval`=0	`seed_edges` fixes $\mathcal{E}_{\text{eval}}$ (reused across conditions). `seed_eval` controls neighbor sampling used to extract layerwise representations before computing Eq. (8). Given fixed checkpoints and fixed $\mathcal{E}_{\text{eval}}$, energy curves are deterministic.

independently across depths ($|\mathcal{S}| = 1000$, 20 resamples) for a representative HypA configuration (ReLU, train seed 0), while keeping the evaluation-time neighbor-sampling seed fixed (`seed_eval`=0). Figure A.6 plots drift_{p90} at the final transition $\ell = L - 1$ as a function of depth L, showing near-identical curves across resamples. Quantitatively, the standard deviation over resamples was 0.088 at $L = 2$, 0.176 at $L = 4$, and below 3×10^{-3} for $L \geq 8$ (see also `jitter_by_depth.csv` in the experiment outputs). This supports that the depthwise transition observed in Figure 7 is not driven by probe-node resampling. In the main experiments, we nevertheless fix $\mathcal{S}$ and $\mathcal{E}_{\text{eval}}$ across all conditions to eliminate any such variance and ensure exact reproducibility from the released artifacts.

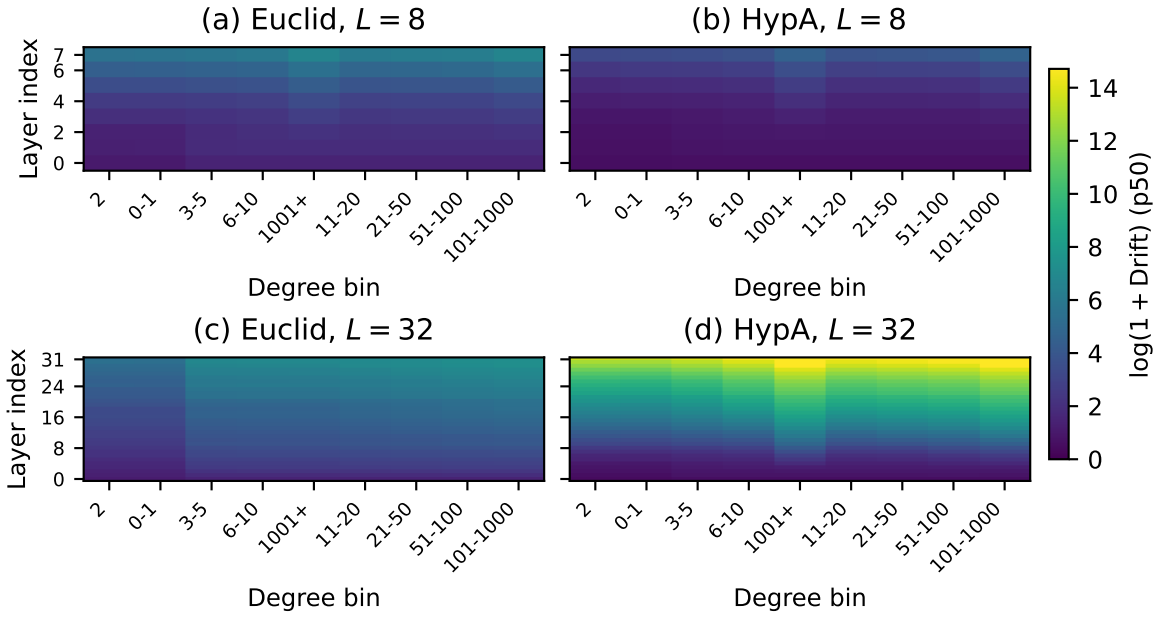

Figure A.2: Degree-resolved drift heatmaps (median, p50) for ReLU-ablated (ReLU→Identity) models: (a) Euclidean, $L = 8$, (b) Hyperbolic (HypA), $L = 8$, (c) Euclidean, $L = 32$, and (d) Hyperbolic (HypA), $L = 32$. Color scales are fixed within this figure and differ from that in Figure A.1 owing to scale differences induced by activation ablation.

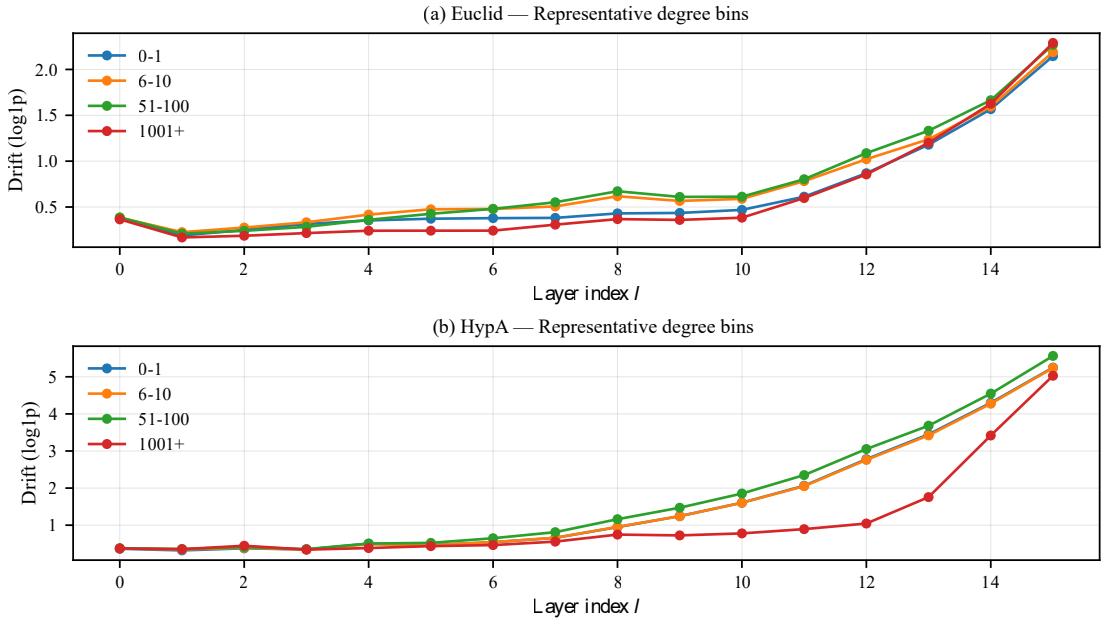

Figure A.3: Representative degree-bin drift trajectories at $L = 16$ (log1p scale): (a) Euclidean and (b) Hyperbolic (HypA) models. Each curve corresponds to a representative degree bin, illustrating distinct onset patterns of late-layer drift amplification.

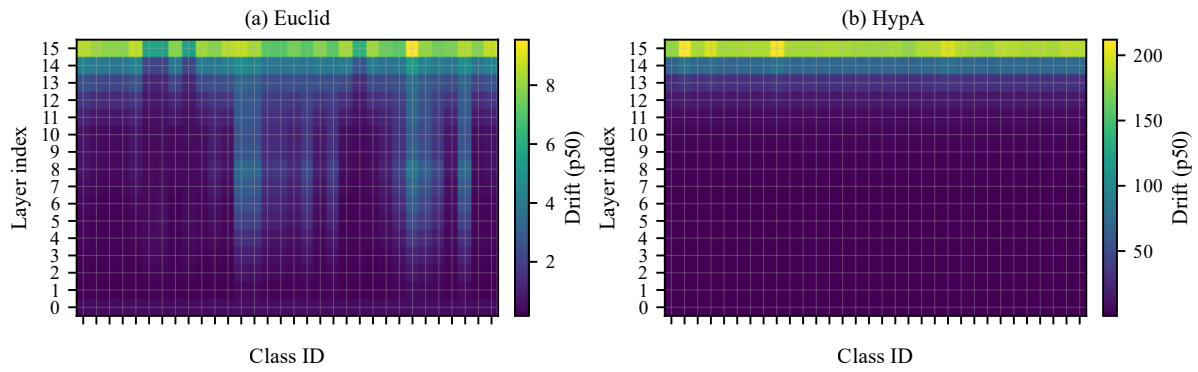

Figure A.4: Class-resolved drift heatmaps at $L = 16$. Each cell shows median per-node drift (p50) for a given class at layer ℓ. The x-axis shows class IDs as indices. The hyperbolic model exhibits broadly elevated late-layer drift across classes, consistent with a global amplification regime. Color scales are set independently for each panel to highlight within-class patterns rather than to make absolute magnitude comparisons across classes.

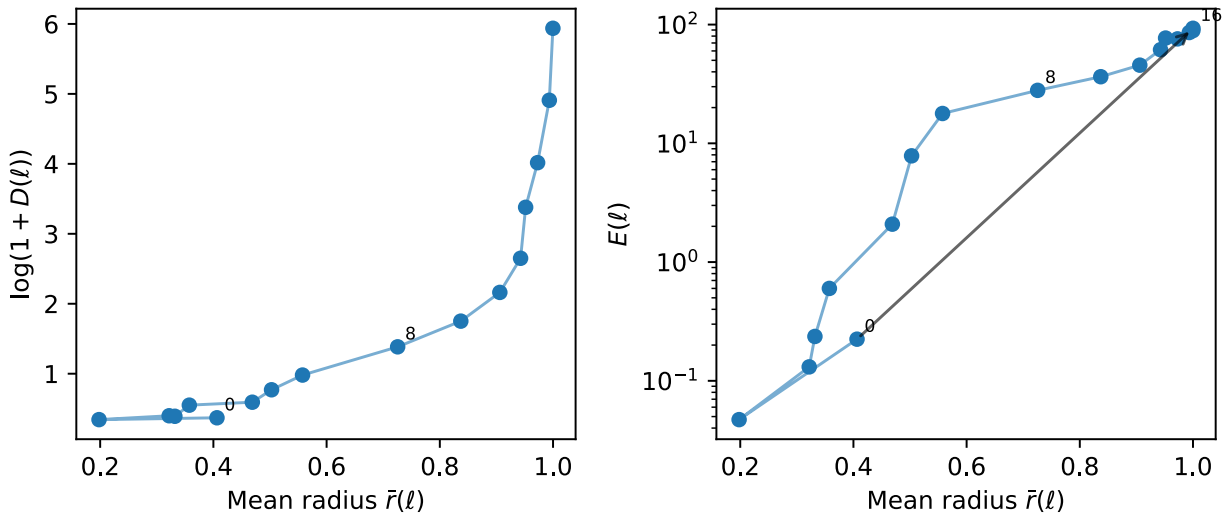

Figure A.5: Boundary association in the hyperbolic model at $L = 16$: (a) interlayer drift $D(\ell)$ and (b) Dirichlet energy $E(\ell)$ plotted against the mean normalized radius $\bar{r}(\ell)$. Both quantities exhibit amplification as representations approach the boundary of the Poincaré ball. Interpretation should be made jointly with boundary diagnostics and coordinate-step separation (Appendix D.5).

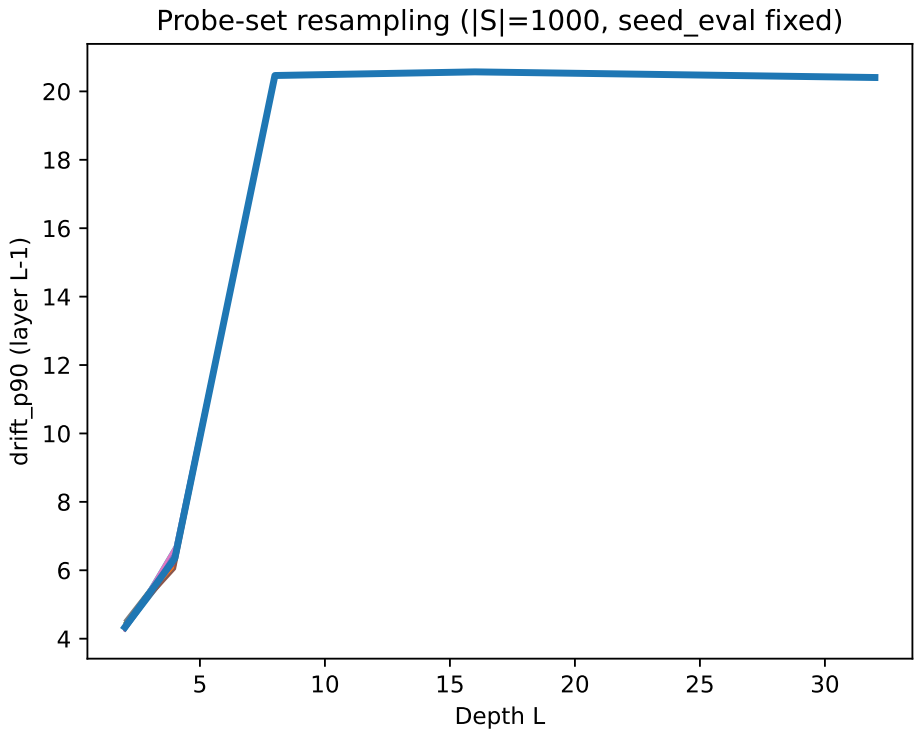

Figure A.6: Probe-set resampling ablation (HypA, ReLU, train seed 0; $|\mathcal{S}|$=1000; `seed_eval`=0). Each curve corresponds to one resample of $\mathcal{S}$ (resampled independently across depths).

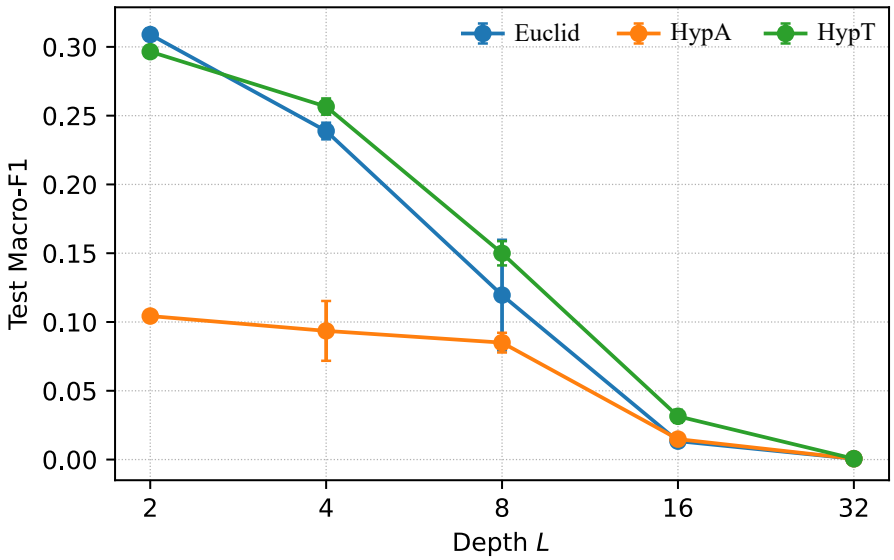

Figure B.1: Macro-F1 (test) versus depth L for Euclidean, HypA, and HypT GraphSAGE. Endpoint performance alone does not reveal the layerwise regime transitions exposed by fixed-point probing (see Figures 1 and 6). The deepest L=32 points should be read only as endpoint-collapse references.

