WHY GRAPH PRETRAINING FAILS TO TRANSFER: CAPABILITY CEILING AND ACCESSIBILITY DEFICIT

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph self-supervised learning (SSL) provides the foundation for reusable graph representations. However, cross-domain pretraining often fails to match target-domain SSL performance, leading to incomplete or negative transfer. We decompose this transfer gap into two distinct mechanisms: *capability loss*, where target-task information is not retained, and *finite-supervision accessibility loss*, where retained information is difficult to extract from limited target labels. To unify both mechanisms, we propose the Task–Geometry Profile (TGP), a theoretical framework that models representation-mode strength alongside target-task signal alignment. Our analysis reveals that positive transfer depends not only on how much task-relevant signal is retained, but also on where that signal is organized in representation geometry. Furthermore, we show how graph aggregation architectures distort this profile through distinct inductive biases: GCN's normalized diffusion induces mode-selective spectral attenuation, whereas GIN's unnormalized accumulation causes degree-driven modal reallocation. Across diverse domains, SSL objectives, backbones, and 120 transfer settings, TGP quantities show systematic correspondence with empirical capability and accessibility behavior. Mechanism-guided interventions further reveal that the two components need not improve together: degree-tempered aggregation reduces the trajectory-level transfer deficit by up to 56% on the classical benchmark suite, while large-scale validation shows substantial absolute utility gains.

1 INTRODUCTION

The emergence of foundation models has shifted representation learning toward acquiring transferable knowledge that generalizes across tasks and domains (Bommasani et al., 2021). In graph machine learning, this paradigm has driven the development of Graph Foundation Models (GFMs) designed to leverage diverse graph data for downstream transfer (Liu et al., 2023a; Mao et al., 2024; Liu et al., 2024). A primary requirement of GFMs is sample-efficient adaptation: pretrained representations should achieve strong generalization with minimal target-domain supervision. Graph self-supervised learning (SSL) offers a compelling approach for this goal by learning structured representations without task annotations (Liu et al., 2023b).

However, cross-domain graph pretraining rarely guarantees positive transfer (Hu et al., 2020b; Qiu et al., 2020). Pretrained models frequently underperform compared to target-matched SSL—where pretraining occurs directly on unlabeled target graphs—and can even exhibit negative transfer under source–target mismatch (Hu et al., 2020b; Wang et al., 2024). This persistent discrepancy presents a fundamental question: *If target-domain SSL establishes the attainable utility of a backbone and objective, why is this utility lost after cross-domain pretraining?*

Existing graph-transfer methods primarily focus on mitigating source–target distribution shift or enhancing downstream adaptation (Wang et al., 2024; Tian et al., 2025; Wang et al., 2026). Our focus is complementary: we decouple whether the degradation in target-matched utility stems from an inherent lack of expressiveness or poor finite-sample accessibility, and trace both bottlenecks directly to representation geometry and graph aggregation dynamics. Extended discussion appears in Appendix A.

Using target-matched SSL as the attainable-utility benchmark, we compare *frozen* representations while varying only the target-label budget. As shown in Fig. 1, source-pretrained representations

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

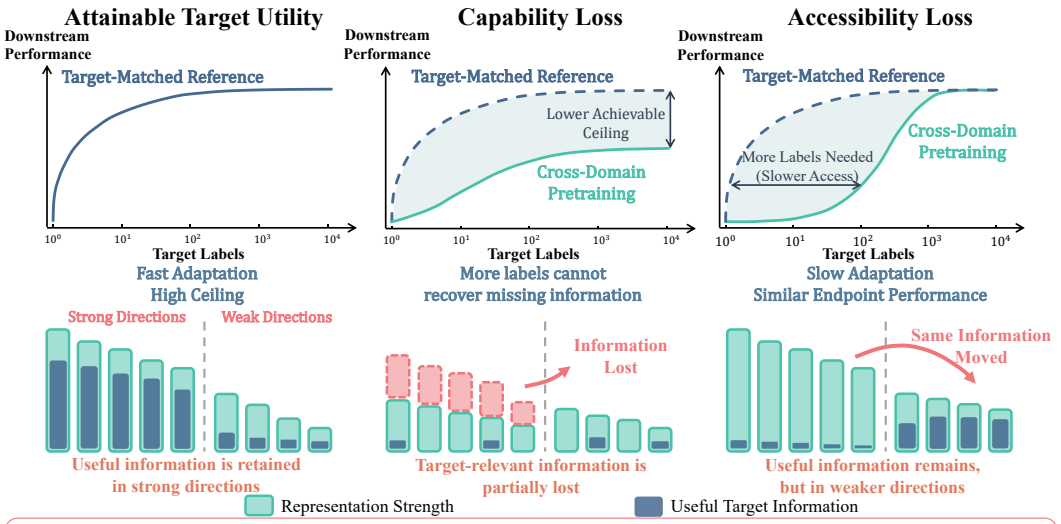

Figure 1: **Two failure modes of cross-domain utility transfer under frozen representations.** Varying only target-label supervision reveals capability loss as a lower attainable ceiling and accessibility loss as slower recovery caused by task signal residing in weaker representation modes.

may require substantially more labels to approach target-matched utility, or remain strictly capped below it even as supervision increases. We formalize these failure modes as *finite-supervision accessibility loss* and *capability loss*, respectively.

To explain these regimes, we introduce the *Task–Geometry Profile* (TGP), which couples representation-mode strength with target-task signal allocation. We show that capability depends on the amount of retained task signal, whereas finite-supervision accessibility depends on its modal placement. We further connect TGP distortion to architectural inductive biases: normalized GCN (Kipf & Welling, 2017) diffusion reshapes representation covariance through spectral filtering, while unnormalized GIN (Xu et al., 2019) aggregation introduces degree-dependent gain.