Table B.1: Macro-F1 (test) versus depth for the same models as Figure B.1. Values report mean±std over the three training seeds (rounded). The L=32 entries are collapse-level endpoint references in this imbalanced many-class setting, and full per-seed values are included in the released CSV diagnostics (Appendix A.5).

Depth	Euclid	HypA	HypT
2	0.309±0.004	0.104±0.002	0.297±0.002
4	0.239±0.006	0.094±0.022	0.257±0.006
8	0.120±0.040	0.085±0.007	0.150±0.009
16	0.013±0.003	0.015±0.001	0.031±0.005
32	0.001±0.000	0.001±0.000	0.001±0.000

B Additional Depth and Decomposition Results

B.1 Macro-F1 versus depth

For reference, Figure B.1 summarizes the test macro-F1 as a function of depth for the Euclidean, hyperbolic (HypA), and tuned hyperbolic (HypT) GraphSAGE models (mean±std over the three training seeds). HypT uses a tangent-space classifier head with α_{in}=1.0, curvature c=3.0, and a learning-rate split (encoder multiplier 1.0, head multiplier 3.0; base learning rate 3×10^{-3}). At L=32, all three families reach collapse-level macro-F1 in the 10^{-3} range; under the strongly imbalanced 316-class setting, these values are consistent with near-single-class prediction and should be read only as endpoint-collapse references. Although the endpoint performance changes with depth, the probe signals in the main text reveal regime transitions in representation dynamics that cannot be reliably inferred from the endpoint performance alone. Full per-seed values are released with the CSV diagnostics to make this variance visible rather than hidden by rounding.

Table B.2: Spearman correlation between test macro-F1 and selected probe summaries across depth on the patent subgraph. Correlations were computed over $L \in \{2, 4, 8, 16, 32\}$ using the mean test macro-F1 over three training seeds. (We present n, the number of depth points used after excluding undefined probe values.) These correlations are descriptive only: each row uses at most five depth points and is not presented as inferential evidence.

Mode	Probe metric	n	Spearman ρ_s
Euclid	$D_{p50}(L-1)$	5	+0.700
Euclid	Sep. ratio	5	-0.700
Euclid	$\sqrt{c}\,\|z_L\|_{p50}$	0	–
Euclid	$\|h_L\|_{p50}$	5	+0.700
HypA	$D_{p50}(L-1)$	5	-0.600
HypA	Sep. ratio	5	+1.000
HypA	$\sqrt{c}\,\|z_L\|_{p50}$	5	-0.975
HypA	$\|h_L\|_{p50}$	5	-0.900
HypT	$D_{p50}(L-1)$	5	-0.600
HypT	Sep. ratio	4	+0.200
HypT	$\sqrt{c}\,\|z_L\|_{p50}$	5	-0.707
HypT	$\|h_L\|_{p50}$	5	-1.000

B.2 Patent-subgraph GCN backbone check

As an architecture robustness check, we repeated the patent-subgraph experiment with a Euclidean GCN backbone on the same patent subgraph under the same fixed-point probing protocol, reusing the same fixed probe-node set $\mathcal{S}$ and fixed evaluation-edge set $\mathcal{E}_{\text{eval}}$ as in the main experiments. Because this GCN variant is evaluated with neighbor-sampling loaders, its degree normalization is approximate on sampled subgraphs; we therefore treat the result as a diagnostic backbone check rather than an optimized benchmark comparison.

The summary plot is shown in Figure 8 in the main text. The key observation is that the qualitative probe–performance decoupling persists beyond GraphSAGE, but with an architecture-dependent internal profile: for L=16, both drift and metric-aware Dirichlet energy peak in mid-depth layers rather than only at the final transition. Table B.2 supplements that figure with depthwise probe–performance correlations computed on the same patent-subgraph sweep; we treat them as descriptive because the sweep contains only five depth points.

B.3 Extended Depth Sweeps

Additionally, we swept depths up to $L = 32$ and observed that the qualitative trends reported in the main text persist at larger depths. All layerwise probe traces (drift, Dirichlet energy, separability, and mean radius for hyperbolic models) are provided in the released CSV diagnostics (Appendix A.5).

B.4 Degree-Resolved Representation Dynamics

We further decomposed the representation drift by the node degree. Heatmaps in Figure A.1 show that higher-degree nodes exhibit earlier stabilization in Euclidean space, whereas hyperbolic models display degree-dependent late-layer amplification, which is consistent with the boundary effects.

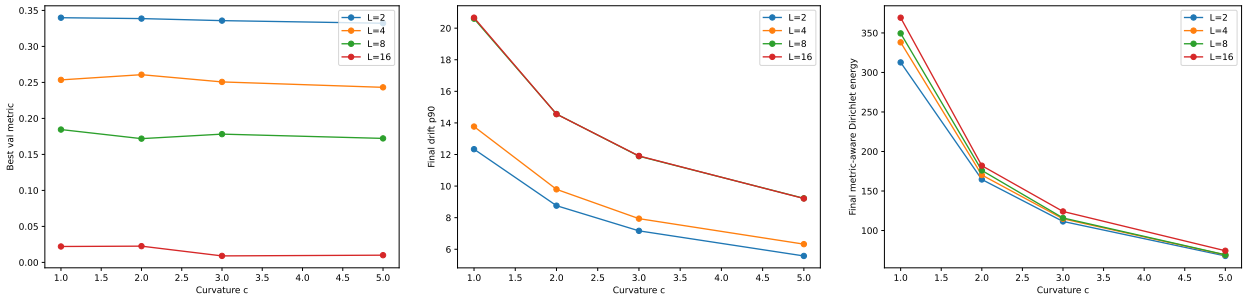

Figure B.2: **Curvature robustness sweep for HypT.** We swept $c \in \{1, 2, 3, 5\}$ across depths up to L=32. Left: best validation metric (reference only for model selection). Middle: final-transition drift $D_{\mathrm{p90}}(L-1)$. Right: last-layer metric-aware Dirichlet energy. Varying c changes absolute probe scales, but does not remove the qualitative very-deep regime transition.

B.5 Class-Resolved Dynamics

To assess whether the representation collapse is driven by specific classes, we analyzed the drift statistics conditioned on CPC subclasses. As shown in Figure A.4, no single class dominated observed amplification patterns, indicating that the effect is global rather than class-specific.

B.6 Alternative Separability Proxies

In addition to the Fisher ratio-based separability metric used in the main text, we evaluated cosine-based and kNN-based proxies. All the variants exhibited consistent qualitative trends, suggesting that the conclusions are insensitive to the definition of specific separability.

Dispersion ratio proxy for Euclidean homogenization. To complement separability, the main text already reports a separability-independent dispersion proxy in Figure 5. That proxy compares within-class and between-class pairwise distances (median over sampled pairs) and their ratio across layers. We do not repeat the same panel here. A decreasing between/within ratio indicates feature homogenization consistent with oversmoothing.

B.7 Curvature robustness sweep for HypT

A recurring concern in Euclid–hyperbolic comparisons is whether the observed very-deep degradation in hyperbolic models is an artifact of a particular curvature choice. To probe this, we swept the Poincaré-ball curvature $c \in \{1, 2, 3, 5\}$ for HypT across depths $L \in \{2, 4, 8, 16, 32\}$ (three training seeds). Figure B.2 summarizes (left) the best validation metric, shown only as a model-selection reference, (middle) the final-transition drift $D_{\mathrm{p90}}(L-1)$, and (right) the last-layer metric-aware Dirichlet energy. Across the tested curvatures, the qualitative very-deep regime persists (e.g., at L=32), while the absolute probe scales vary with c, supporting the view that the late-depth failure is not attributable to a single curvature setting.

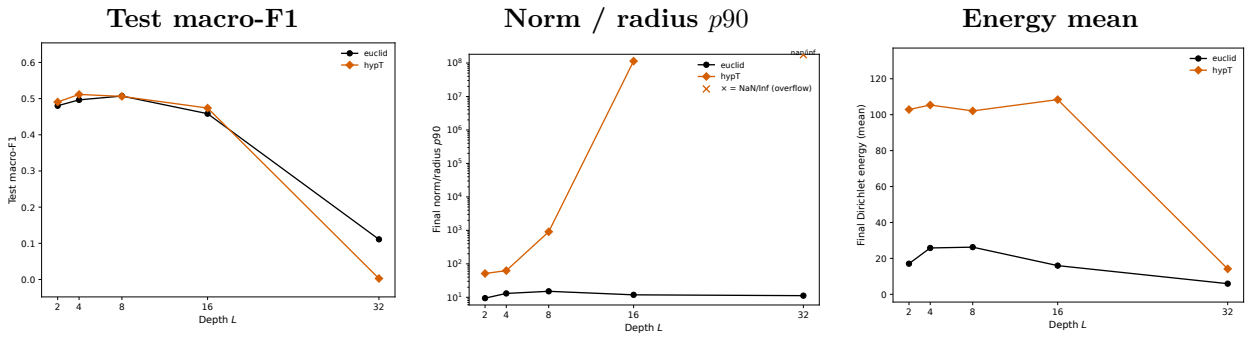

Figure C.1: **OGBN-Arxiv (GraphSAGE) reference diagnostics.** Left: test macro-F1 versus depth. Middle: final norm/radius $p90$ versus depth (log scale), where $\times$ marks non-finite values. Right: final metric-aware Dirichlet energy mean versus depth. Compared with the Euclidean model, the hyperbolic model exhibits a sharper late-depth failure and a much stronger rise in the norm/radius proxy before the final collapse.

C Public Benchmark Validation and Reference Results

This appendix reports the reference results for the public benchmark graphs (Cora and Texas). The purpose is not to analyze geometric mechanisms in detail or to optimize benchmark performance, but to verify that the qualitative probe—performance decoupling observed on the patent citation graph—also arises on standard datasets. All experimental settings and probe definitions were identical to those used in the main experiments unless otherwise stated.

C.1 Depth Sweeps and Layerwise Diagnostics

Figures C.7 and C.8 show class separability at the final layer as a function of depth. Across both benchmarks, separability exhibited a nonmonotonic dependence on depth, qualitatively mirroring the trends observed in the patent citation graph.

Figures C.9 and C.10 show layerwise interlayer drift and class separability at a fixed depth ($L = 16$). These probes illustrate that stable endpoint performance (or small performance changes) does not necessarily coincide with stable layerwise representation dynamics.

For completeness, we also present the metric-aware Dirichlet energy for Cora (Figure C.11) and show how the validation accuracy can obscure substantial variations in separability (Figure C.12). Additional OGBN-Arxiv reference checks are reported in Figures C.1 and C.2. We also report exploratory OGBN-Arxiv intervention pilots in Figure C.3 and tables C.1 to C.3, together with appendix sanity and robustness checks in Figures C.4 to C.6 and table C.4.

C.2 OGBN-Arxiv reference checks on GraphSAGE and GCN

This subsection reports additional public-benchmark checks on OGBN-Arxiv using the same fixed-point probing protocol. We include both the GraphSAGE depth sweep used in the main paper and a GCN backbone reference check. These figures are intended as diagnostic reference plots rather than optimized benchmark comparisons.

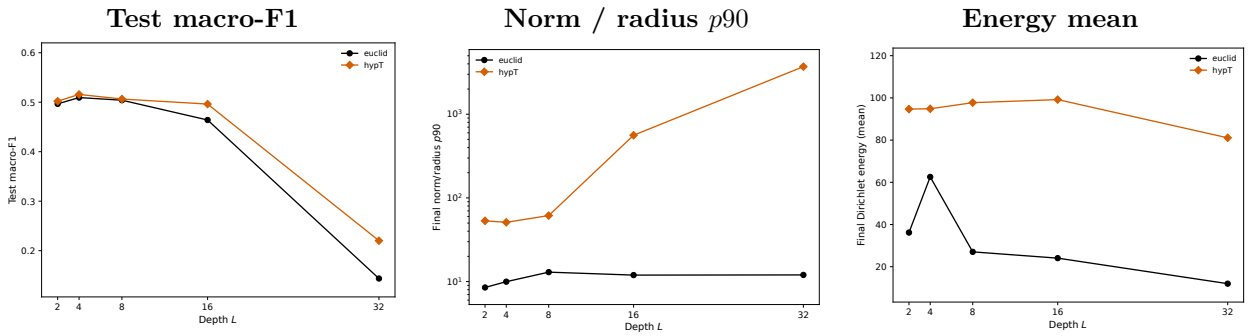

Figure C.2: **OGBN-Arxiv (GCN) reference diagnostics.** The same protocol applied to a GCN backbone provides an architecture check on the depth signatures. Left: test macro-F1 versus depth. Middle: final norm/radius $p90$ versus depth (log scale), with $\times$ indicating non-finite values. Right: final metric-aware Dirichlet energy mean versus depth.

Setting	α_{in}	c	tMLP	$r_{\text{p90}}^{\text{late}}$	test macro-F1
`a0.3_c3.0_tmlp0`	0.3	3.0	0	1.000	0.484
`a1.0_c3.0_tmlp0`	1.0	3.0	0	1.000	0.478
`a0.1_c1.0_tmlp0`	0.1	1.0	0	1.000	0.483
`a0.3_c1.0_tmlp0`	0.3	1.0	0	1.000	0.489
`a1.0_c1.0_tmlp0`	1.0	1.0	0	1.000	0.486

Table C.1: **OGBN-Arxiv L=16 HypT baseline candidates.** A small search over α_{in}, curvature, and tangent MLP width identified several *boundary-saturated but non-collapsed* configurations. All five top-ranked runs retained $r_{\text{p90}}^{\text{late}} \approx 1$ throughout the late window while keeping test macro-F1 in the 0.48–0.49 range. These candidates show that a boundary-pressure regime can be realized on OGBN-Arxiv at L=16 under the same fixed-point probing protocol.

C.3 OGBN-Arxiv intervention pilots on HypT

To probe whether the boundary-control interpretation transfers beyond the patent graph, we ran a small set of exploratory radius-penalty pilots on OGBN-Arxiv with HypT. Because an exploratory L=32 search frequently produced an *origin-side collapse* before the intended late-only window, we concentrated the most interpretable controlled comparisons on L=16, where a small baseline search identified boundary-saturated yet non-collapsed configurations. In this subsection, $r_{\text{p90}}^{\text{late}}$ denotes the mean of the layerwise upper-tail radius summary $r_{\text{p90}}(\ell)$ over the designated late-layer window, and $\text{frac}_{\text{sat,last}}$ denotes the corresponding last-layer saturation ratio.

For the intervention pilots, we selected two representative L=16 baselines that span a more performance-oriented operating point (`performance`: `a0.3_c3.0_tmlp0`) and a more stable operating point (`stable`: `a1.0_c1.0_tmlp0`). Figure C.3 summarizes the OGBN-Arxiv L=16 feasibility and intervention outcomes at a glance, while Tables C.2 and C.3 report the corresponding numeric details for the initial three-seed late-only pilot and the always-on seed-0 sweep.

We then made the schedule control explicit by repeating late-only sweeps with start fractions 0.75, 0.60, and 0.50. Across these runs, $r_{\text{p90}}^{\text{late}}$ and $\text{frac}_{\text{sat,last}}$ remained at 1.0 in all tested seeds and conditions, even though the training logs showed that the late-only switch became active and the radius-penalty term became nonzero once the scheduled start epoch was reached. In

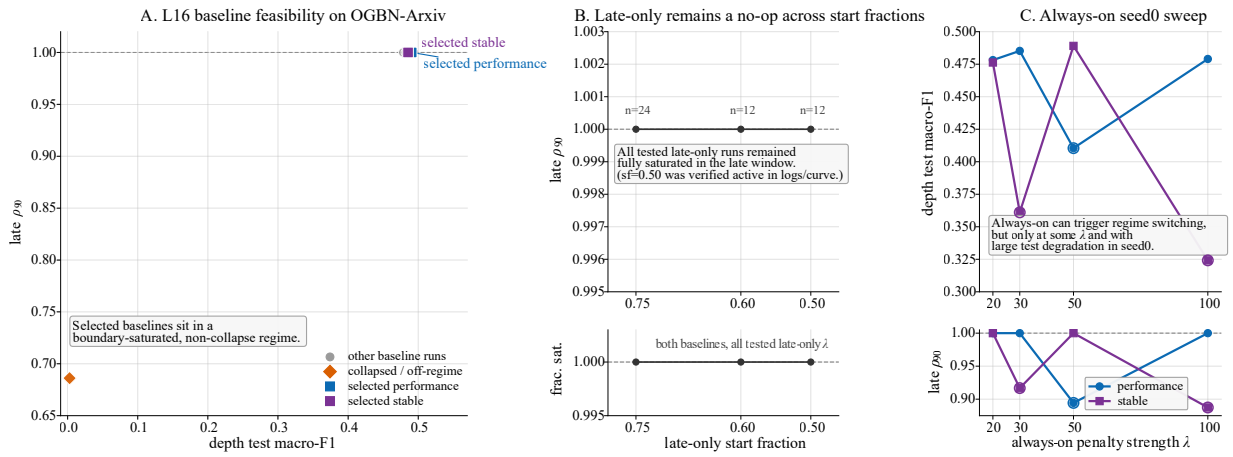

Figure C.3: **Three-panel summary of OGBN-Arxiv L=16 intervention pilots on representative HypT baselines.** (A) Baseline feasibility: the selected `performance` and `stable` settings lie in a *boundary-saturated but non-collapsed* regime at L=16. (B) Late-only radius penalties remain effectively no-op across start fractions 0.75, 0.60, and 0.50: the boundary proxies stay saturated even when the scheduled penalty becomes active in the training logs. (C) The always-on seed-0 sweep over $\lambda \in \{20, 30, 50, 100\}$ is strongly nonmonotonic: some λ values move the boundary proxies, but only with a substantial test macro-F1 cost, whereas adjacent λ values are complete no-ops. Taken together, the OGBN-Arxiv pilots support timing-sensitive and basin-like regime switching rather than a smooth, performance-preserving control band.

Label	Condition	test macro-F1	$r_{\mathrm{p90}}^{\mathrm{late}}$	$\mathrm{frac}_{\mathrm{sat,last}}$
performance	$\lambda = 0$	0.483 ± 0.006	1.000	1.000
performance	$\lambda = 10$ late-only	0.487 ± 0.006	1.000	1.000
performance	$\lambda = 100$ always-on	0.437 ± 0.079	0.956	0.672
stable	$\lambda = 0$	0.480 ± 0.005	1.000	1.000
stable	$\lambda = 10$ late-only	0.481 ± 0.007	1.000	1.000
stable	$\lambda = 100$ always-on	0.457 ± 0.050	0.958	0.674

Table C.2: **Initial OGBN-Arxiv L=16 intervention pilot on two representative HypT baselines (three seeds).** The weak late-only control is effectively a no-op: both $r_{\mathrm{p90}}^{\mathrm{late}}$ and the last-layer saturation ratio remain at 1.0 while endpoint performance stays near the λ=0 baseline. The stronger always-on control partially reduces the boundary proxies, but it also increases variance and degrades endpoint performance. These runs are therefore consistent with a timing-sensitive intervention effect, but not yet with a performance-preserving public control band.

other words, under the tested public L=16 settings, applying the same boundary penalty from the last 50% of training was still effectively a no-op.

Finally, we ran an always-on seed-0 sweep over $\lambda \in \{20, 30, 50, 100\}$ to test whether the OGBN-Arxiv response is smooth in the intervention strength. The outcome was strongly nonmonotonic (Table C.3): some λ values moved the boundary proxies substantially, but only at the cost of a large drop in test macro-F1, while adjacent λ values were complete no-ops.

Taken together, these OGBN-Arxiv pilots support three points that are consistent with the main patent-subgraph narrative while remaining appropriately cautious. First, public benchmark hyperbolic models can also realize a boundary-saturated yet non-collapsed regime. Second, once such a regime is established, late-only boundary control can be genuinely active

Label	λ	Δ test	$r_{\mathrm{p90}}^{\mathrm{late}}$	$\mathrm{frac}_{\mathrm{sat,last}}$	Interpretation
performance	20	+0.000	1.000	1.000	no-op
performance	30	+0.007	1.000	1.000	no-op
performance	50	−0.067	0.894	0.026	moved / costly
performance	100	+0.001	1.000	1.000	no-op
stable	20	+0.000	1.000	1.000	no-op
stable	30	−0.115	0.917	0.063	moved / costly
stable	50	+0.013	1.000	1.000	no-op
stable	100	−0.152	0.887	0.026	moved / costly

Table C.3: **Always-on OGBN-Arxiv L=16 seed-0 sweep relative to the corresponding λ=0 baseline.** The OGBN-Arxiv intervention response is highly nonmonotonic. For the performance-oriented baseline, only λ=50 moved the boundary proxies, and it did so with a marked loss in test macro-F1. For the stable baseline, λ=30 and λ=100 both moved the proxies, again with substantial endpoint degradation. We therefore interpret these pilots as supportive diagnostic evidence for basin-like regime switching, not as evidence for a smooth or dataset-universal intervention threshold.

yet still fail to move the representation trajectory. Third, always-on controls can induce regime switches, but the resulting response is strongly timing-sensitive and nonmonotonic under the tested HypT settings.

C.3.1 Boundary-intervention sanity checks and robustness at L=16

We performed a focused set of sanity and robustness checks to evaluate whether the OGBN-Arxiv L=16 boundary-saturated but non-collapsed regime can be reliably controlled by the same radius-based penalty family. We considered the same two representative HypT baselines used in the OGBN-Arxiv intervention pilot: a performance-oriented setting and a stability-oriented setting. For late-only interventions, we swept the penalty onset and confirmed from the training logs and curve files that the regularizer was implemented correctly: the late-only switch became active at the specified epoch and the radius-penalty term became nonzero in the objective. Despite this correct activation, the boundary proxies stayed saturated throughout training, with r_{p90} and $\mathrm{frac}_{\mathrm{sat}}$ remaining essentially at 1.0. This shows that late-only control is genuinely active yet still ineffective in moving the trained system away from the boundary-saturated regime.

We then examined cross-seed robustness for the always-on controls. The initial seed-0 sweep suggested several nonmonotonic moved cases, but these did not generalize across seeds. When the informative always-on settings were rerun for seeds 1 and 2, all of them behaved as no-ops with saturated boundary proxies and near-baseline test performance. We therefore interpret the initial seed-0 pattern not as evidence of a robust public control band, but rather as seed-conditioned fragility. Figure C.5 summarizes this contrast: late-only conditions remain pinned to the saturated boundary level across all tested start fractions, while the moved always-on cases are confined to the original seed-0 exploratory sweep.

To probe the seed-0 anomalies more directly, we repeated the anomalous conditions multiple times. The performance-oriented `lam50` case did not reproduce the earlier moved behavior, and the stability-oriented `lam30` case moved only rarely. By contrast, the stability-oriented `lam100` case consistently reduced the boundary proxies in all repeats, but this came with a marked drop in validation/test macro-F1. Table C.4 collects the family-level summary, while Figure C.6 shows the repeated stable-`lam100` trajectories directly. A more detailed row-wise

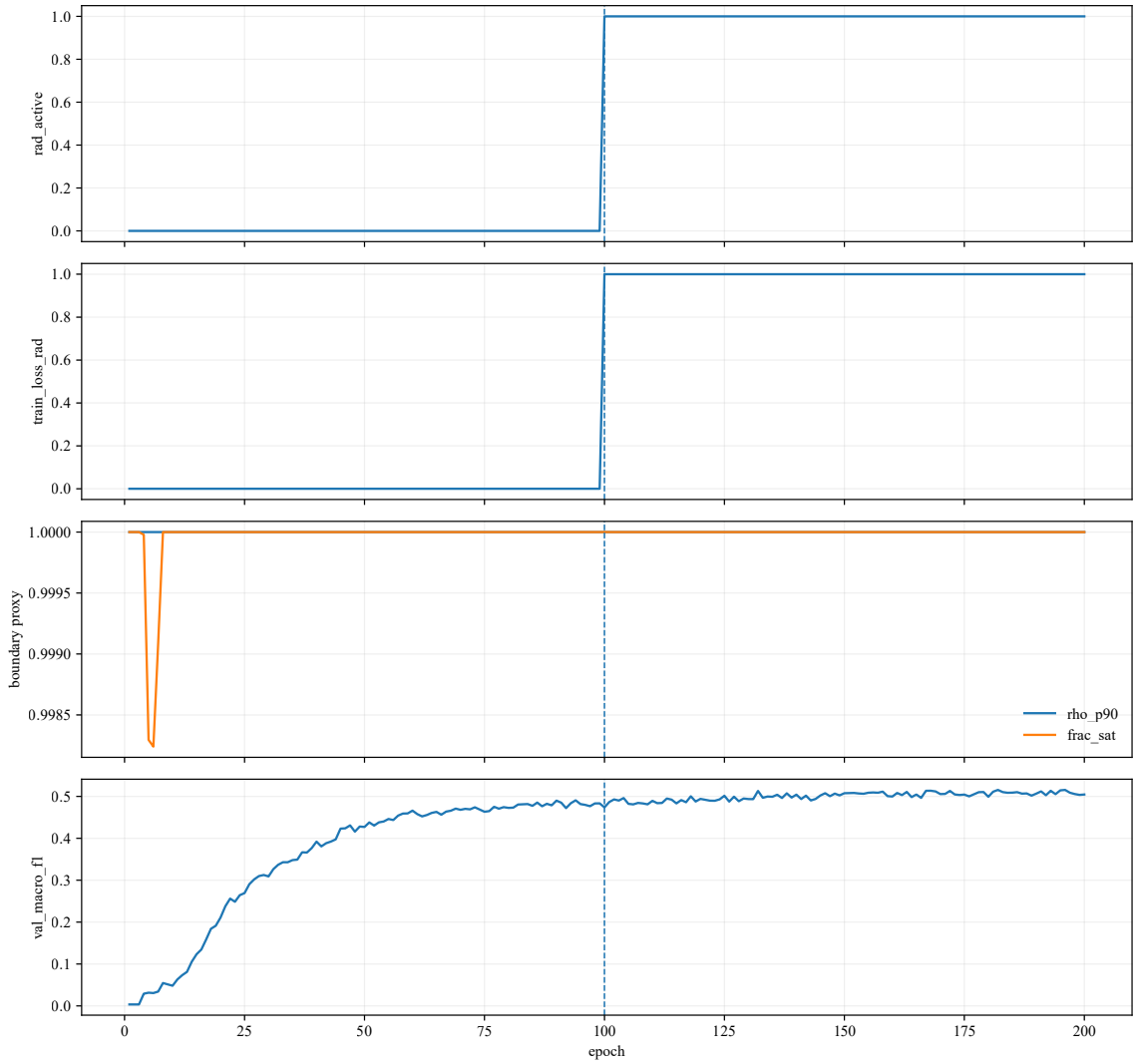

Figure C.4: **Late-only penalty activates, but boundary proxies do not move.** Shown is a representative late-only run at depth 16 (`sf=0.5`, `lam100`). The regularizer activates at the prescribed epoch, as indicated by the jump in `rad_active` and the nonzero `train_loss_rad`. However, both boundary proxies remain saturated ($r_{\mathrm{p90}} \approx 1$, $\mathrm{frac}_{\mathrm{sat}} \approx 1$), and validation macro-F1 continues along its usual trajectory. This confirms that the late-only intervention is implemented correctly, but ineffective in shifting the model away from the boundary-saturated regime.

CSV summary is released with the artifact bundle; in the paper appendix we show only the compact family-level table to keep the narrative readable.

Taken together, these checks support two conclusions. First, the OGBN-Arxiv L=16 boundary-saturated regime is hard to perturb once it has formed: late-only penalties remain ineffective even when they are demonstrably active. Second, while always-on penalties can trigger a regime switch in some cases, the response is fragile across seeds and, in the one condition that reproduces robustly, does not preserve predictive performance. We therefore use these appendix experiments as supporting evidence for the hardness and fragility of the regime, rather than as a positive intervention result.

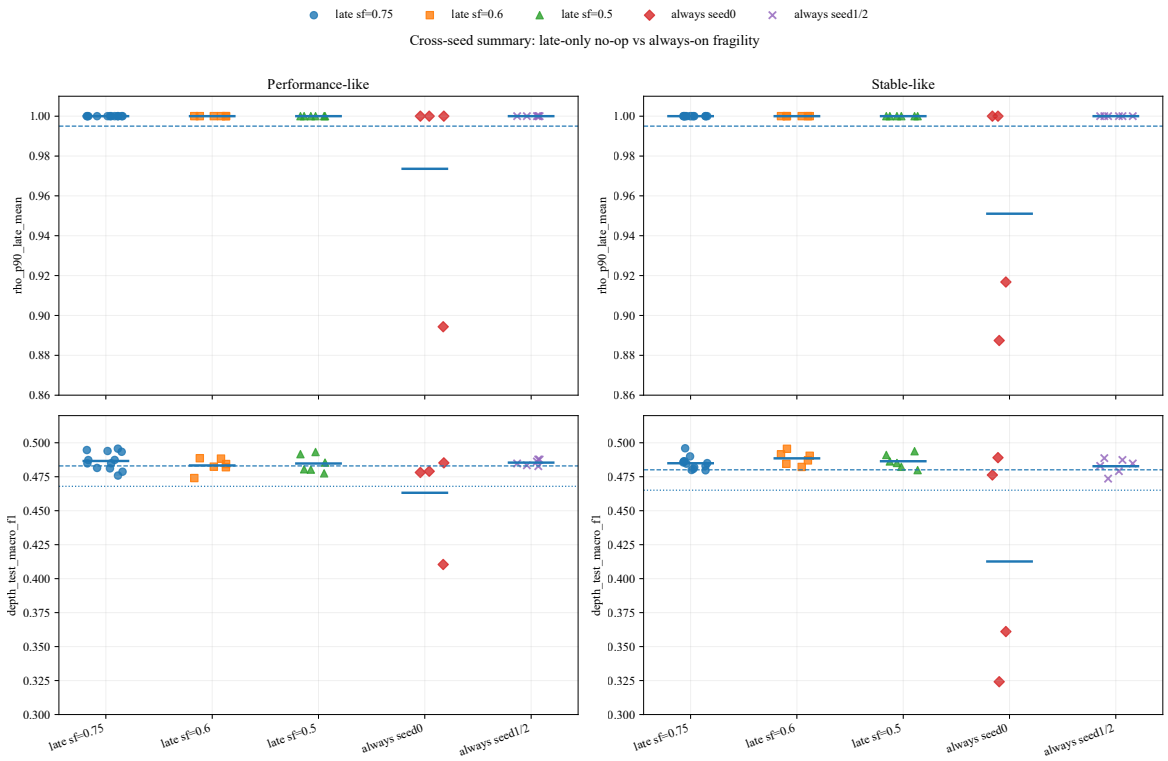

Figure C.5: **Late-only interventions are robustly inactive, whereas always-on responses are seed-fragile.** Cross-seed summary of boundary proxy ($r^{\text{late}}_{\text{p90}}$) and test macro-F1 for the late-only sweeps (`sf=0.75`, `0.6`, `0.5`) and the always-on sweeps. All late-only conditions remain at the saturated boundary level across seeds. The moved cases seen in the initial always-on seed-0 sweep do not reproduce for seeds 1 and 2, indicating that the apparent response is not a robust cross-seed control effect.

Family / setting	n	moved	$r^{\text{late}}_{\text{p90}}$	$\text{frac}_{\text{sat,last}}$	Δ test vs. λ=0
late-only, sf= 0.75	24	0/24	1.000	1.000	+0.004
late-only, sf= 0.60	12	0/12	1.000	1.000	+0.004
late-only, sf= 0.50	12	0/12	1.000	1.000	+0.004
always-on, seeds 1–2 ($\lambda \in \{30, 50, 100\}$)	12	0/12	1.000	1.000	+0.003
repeat, performance-like `lam50`	8	0/8	1.000	1.000	−0.001
repeat, stable-like `lam30`	8	1/8	0.990	0.883	−0.012
repeat, stable-like `lam100`	8	8/8	0.889	0.022	−0.098

Table C.4: **Summary of boundary-intervention robustness at depth** 16. For each intervention family, we report the number of moved runs, the late-window boundary proxy, the last-layer saturation ratio, and the mean test macro-F1 change relative to the matched λ=0 baseline. Late-only settings never produce a moved run, informative always-on settings do not reproduce across seeds 1 and 2, and only the repeated stable-like `lam100` condition shows a consistent regime switch, albeit with a clear performance drop.

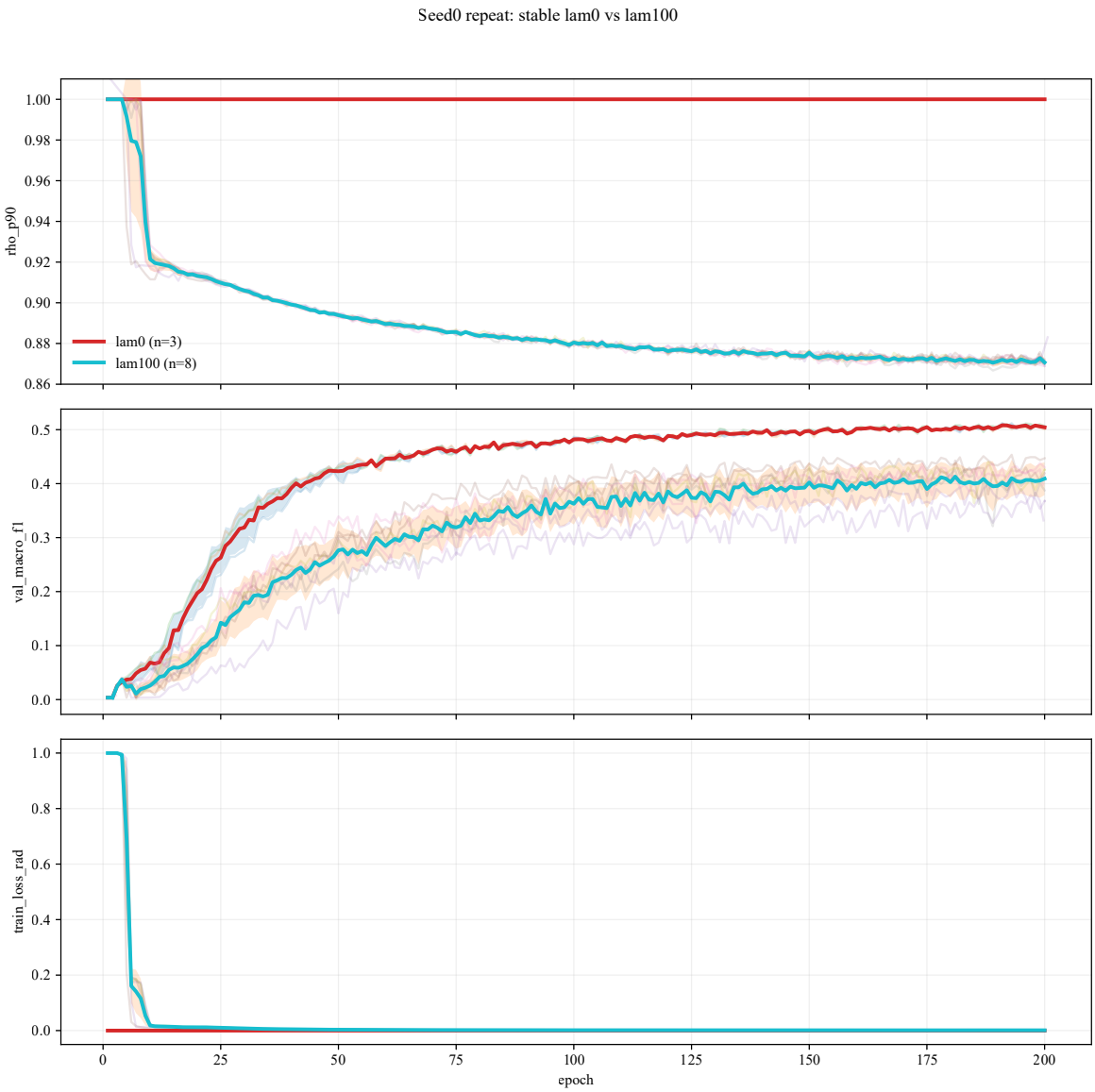

Figure C.6: **In seed-0 repeats, stable `lam100` consistently lowers the boundary proxy but degrades performance.** Mean trajectories with standard-deviation bands for repeated seed-0 runs comparing stable-like `lam0` and stable-like `lam100`. The `lam100` condition consistently drives r_{p90} downward and keeps the RAD loss active, whereas `lam0` stays boundary-saturated throughout. However, this boundary reduction comes with substantially lower validation macro-F1, showing that the repeatable regime switch does not preserve predictive performance.

C.4 Summary

In summary, these public benchmark results provide reference evidence that fixed-point probing can surface depthwise representation changes that are not consistently reflected in endpoint performance, even on standard benchmark datasets. The additional OGBN-Arxiv reference checks and intervention pilots further indicate that similar qualitative depth signatures are visible beyond the patent graph, while also showing that boundary control on the public benchmark can be strongly timing-sensitive, seed-fragile, and difficult to reconcile with performance retention under the tested HypT settings.

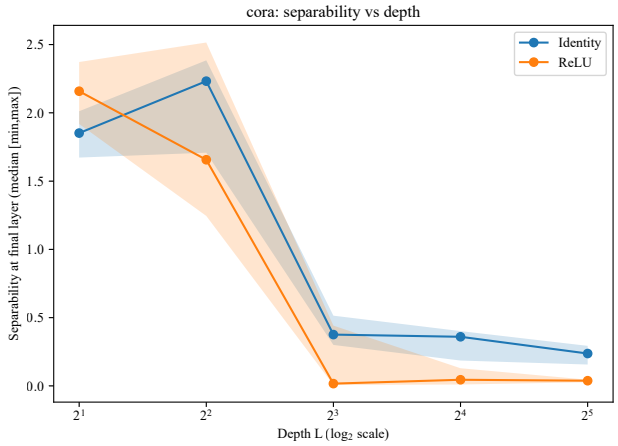

Figure C.7: Class separability at the final layer as a function of depth on the Cora citation network. Depth is shown on a $\log_2$ scale. Shaded area shows the min–max range across random seeds and data splits.

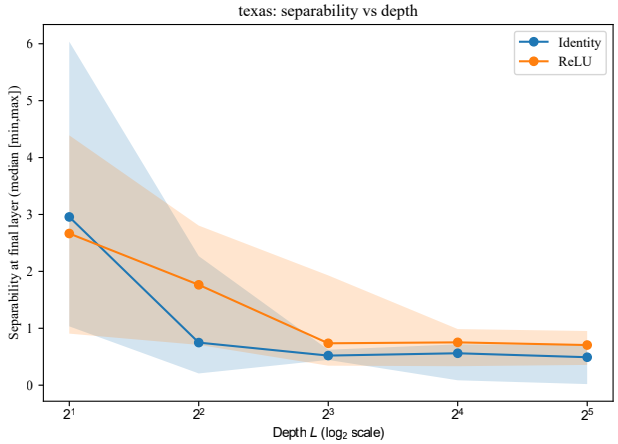

Figure C.8: Class separability at the final layer as a function of depth on the Texas web graph. Shaded area shows the min–max range across random seeds and data splits.

Table C.5: Test accuracy versus depth on the Cora citation network. Values show the median [min, max] across random seeds and data splits (reported for reference only).

depth	Identity	ReLU
2	0.7180 [0.7010, 0.7450]	0.7130 [0.6890, 0.7550]
4	0.7070 [0.5790, 0.7510]	0.4900 [0.3700, 0.5600]
8	0.1440 [0.0910, 0.1490]	0.3190 [0.1030, 0.3190]
16	0.1300 [0.0640, 0.3190]	0.3190 [0.1300, 0.3190]
32	0.1300 [0.0910, 0.3190]	0.3190 [0.1440, 0.3190]

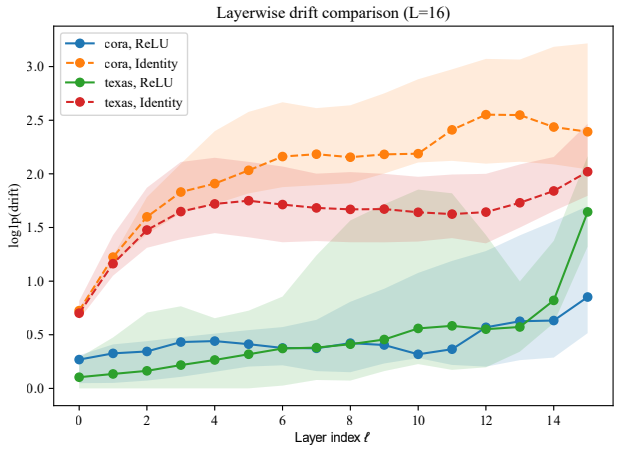

Figure C.9: Layerwise interlayer drift at fixed depth $L = 16$ on public benchmarks. Curves compare ReLU and identity activations on the same axes. Drift is shown using $\mathrm{log1p}(x) = \log(1 + x)$. Shaded area shows the min–max range across random seeds and data splits.

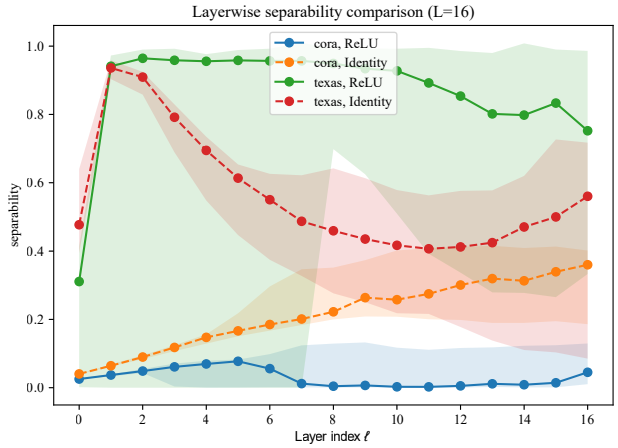

Figure C.10: Layerwise class separability at fixed depth $L = 16$ on public benchmarks. Shaded area shows the min–max range across random seeds and data splits.

Table C.6: Test macro-F1 versus depth on the Cora citation network. Values show the median [min, max] across random seeds and data splits (reported for reference only).

depth	Identity	ReLU
2	0.7071 [0.7048, 0.7318]	0.7061 [0.6996, 0.7495]
4	0.6832 [0.4936, 0.7381]	0.4210 [0.2345, 0.5465]
8	0.0360 [0.0238, 0.0371]	0.0691 [0.0267, 0.0691]
16	0.0329 [0.0172, 0.0691]	0.0691 [0.0329, 0.0691]
32	0.0329 [0.0238, 0.0987]	0.0691 [0.0360, 0.0691]

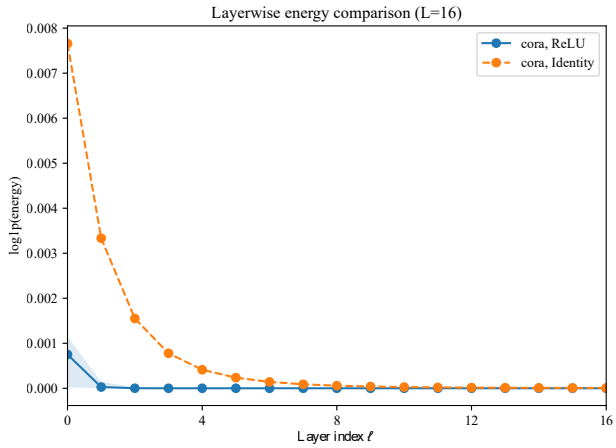

Figure C.11: Layerwise metric-aware Dirichlet energy at fixed depth $L = 16$ on the Cora dataset. Energy is shown using $\log1p(x)$ for numerical stability. Shaded area shows the min–max range across random seeds and data splits.

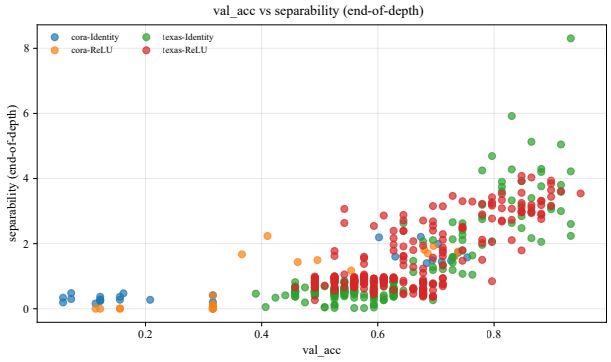

Figure C.12: Validation accuracy versus class separability on public benchmarks. Despite similar accuracy values, separability varies substantially, illustrating that endpoint performance alone does not reliably reflect internal representation quality.

Table C.7: Test accuracy versus depth on the Texas web graph. Values show the median [min, max] across random seeds and data splits (reported for reference only).

depth	Identity	ReLU
2	0.8108 [0.5946, 0.9189]	0.7838 [0.5676, 0.8919]
4	0.6216 [0.4595, 0.7838]	0.6757 [0.5676, 0.8108]
8	0.5946 [0.4865, 0.6486]	0.6216 [0.5405, 0.7297]
16	0.5946 [0.4865, 0.6486]	0.6216 [0.4865, 0.7297]
32	0.5946 [0.4324, 0.6486]	0.6216 [0.5135, 0.7297]

Table C.8: Test macro-F1 versus depth on the Texas web graph. Values show the median [min, max] across random seeds and data splits (reported for reference only).

depth	Identity	ReLU
2	0.6726 [0.3212, 0.8157]	0.6255 [0.2615, 0.8605]
4	0.2703 [0.1309, 0.6846]	0.4262 [0.2615, 0.7560]
8	0.1864 [0.1309, 0.2652]	0.2914 [0.1533, 0.4019]
16	0.1826 [0.1309, 0.3764]	0.2836 [0.1533, 0.4037]
32	0.1864 [0.1309, 0.2981]	0.2890 [0.1533, 0.4212]

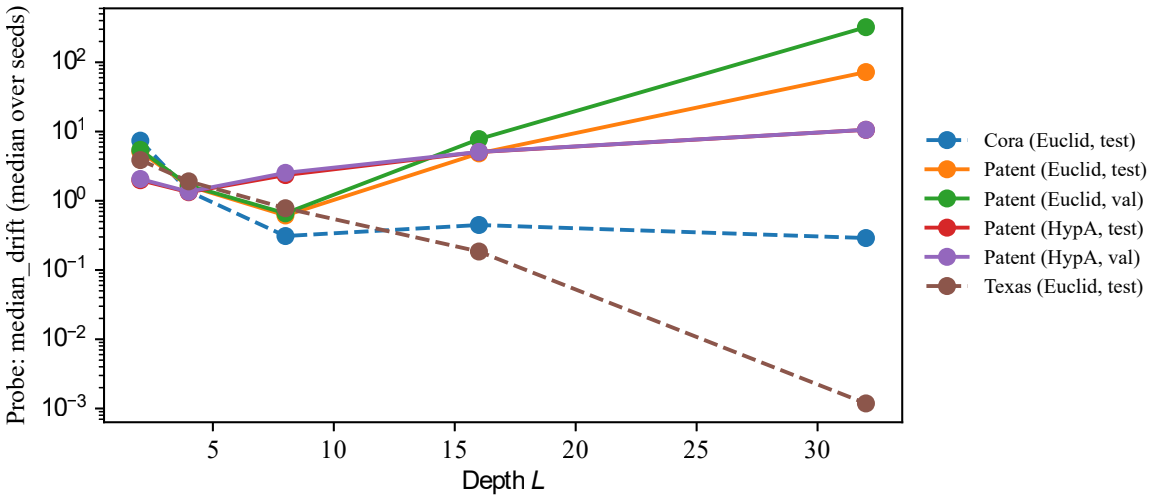

Figure C.13: Median interlayer drift as a function of depth, evaluated on validation/test splits, where available. Curves show the median over random seeds. This plot is included as a compact reference for split-wise probe behavior across datasets.

D Technical details and probe interpretation

Dirichlet energy evaluation protocol. For the representation $z^{(\ell)}$, we computed the Dirichlet energy

$$E^{(\ell)} = \frac{1}{|E_{\text{eval}}|} \sum_{(u,v)\in E_{\text{eval}}} d\Big(z_u^{(\ell)}, z_v^{(\ell)}\Big)^2,$$

where d denotes the geometry-specific distance (Euclidean for EUCLID and Poincaré for HYPA/HYPT). In the patent subgraph, we evaluated $E^{(\ell)}$ on a fixed set of $|E_{\text{eval}}| = 200{,}000$ edges (sampled once and symmetrized) and reported the *last layer* value $E^{(L)}$. Unless otherwise stated, the results are aggregated into three training seeds (mean±std).

Reporting conventions and endpoint metrics. In Appendix D, we present multiple quantities for completeness. Unless otherwise stated, the *accuracy* values (val/test) are provided for reference only; *macro-F1* is the primary endpoint performance metric used and discussed in the main text, which is consistent with the class-imbalanced setting of the patent citation graph.

The per-seed results are summarized as the mean±std over three training seeds. All underlying per-seed probe values and endpoint metrics are included in the released CSV diagnostics (Appendix A.5), combined with scripts that deterministically regenerate every reported plot and table. We have omitted additional seed-wise tables to avoid redundancy because rounding in compact tables can obscure small but real differences.

D.1 Operational definition of fixed-point-like plateaus

Definition (Operational fixed-point-like plateau). In this study, we adopt the *operational notion of fixed-point-like (near-stationary) plateau behavior* defined in the observed representation space $\mathbf{z}^{(\ell)}$ (Eq. (2)), rather than in internal latent coordinates. A layer regime exhibits fixed-point-like dynamics when the interlayer drift summary (e.g., $D_{\text{mean}}(\ell)$ or $D_{\text{p50}}(\ell)$ in Eqs. (4)–(5)) decreases and remains small over consecutive layers. In the main text, we present $D_{\text{p50}}(\ell)$ unless otherwise noted. This definition identifies plateaus in the depthwise representation dynamics and does not imply convergence to a true attractor or the existence of a dynamical-systems fixed-point theorem for the learned update maps.

For auditability, when an explicit thresholded rule is needed, we call a block of k consecutive layers plateau-like if

$$D_{\text{p50}}(\ell) < \tau_D \qquad \text{for all } \ell \in \{\ell_0, \ldots, \ell_0 + k - 1\},$$

where k is fixed in advance and τ_D is a low-drift threshold chosen *a priori* within an experiment family. In our terminology, we use $k = 3$ as a minimal consecutive-layer requirement. Because probe scales differ across geometries and curvature settings, we do not promote a single universal numerical threshold across all experiments. This thresholded rule is included only to make the operational terminology auditable; the reported analyses rely on the full drift curves and on joint probe couplings rather than on thresholded plateau counts.

This definition is canonical and is referenced throughout the main text and appendices. In hyperbolic space, drift should be interpreted jointly with boundary indicators and coordinate-

level diagnostics (Section D.5), because geodesic distances may be amplified near the Poincaré boundary.

D.2 Mean computation in hyperbolic space

For the hyperbolic models, the class prototypes were computed using a tangent-space mean at the origin of the Poincaré ball. Specifically, the embeddings were first mapped to the tangent space using the Riemannian logarithmic map $\log_0(\cdot)$, averaged in the Euclidean space, and mapped back using the exponential map $\exp_0(\cdot)$. This choice provides numerical stability and scalability while yielding behavior qualitatively similar to the Fréchet mean in our setting. Because this approximation is anchored at the origin, we do not present it as an origin-invariant geometric summary; its role here is operational and comparative, with the same construction reused across depths and model variants within the protocol.

D.3 Metric-aware Dirichlet energy

We employed the metric-aware Dirichlet energy to quantify the graph-local representation roughness. This formulation generalizes the classical Euclidean Dirichlet energy by replacing squared Euclidean distances with squared distances under the representation metric $d(\cdot, \cdot)$. Consequently, smoothness is measured in a geometry-consistent manner and is naturally connected to Laplacian-based analyses of oversmoothing when instantiated in the Euclidean space.

D.4 Poincaré ball geometry and distance

We summarize the Poincaré ball geometry used throughout this study and fix the notation for the hyperbolic distance d_{Hyp} referenced in the main text.

Poincaré ball. The d-dimensional Poincaré ball is

$$\mathbb{B}^d = \{u \in \mathbb{R}^d : \|u\|_2 < 1\},$$

equipped with a Riemannian metric with constant negative curvature. Unless otherwise stated, we used curvature $c = 1.0$ (unit ball).

General curvature. Generally, for a curvature $c > 0$, the d-dimensional Poincaré ball is defined as

$$\mathbb{B}_c^d = \{u \in \mathbb{R}^d : \|u\|_2 < 1/\sqrt{c}\},$$

with a constant negative curvature $-c$. The corresponding geodesic distance is expressed by

$$d_{\mathbb{B}_c}(u, v) = \frac{1}{\sqrt{c}} \operatorname{arcosh}\left(1 + 2c \frac{\|u - v\|_2^2}{(1 - c\|u\|_2^2)(1 - c\|v\|_2^2)}\right). \tag{19}$$

Throughout the main text, we present expressions in unit-ball form ($c = 1$) for readability. All the hyperbolic probes were evaluated using the appropriate curvature for each model (e.g., $c = 1$ for HypA and $c = 3$ for HypT; see Table A.3). The boundary proximity is reported using the normalized radius $\sqrt{c}\,\|z\|$, such that the values approaching 1 consistently indicate saturation across the curvatures.

Geodesic distance. For $u, v \in \mathbb{B}^d$, the geodesic distance is expressed in closed form as follows:

$$d_{\mathbb{B}}(u, v) = \operatorname{arcosh}\left(1 + 2\frac{\|u - v\|_2^2}{(1 - \|u\|_2^2)(1 - \|v\|_2^2)}\right). \tag{20}$$

Throughout the paper, we denote the Poincaré ball geodesic distance by

$$d_{\mathrm{Hyp}}(u, v) := d_{\mathbb{B}}(u, v). \tag{21}$$

Exponential and logarithmic maps at the origin. Let $0 \in \mathbb{B}^d$ be the origin. Riemannian exponential and logarithmic maps at the origin are

$$\exp_0(v) = \tanh(\|v\|_2)\frac{v}{\|v\|_2}, \qquad \log_0(u) = \operatorname{artanh}(\|u\|_2)\frac{u}{\|u\|_2}, \tag{22}$$

with the conventions $\exp_0(0) = 0$ and $\log_0(0) = 0$.

Relation to observed representations. In hyperbolic GNN models, internal activations are parameterized in the Euclidean tangent space and mapped to the Poincaré ball using an exponential map of the origin. As described in Section 4, we defined the observed representations $\mathbf{z}_i^{(\ell)} = \exp_0(\tilde{\mathbf{h}}_i^{(\ell)}) \in \mathbb{B}^d$ and evaluated all geometric probes, including drift, Dirichlet energy, separability, and boundary monitoring, on $\{\mathbf{z}_i^{(\ell)}\}$ using d_{Hyp}.

D.5 Boundary saturation and metric amplification

Boundary monitoring. To monitor boundary effects, we track the curvature-normalized radius

$$r_i^{(\ell)} := \sqrt{c}\,\|\mathbf{z}_i^{(\ell)}\|_2 \in [0, 1),$$

which provides a curvature-comparable notion of boundary proximity. We summarize it by the mean normalized radius

$$\bar{r}(\ell) := \frac{1}{|\mathcal{S}|}\sum_{i \in \mathcal{S}} r_i^{(\ell)},$$

and, where relevant, by upper-tail summaries such as

$$r_{\mathrm{p90}}(\ell) := \mathrm{p90}_{i \in \mathcal{S}}\, r_i^{(\ell)},$$

together with the boundary occupancy ratio from Section 4.4. Figure D.1 shows the median normalized radius

$$r_{\mathrm{p50}}(L) := \mathrm{p50}_{i \in \mathcal{S}}\, r_i^{(L)}$$

at the final layer as a function of trained depth for HypA ($c = 1$) and HypT ($c = 3$) on the patent subgraph (representative run). Values close to 1 indicate boundary saturation in a curvature-comparable way, and saturation is already visible at shallow depths for both hyperbolic variants.

Local metric amplification. The Poincaré ball model is endowed with a conformal Riemannian metric

$$g_{\mathbf{x}} = \lambda(\mathbf{x})^2 I, \qquad \lambda(\mathbf{x}) = \frac{2}{1 - \|\mathbf{x}\|_2^2},$$

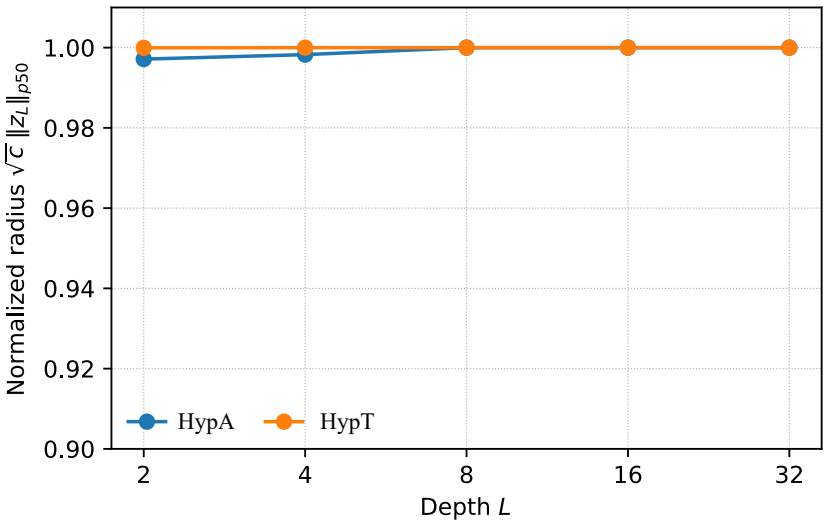

Figure D.1: Median normalized radius $r_{\mathrm{p50}}(L)$ at the final layer versus trained depth on the patent subgraph (representative run). Values near 1 indicate boundary saturation.

diverging as $\|\mathbf{x}\|_2 \to 1$. For sufficiently small updates,

$$d_{\mathrm{Hyp}}(\mathbf{x}, \mathbf{x} + \Delta\mathbf{x}) \approx \lambda(\mathbf{x})\|\Delta\mathbf{x}\|_2.$$

Thus, a large measured drift can arise from metric amplification near the boundary, even when the intrinsic coordinate updates remain modest. The same mechanism amplifies the squared-distance quantities (e.g., metric-aware Dirichlet energy) by approximately $\lambda(\mathbf{x})^2$ in small increments.

D.5.1 Layerwise coupling plot: boundary proximity versus drift/energy

To directly visualize how boundary proximity relates to the measured dynamics, we construct a layerwise coupling plot for the same HypA configuration highlighted in Figure 2. For each layer ℓ, we compute the upper-tail boundary-proximity proxy

$$r_{\mathrm{p90}}(\ell) := \mathrm{p90}_{i \in \mathcal{S}} \sqrt{c}\,\|\mathbf{z}_i^{(\ell)}\|_2$$

together with the interlayer drift $D(\ell)$ and Dirichlet energy $E(\ell)$ on the same fixed evaluation subsets. Figure 2 shows that both drift and energy rise sharply as r_{p90} approaches 1, consistent with a boundary-pressure regime in which boundary-associated metric amplification dominates the measured signals.

Empirical separation of coordinate steps and geodesic drift. To disambiguate metric amplification from large coordinate changes, we present (i) Euclidean coordinate-step magnitude in the Poincaré ball and (ii) the corresponding geodesic drift across layers, combined with the mean normalized radius (Figure D.3). We observed that the geodesic drift can grow sharply as the embeddings approach the boundary, even when coordinate-step magnitudes do not increase comparably, which is consistent with amplification by the local metric factor.

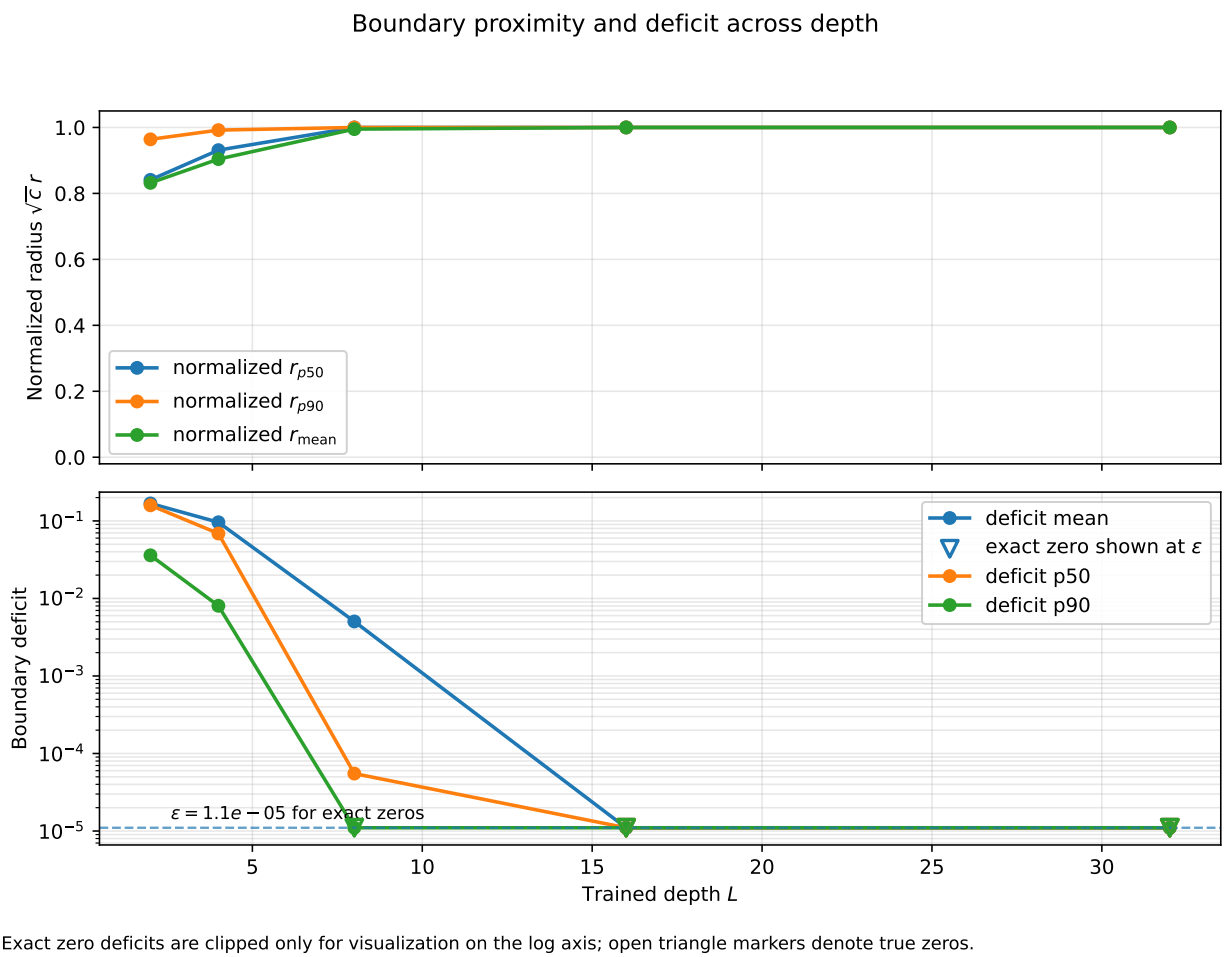

Figure D.2: **Boundary proximity and deficit across trained depths on the same deterministic subset as Figure 2.** For HypA, we summarize hyperbolic boundary statistics across trained depths for the same 2,000-node deterministic subset drawn from the fixed probe set $\mathcal{S}$ (top-10 classes; cap 200 nodes per class). Top: curvature-normalized radius summaries (mean, p50, and p90). Bottom: boundary-deficit summaries (mean, p50, and p90), where $\Delta_i^{(\ell)} = 1 - \sqrt{c}\,\|\mathbf{z}_i^{(\ell)}\|_2$. This figure preserves the original same-subset depth-sweep boundary summary corresponding to the geometry panels in Figure 2. For log-scale visualization of the boundary deficit, exact zeros are plotted at $\epsilon = 1.1 \times 10^{-5}$ and marked by open triangles.

Hyperbolic baselines and the role of HypT. We consider two hyperbolic variants. HypA is a representative off-the-shelf hyperbolic instantiation in a shared training protocol. HypT is a minimally tuned hyperbolic control (tangent-space classifier head, input scaling α_{in}, curvature c, and a learning-rate split between the encoder and head) introduced to ensure competitive shallow-depth performance. We included HypA to reflect typical hyperbolic behavior and HypT to argue against shallow-depth underoptimization as the primary explanation for the observed boundary-associated amplification.

Curvature-normalized boundary deficit across c. Raw Poincaré radii are not directly comparable across curvatures because the boundary radius depends on c: $\mathbb{B}_c^d = \{u \in \mathbb{R}^d : c\|u\|_2^2 < 1\}$ has boundary at $\|u\|_2 = 1/\sqrt{c}$. We therefore normalize radii by defining the curvature-

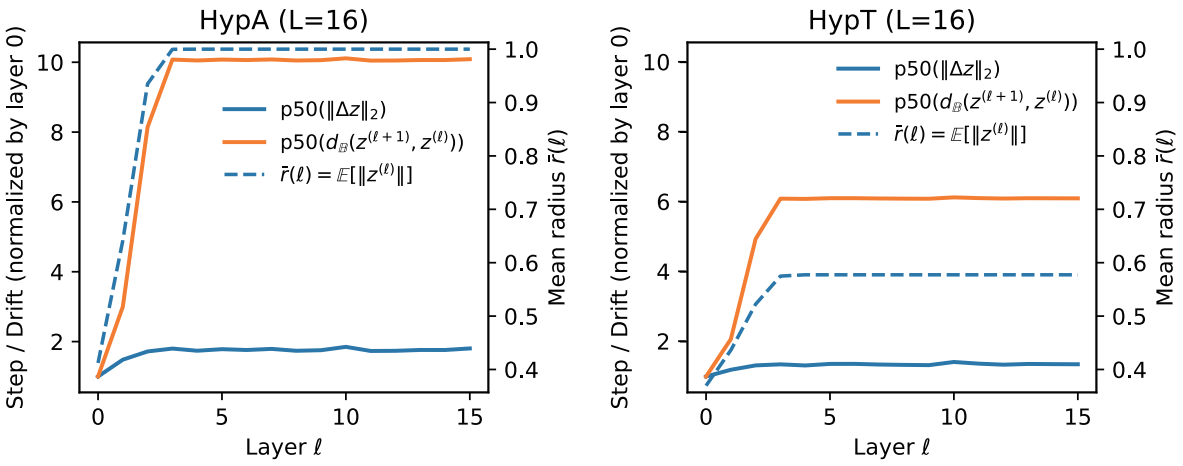

Figure D.3: Separating coordinate updates from metric-induced drift in hyperbolic GNNs. Layerwise comparison of (left) a representative hyperbolic baseline (HypA) and (right) a minimally tuned hyperbolic control (HypT) at depth $L = 16$ on the patent subgraph. For each layer ℓ, we present the median Euclidean coordinate-step magnitude $p50(\|\Delta z\|_2)$ and the median geodesic drift $p50(d_{\mathbb{B}_c}(z^{(\ell+1)}, z^{(\ell)}))$, both normalized by their layer-0 values (left axis), together with the mean normalized radius $\bar{r}(\ell) = \mathbb{E}_{i\in\mathcal{S}}[\sqrt{c}\,\|\mathbf{z}_i^{(\ell)}\|_2]$ (right axis). In HypA, geodesic drift increases sharply as embeddings approach the boundary ($\bar{r} \to 1$) despite modest coordinate-level updates. HypT, introduced to discard trivial underoptimization, exhibits qualitatively similar boundary-associated drift amplification. Overall, these trends indicate that the observed late-layer amplification reflects metric amplification near the boundary rather than unusually large coordinate updates.

normalized radius

$$\tilde{r}_i^{(\ell)} := \sqrt{c}\,\|\mathbf{z}_i^{(\ell)}\|_2 \in [0, 1),$$

and the corresponding *boundary deficit*

$$\Delta_i^{(\ell)} := 1 - \tilde{r}_i^{(\ell)}.$$

In this parameterization, $\Delta \approx 0$ indicates boundary saturation in a curvature-comparable way. Figure D.4 reports bulk deficit summaries across curvatures, showing that upper-tail proximity can saturate while bulk deficit remains informative for distinguishing depth regimes.

Heterogeneity of the extreme-depth regime. To summarize distributional distortion beyond bulk averages, we compare the median and mean deficits:

$$\Delta_{\text{p50}}^{(\ell)} := \text{median}_{i\in S}\,\Delta_i^{(\ell)}, \qquad \Delta_{\text{mean}}^{(\ell)} := \frac{1}{|S|}\sum_{i\in S}\Delta_i^{(\ell)},$$

and define a simple heterogeneity proxy

$$H_\Delta^{(\ell)} := \Delta_{\text{p50}}^{(\ell)} - \Delta_{\text{mean}}^{(\ell)}.$$

The corresponding export artifact labels this same summary as `gap_deficit_p50_mean`; throughout the paper we refer to it uniformly as H_Δ. Figure D.5 shows that H_Δ stays near zero for shallower depths but becomes qualitatively distinct at very large depth (L=32), consistent with a boundary-dominated regime whose distributional shape differs from shallower settings.

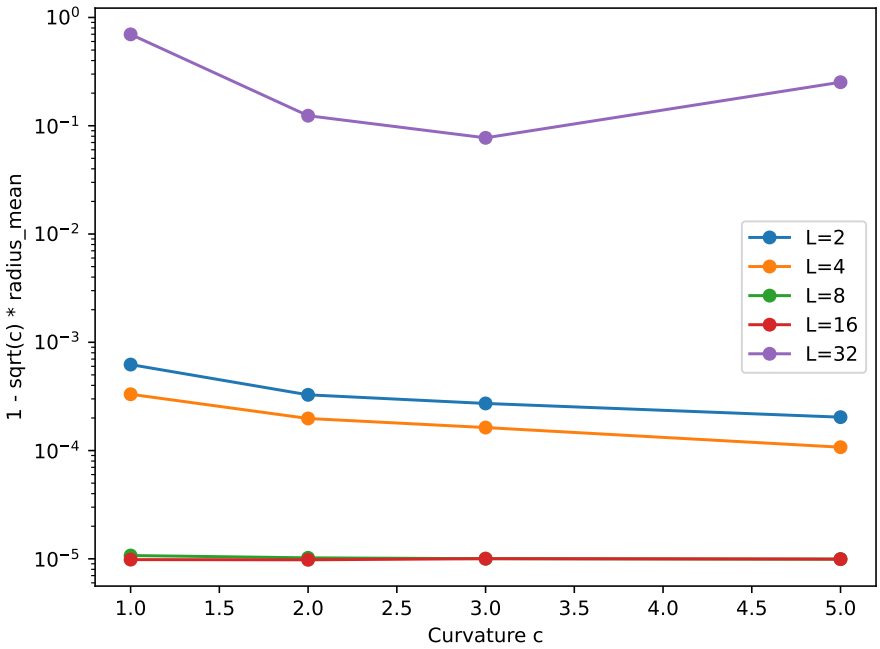

Figure D.4: **Curvature-normalized boundary deficit (bulk).** Mean boundary deficit $\Delta_{\text{mean}} = 1 - \sqrt{c}\, r_{\text{mean}}$ across curvatures c for HypT. A log scale highlights that bulk deficits can remain discriminative even when upper-tail proximity saturates, supporting a heterogeneous boundary-dominated regime at extreme depth.

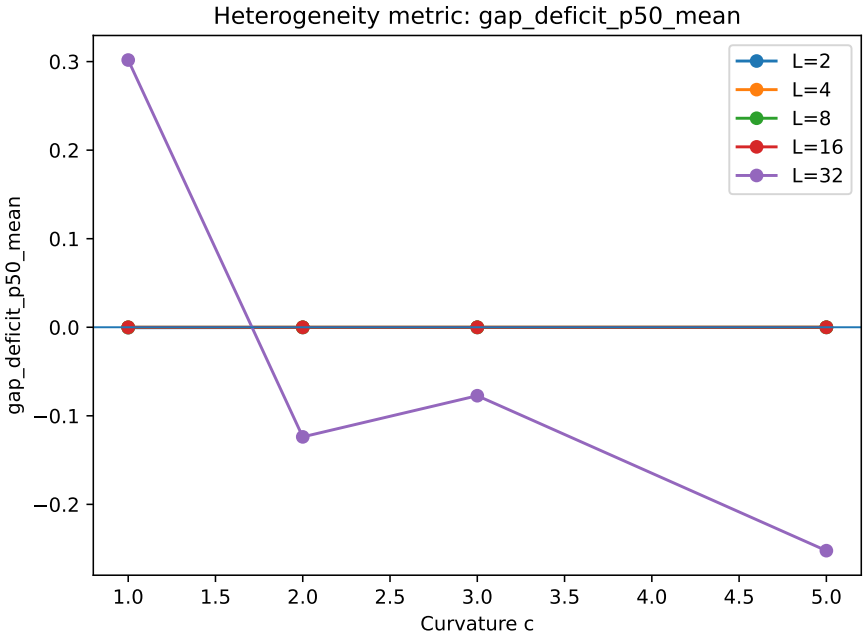

Figure D.5: **Boundary-deficit heterogeneity across curvatures.** We plot $H_\Delta = \Delta_{\text{p50}} - \Delta_{\text{mean}}$ across curvatures for HypT (the exported figure file labels the same quantity as `gap_deficit_p50_mean`). Shallower depths remain close to zero under this summary, whereas L=32 is clearly separated, indicating a qualitatively distinct very-deep regime with distributional distortion (tail saturation but nontrivial bulk).

D.6 Disentangling metric amplification from intrinsic representation change

A central concern in hyperbolic representation analysis is whether large measured interlayer drift reflects a genuine representational change or amplification induced by the Poincaré metric near the boundary. To disambiguate these effects, we explicitly compared multiple notions of drift computed for the same learned representation.

Euclidean-coordinate drift. First, we computed the coordinate-level drift by measuring the Euclidean displacement. $\|\Delta z^{(\ell)}\|_2 = \|z^{(\ell+1)} - z^{(\ell)}\|_2$ in ambient space. This quantity reflects the intrinsic update magnitude, which is independent of hyperbolic metrics.

Geometry-aware (hyperbolic) drift. We then computed the corresponding geodesic drift using the Poincaré distance $d_{\mathrm{Hyp}}(z^{(\ell+1)}, z^{(\ell)})$, which is a geometrically consistent concept in the main text.

Comparison and interpretation. As shown in Figure D.3, geodesic drift can grow sharply as representations approach the boundary even when Euclidean-coordinate updates remain modest. This separation indicates that the observed late-layer drift amplification in hyperbolic models cannot be solely attributed to remarkably large coordinate updates. Instead, it reflects the metric amplification induced by boundary proximity.

Therefore, we interpret boundary-associated drift amplification as a *geometry-driven amplification regime* rather than a pure measurement artifact. In addition, we emphasize that coordinate-level drift provides a complementary diagnostic and that geometry-aware and coordinate-level probes should be interpreted jointly, rather than in isolation.

D.7 Seed sensitivity and determinism

All the reported mean±std values were computed for the three training seeds. Given a fixed checkpoint and released probe/evaluation subsets (Appendix A.1), the probe computation is deterministic. For transparency and precise reproduction, the reproducibility bundle released full-per-seed CSV diagnostics for every depth, geometry, and activation setting (Appendix A.5), combined with scripts that regenerate all figures and tables.

D.8 Architectural scope

Our experiments focused on GraphSAGE-style message passing. Although the probing methodology is architecture agnostic, we do not claim that the observed regime boundaries are architecture universal. Extending the present analysis to attention-based or spectral/diffusion GNNs constitutes a next step and may reveal architecture-dependent variations in the regime structure.

D.9 Probe as diagnostic tool

We emphasize that fixed-point probing is a diagnostic instrument rather than a training objective. The incorporation of probe-derived signals as regularizers remains an interesting direction for future research.

Table E.1: Scope checks and the claims they support. These checks support protocol portability and bounded scope, not dataset-universal quantitative laws.

scope check	dataset/backbone	claim supported
Patent GraphSAGE stress test	Patent citation / GraphSAGE	Main bounded case-study mechanism.
Patent GCN backbone check	Patent citation / Euclidean GCN	The protocol is not tied to a single GraphSAGE configuration.
Public benchmark fixed-point checks	Cora / Texas, Euclidean public checks	The same fixed-point probing workflow can be applied beyond the patent graph.
OGBN-Arxiv reference checks	OGBN-Arxiv / GraphSAGE and GCN references	Additional public reference diagnostics for depth-wise representation changes.

Table E.2: What the scope checks do not claim. This table is included to make the boundary of the paper's claims explicit.

scope check	claim not supported
Patent GraphSAGE stress test	A dataset-universal law of hyperbolic depth failure.
Patent GCN backbone check	The same layerwise onset or degeneration location for every backbone.
Public benchmark fixed-point checks	Universal quantitative thresholds for drift, energy, separability, or boundary proximity.
OGBN-Arxiv reference checks	An optimized benchmark comparison or a performance-preserving universal intervention threshold.

E Robustness, sensitivity, and scope checks

This appendix collects the additional checks that make the protocol-level uncertainty, robustness, runtime cost, and scope boundaries explicit. The purpose is not to expand the patent-citation case study into a dataset-universal law. Instead, the checks below separate three claims: (i) fixed-point probing is a portable post-training protocol, (ii) the main patent-graph mechanism is a bounded stress-test finding, and (iii) the quantitative thresholds and layerwise onsets are not claimed to be universal. Unless stated otherwise, values are reported as mean±std over training seeds and/or resampling seeds as indicated by the table captions.

E.1 Scope of claims and portability checks

The main case study remains the patent GraphSAGE stress test. The additional GCN-backbone and public-benchmark checks are included to demonstrate that the protocol can be applied beyond a single GraphSAGE configuration, while also making explicit what these checks do *not* establish. Tables E.1 and E.2 summarize this boundary.

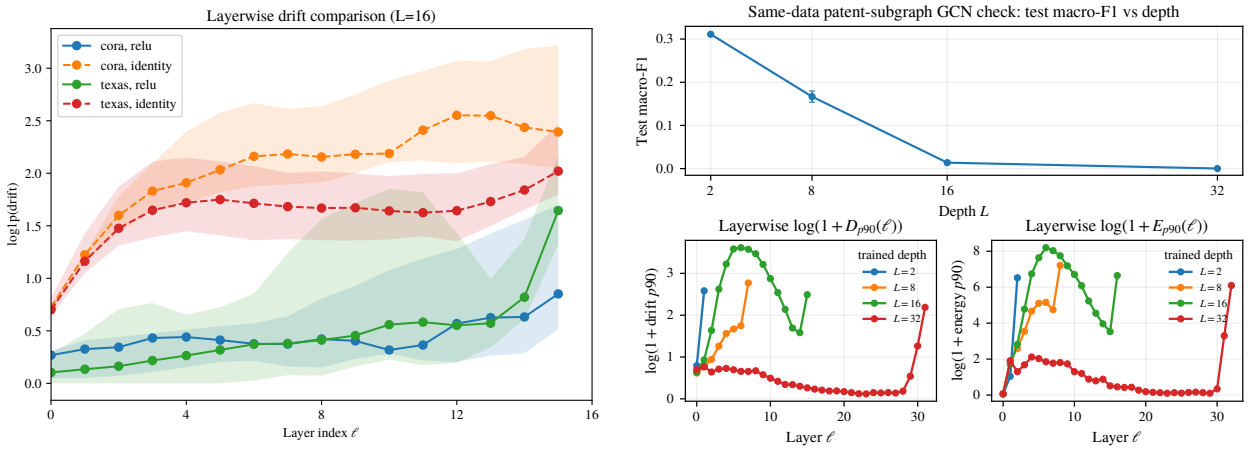

Figure E.1: Representative portability/scope artifacts. Left: public benchmark fixed-point probe comparison at $L = 16$. Right: Euclidean GCN backbone check on the patent subgraph. These are scope checks for protocol portability rather than evidence of universal quantitative thresholds.

E.2 Seed-aggregated uncertainty for the core diagnostics

Several mechanism figures in the main text are intentionally shown as representative layerwise traces. To make uncertainty explicit, we added seed-aggregated counterparts for the core summaries. Table E.3 reports representative final-layer or final-transition values at $L \in \{16, 32\}$; Figure E.2 shows the corresponding depth-sweep summaries with standard-deviation bars.

Table E.3: Seed-aggregated core diagnostic summaries at representative depths. Drift is the final-transition median drift $D_{p50}(L-1)$; energy and separability are last-layer summaries. The radius column is reported only for hyperbolic models.

mode	L	drift p50	energy mean	Sep	radius p90	n
EUCLID	16	5.761 ± 0.634	52.783 ± 8.023	1.406 ± 0.036	–	3
EUCLID	32	4.017 ± 1.123	23.812 ± 7.751	1.595 ± 0.146	–	3
HYPA	16	20.474 ± 0.036	261.480 ± 12.924	1.353 ± 0.096	1.000 ± 0.000	3
HYPA	32	20.536 ± 0.090	245.196 ± 32.467	1.237 ± 0.296	1.000 ± 0.000	3
HYPT	16	11.831 ± 0.050	122.766 ± 3.593	1.197 ± 0.047	0.577 ± 0.000	3
HYPT	32	10.189 ± 2.724	69.396 ± 57.506	–	0.577 ± 0.000	3

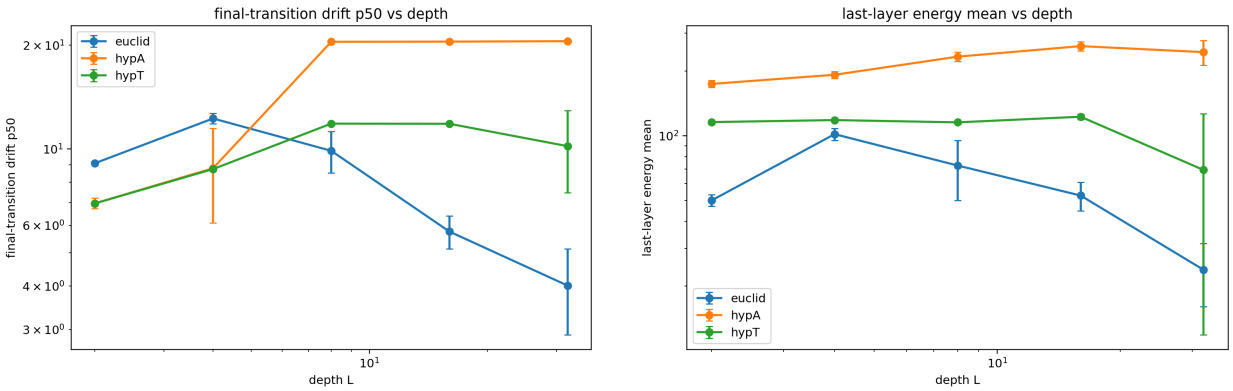

Figure E.2: Seed-aggregated final summaries with standard-deviation bars. Left: final-transition drift p50 versus trained depth. Right: last-layer Dirichlet energy mean versus trained depth. These plots complement the representative layerwise mechanism traces in the main text.

E.3 Probe-node and evaluation-edge budget sensitivity

We varied the probe-node budget $|\mathcal{S}| \in \{1k, 5k, 10k, 50k\}$ and the evaluation-edge budget $|\mathcal{E}_{\text{eval}}| \in \{10k, 50k, 100k, 200k\}$. For probe nodes, the final-transition drift estimates and qualitative ordering were stable across the tested budgets; increasing $|\mathcal{S}|$ mainly reduced resampling variability. For Dirichlet energy, the 10k-edge estimates were already close to the 200k default, with relative changes below 0.3% in all mode/depth conditions in Table E.6.

Practical recommendation. Practitioners can start with a small pilot budget, e.g., $|\mathcal{S}| = 1k$ and $|\mathcal{E}_{\text{eval}}| = 10k$, repeat the probe with a few resampling seeds, and increase the budget until the qualitative regime label and the rank/order of depthwise summaries are stable. Once selected, $\mathcal{S}$ and $\mathcal{E}_{\text{eval}}$ should be fixed and reported for all comparisons.

Table E.4: Probe-node budget sensitivity. The table compares the smallest and largest tested probe budgets. The last column is the relative change in final-transition drift p50 from $|\mathcal{S}| = 1k$ to $50k$.

mode	L	$\lvert\mathcal{S}\rvert = 1k$ drift	$\lvert\mathcal{S}\rvert = 50k$ drift	$\lvert\mathcal{S}\rvert = 1k$ Sep	$\lvert\mathcal{S}\rvert = 50k$ Sep	Δ drift (%)
EUCLID	16	6.970 ± 0.417	6.901 ± 0.410	1.267 ± 0.038	1.169 ± 0.030	-0.99
EUCLID	32	4.961 ± 0.424	4.890 ± 0.369	1.225 ± 0.111	1.119 ± 0.105	-1.43
HYPA	16	20.480 ± 0.016	20.477 ± 0.018	1.189 ± 0.098	1.119 ± 0.068	-0.01
HYPA	32	20.536 ± 0.078	20.536 ± 0.078	1.159 ± 0.076	1.073 ± 0.036	0.0
HYPT	16	11.839 ± 0.041	11.838 ± 0.039	1.058 ± 0.052	0.978 ± 0.041	-0.01
HYPT	32	7.841 ± 5.881	7.841 ± 5.881	–	–	0.0

Table E.5: Within-checkpoint probe-node resampling variability for drift p50. For each training checkpoint and budget, we repeated the probe with three independent probe-sampling seeds and then summarized the within-checkpoint standard deviation.

$\lvert\mathcal{S}\rvert$	mean resampling std	std across ckpts	n
1,000	0.0248	0.0478	18
5,000	0.0112	0.0213	18
10,000	0.0079	0.0132	18
50,000	0.0030	0.0053	18

Table E.6: Evaluation-edge budget sensitivity for Dirichlet energy. The 200k default is conservative: in these runs the 10k and 200k estimates differ by less than 0.3% in every condition.

mode	L	$\lvert\mathcal{E}_{eval}\rvert = 10k$ energy	$\lvert\mathcal{E}_{eval}\rvert = 200k$ energy	Δ energy (%)
EUCLID	16	52.817 ± 6.878	52.781 ± 6.948	-0.07
EUCLID	32	23.843 ± 6.734	23.813 ± 6.714	-0.13
HYPA	16	261.834 ± 11.224	261.477 ± 11.195	-0.14
HYPA	32	245.888 ± 28.217	245.194 ± 28.123	-0.28
HYPT	16	122.753 ± 3.110	122.768 ± 3.110	0.01
HYPT	32	69.243 ± 49.719	69.396 ± 49.802	0.22

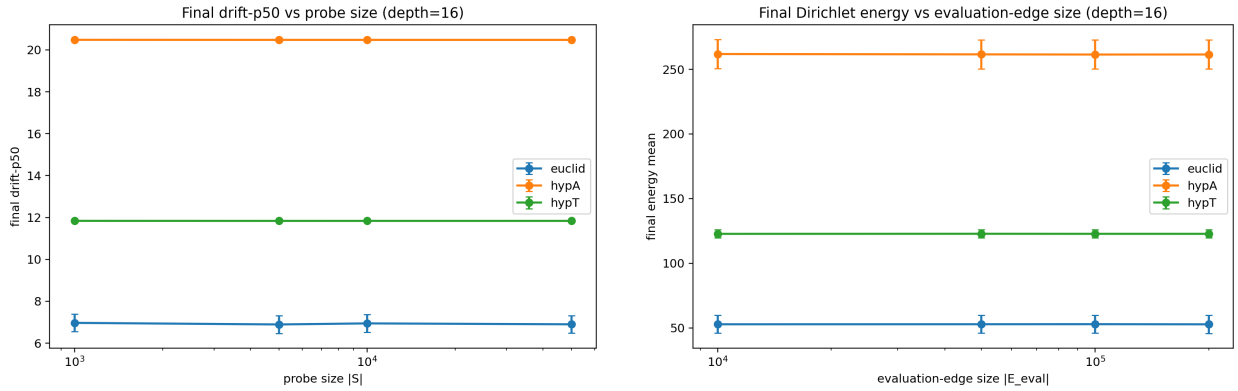

Figure E.3: Budget sensitivity at $L = 16$. Left: final-transition drift p50 versus probe-node budget. Right: final Dirichlet energy mean versus evaluation-edge budget. Error bars show standard deviations over training and resampling seeds.

E.4 Wall-clock runtime and monitoring cost

The probes require no backpropagation and operate on fixed subsets. With the default budgets $|\mathcal{S}| = 50k$ and $|\mathcal{E}_{\text{eval}}| = 200k$, the full diagnostic takes approximately 17–24 seconds per checkpoint in our environment (Table E.7). The small pilot budget takes approximately 9–10 seconds per checkpoint. In practice, the full default probe is most naturally run at validation intervals or saved checkpoints, while the smaller pilot budgets are cheap enough for lightweight training-time monitoring; we do not recommend running the full default probe after every mini-batch.

Table E.7: Default per-checkpoint runtime in seconds at $|\mathcal{S}| = 50k$ and $|\mathcal{E}_{\text{eval}}| = 200k$. The total combines the probe pass and the Dirichlet-energy pass by quadrature for the displayed standard deviation.

mode	L	probe time	energy time	total time
EUCLID	16	6.795 ± 0.253	9.929 ± 0.167	16.724 ± 0.303
EUCLID	32	9.178 ± 0.186	14.444 ± 0.131	23.622 ± 0.228
HYPA	16	6.897 ± 0.157	9.693 ± 0.170	16.590 ± 0.231
HYPA	32	9.283 ± 0.250	14.417 ± 0.340	23.700 ± 0.422
HYPT	16	6.830 ± 0.229	9.706 ± 0.190	16.536 ± 0.298
HYPT	32	9.276 ± 0.123	14.719 ± 1.444	23.995 ± 1.449

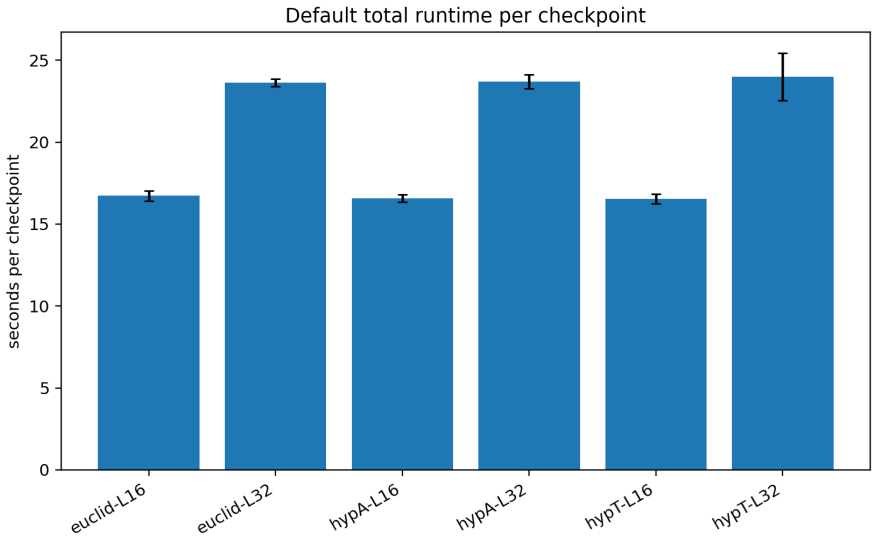

Figure E.4: Default total runtime per checkpoint at $|\mathcal{S}| = 50k$ and $|\mathcal{E}_{\text{eval}}| = 200k$.

E.5 Revised boundary-intervention interpretation for Figure 3

We make the schedule explicit: $\lambda = 10$ is a late-only/delayed intervention, whereas $\lambda = 100$ is applied globally from epoch 0. We also avoid interpreting the intervention as a smooth or performance-preserving control law. A terminal-window average over layers 24–32 can be misleading because the $\lambda = 0$ trace contains a terminal drop/collapse artifact; therefore, Table E.8 reports the primary numeric summary over the pre-terminal saturated window, layers 5–28. The interpretation should also account for the different metric sensitivity of the two probes: local geodesic drift scales approximately as $\lambda(x)\|\Delta x\|_2$, while the squared distances in metric-aware Dirichlet energy are approximately $\lambda(x)^2$-sensitive. Thus a radius perturbation can reduce the high-energy portion of the trajectory more clearly than it reduces median drift, whose layerwise value is also affected by direction and timing.

The revised interpretation is consequently cautious: late-only control remains boundary-saturated and does not materially suppress the amplification. Strong global control from

Table E.8: Pre-terminal saturated-window summary for the boundary-intervention diagnostic at $L = 32$ for HypA. The global $\lambda = 100$ condition reduces the high-energy portion of the trajectory in this window, whereas the drift response is not a clean monotonic attenuation. This asymmetry is expected because Dirichlet energy is a squared-distance statistic, while drift is a first-order displacement statistic; the intervention is therefore interpreted as a diagnostic perturbation rather than as a remedy or control law.

condition	schedule	window	r_{p90}	drift p50	energy mean	energy max	use
$\lambda = 0$, no radius control	none	5–28	1.000 ± 0.000	20.149 ± 1.699	311.134 ± 28.291	358.352	primary
$\lambda = 10$, late-only control	late-only / delayed	5–28	1.000 ± 0.000	20.491 ± 0.138	297.545 ± 37.358	350.118	primary
$\lambda = 100$, global from epoch 0	global / always-on from epoch 0	5–28	1.000 ± 0.000	19.943 ± 2.296	257.816 ± 31.666	305.717	primary

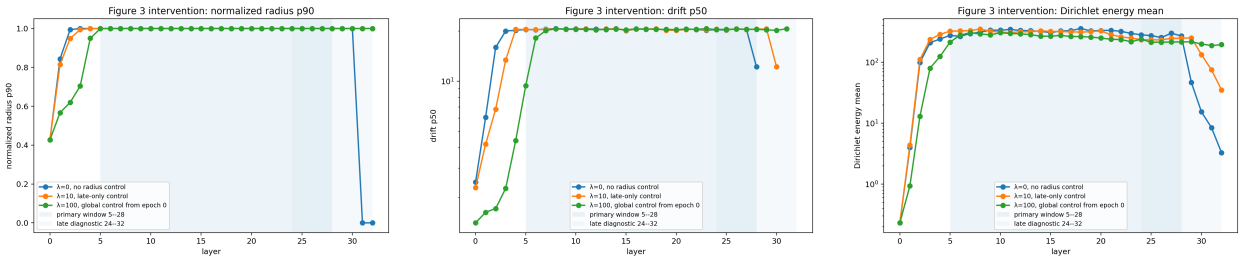

Figure E.5: Boundary-intervention diagnostics with the primary pre-terminal window shaded. Left: upper-tail normalized radius. Middle: drift p50. Right: Dirichlet energy mean. The weak late-only intervention remains boundary-saturated, while the always-on $\lambda = 100$ condition reduces the high-energy portion of the saturated trajectory.

epoch 0 changes the saturated trajectory and reduces the high-energy portion of the curve, but the response is timing-sensitive and nonmonotonic, especially for drift. This difference is compatible with the conformal-factor asymmetry above: the energy probe is a squared-distance statistic and is expected to respond more sharply to changes in boundary pressure than the median drift probe. We therefore treat this intervention as diagnostic evidence consistent with a boundary contribution, not as a robust radius-control law or a performance-preserving regularizer.

E.6 Hidden-width sensitivity

To address the reviewer's question about embedding dimension, we interpret the queried dimension as the GraphSAGE encoder hidden width and add a compact sweep over $d \in \{64, 128, 256\}$ at $L \in \{16, 32\}$ for all three geometries. This sweep varies the encoder hidden width while keeping both the raw input text-feature dimension and the output representation dimension fixed (output dimension 64). Thus, it tests sensitivity to encoder capacity rather than a change to the input text embedding or the final output representation dimension. Across widths, the absolute scale of the probes and endpoint metrics varies, but changing the hidden width does not remove the severe $L = 32$ degradation regime. A sweep over the final output representation dimension is a separate axis and is not claimed here.

Table E.9: Hidden-width sensitivity sweep. The swept dimension is the GraphSAGE encoder hidden width; the raw input text-feature dimension and the output representation dimension are fixed, with output dimension 64. Values are mean±std over three training seeds.

d	mode	L	test macro-F1	drift p50	energy mean	Sep
64	Euclid	16	0.008 ± 0.003	5.173 ± 0.424	26.040 ± 3.564	1.164 ± 0.019
64	Euclid	32	0.001 ± 0.001	2.701 ± 1.730	34.189 ± 12.966	0.825 ± 0.257
64	HypA	16	0.007 ± 0.003	20.556 ± 0.006	288.151 ± 2.227	1.134 ± 0.029
64	HypA	32	0.001 ± 0.000	6.825 ± 11.822	97.735 ± 134.797	–
64	HypT	16	0.015 ± 0.005	11.840 ± 0.015	118.099 ± 0.380	0.981 ± 0.044
64	HypT	32	0.001 ± 0.000	11.834 ± 0.094	78.575 ± 35.980	–
128	Euclid	16	0.013 ± 0.003	6.721 ± 0.821	51.071 ± 15.287	1.163 ± 0.035
128	Euclid	32	0.001 ± 0.000	4.564 ± 1.926	16.876 ± 13.097	1.143 ± 0.119
128	HypA	16	0.014 ± 0.004	20.482 ± 0.049	264.856 ± 12.673	1.140 ± 0.045
128	HypA	32	0.001 ± 0.000	20.519 ± 0.118	181.991 ± 41.026	–
128	HypT	16	0.023 ± 0.005	11.816 ± 0.023	122.181 ± 2.183	0.982 ± 0.019
128	HypT	32	0.001 ± 0.000	10.232 ± 2.768	62.976 ± 47.709	–
256	Euclid	16	0.012 ± 0.008	7.779 ± 0.700	71.797 ± 7.830	1.153 ± 0.023
256	Euclid	32	0.001 ± 0.000	3.255 ± 1.590	50.598 ± 19.797	0.859 ± 0.294
256	HypA	16	0.020 ± 0.001	20.480 ± 0.070	255.395 ± 7.582	1.039 ± 0.036
256	HypA	32	0.001 ± 0.000	13.601 ± 11.780	98.472 ± 75.789	–
256	HypT	16	0.030 ± 0.009	11.813 ± 0.047	122.427 ± 1.730	1.018 ± 0.014
256	HypT	32	0.001 ± 0.000	10.195 ± 2.729	47.466 ± 33.475	–

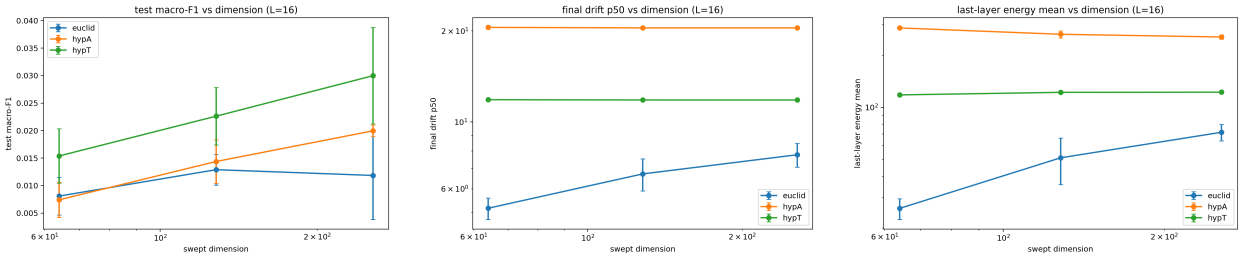

Figure E.6: Hidden-width sensitivity at $L = 16$. Changing hidden width affects endpoint metrics and absolute probe scales, but does not remove the qualitative geometry-dependent differences.