Across diverse SSL objectives, domains, backbones, and probes, our experiments reveal widespread capability and accessibility deficits and systematic correspondence with TGP diagnostics. Controlled interventions further expose that the two components can trade off even when overall transferable utility improves.

Our contributions are threefold. (1) We decompose cross-domain utility transfer into capability loss and finite-supervision accessibility loss relative to target-matched SSL. (2) We present TGP, a theoretical framework showing that sample efficiency is determined by the alignment between feature singular values and target-task label projections. (3) We characterize how normalized diffusion and degree-dependent aggregation reshape the Task–Geometry profile, and use controlled interventions to demonstrate how these mechanisms can guide architecture-level improvements in transferable graph representations.

2 Measuring Utility Transfer

2.1 Cross-Domain Evaluation Protocol

We evaluate cross-domain SSL against an objective- and backbone-matched target reference to quantify the utility lost under cross-domain pretraining and characterize this deficit through capability and finite-supervision accessibility. Consider a graph $\mathcal{G} = (V, E, \mathbf{X})$ with $N = |V|$ nodes and $\mathbf{X} \in \mathbb{R}^{N \times d_f}$. We use three encoder roles $r \in \{0, s, t\}$: randomly initialized f_0, source-pretrained f_s, and target-matched f_t. The pretrained encoders share the SSL objective, backbone, and training setup, differing only in the pretraining domain. Each frozen encoder produces target representations

$\mathbf{Z}_r = f_r(\mathcal{G}) \in \mathbb{R}^{N\times d}$. For probe family $\mathcal{H} \in \{\text{Linear}, \text{MLP}\}$, let $L_r(n)$ denote performance with n labeled target nodes.

2.2 Utility Transfer Decomposition

At target-label budget n, we define the utility transfer deficit as

$$G(n) = L_t(n) - L_s(n), \tag{1}$$

where $G(n) > 0$ indicates utility attainable by target-matched SSL but not retained after source-domain pretraining.

Definition 1 (Empirical Capability–Accessibility Decomposition)**.** Let $n_{\min}$ and $n_{\max}$ denote the smallest and largest evaluated target-label budgets. We define

$$\Delta_C = G(n_{\max}), \qquad A_G = \frac{1}{\log n_{\max} - \log n_{\min}} \int_{n_{\min}}^{n_{\max}} G(n)\, d(\log n), \tag{2}$$

and $\Delta_A = A_G - \Delta_C$.

Here, A_G is the normalized area between source- and target-pretrained learning curves on the log-label scale, while Δ_C is the endpoint deficit and Δ_A its finite-supervision excess. Thus Δ_C serves as the empirical proxy for population capability. Section 3 formalizes this distinction, with the exact learning-curve identity deferred to Appendix D.1.

3 A Task–Geometry Theory of Utility Transfer

We now formalize capability as the population-optimal utility of a fixed representation and accessibility as its additional finite-supervision cost. Our theory is developed in a linear representation-level setting: population capability is characterized by the optimal linear predictor, and finite-supervision accessibility by a spectral ridge surrogate. We additionally evaluate nonlinear MLP probes to test the robustness of these predictions beyond the linear setting. Fix a target graph $\mathcal{G}$ and a representation role r. Let $\mathbf{z}_r \in \mathbb{R}^d$ denote the centered representation of a uniformly sampled target node, and let $y \in \mathbb{R}$ denote its centered target label. We define the population-optimal linear predictor as

$$\boldsymbol{\beta}_r^\star = \arg\min_{\boldsymbol{\beta}} \mathbb{E}\left[(y - \mathbf{z}_r^\top \boldsymbol{\beta})^2\right]. \tag{3}$$

Let $\xi_r = y - \mathbf{z}_r^\top \boldsymbol{\beta}_r^\star$ denote the corresponding residual. By optimality, $\mathbb{E}[\mathbf{z}_r \xi_r] = \mathbf{0}$, and we define its residual variance as $\sigma_r^2 := \mathbb{E}[\xi_r^2]$. The population projection and profile construction are detailed in Appendix D.2.1.

3.1 Task–Geometry Profile

Here, a *representation mode* refers to an eigendirection of the feature covariance $\boldsymbol{\Sigma}_r \in \mathbb{R}^{d\times d}$. Let $\boldsymbol{\Sigma}_r = \mathbb{E}[\mathbf{z}_r \mathbf{z}_r^\top] = \mathbf{Q}_r \boldsymbol{\Lambda}_r \mathbf{Q}_r^\top$, where $\boldsymbol{\Lambda}_r = \text{diag}(\lambda_{r,1}, \ldots, \lambda_{r,d})$ and $\mathbf{q}_{r,i}$ is the eigenvector associated with $\lambda_{r,i}$. Expanding the optimal task direction in this eigenbasis gives $\boldsymbol{\beta}_r^\star = \sum_{i=1}^d b_{r,i} \mathbf{q}_{r,i}$.

Definition 2 (Task–Geometry Profile)**.** For representation role r, the target-task signal carried by mode i is $\psi_{r,i} := \lambda_{r,i} b_{r,i}^2$. The TGP is

$$\mathcal{P}_r = \{(\lambda_{r,i}, \psi_{r,i})\}_{i=1}^d, \tag{4}$$

with total target-task signal $\Psi_r := \sum_{i=1}^d \psi_{r,i}$. The profile jointly characterizes the statistical strength of each representation mode and the target-task signal carried by that mode.

The TGP is the representation-level interface used throughout the following analysis. Its total signal Ψ_r determines population capability, whereas the placement of $\{\psi_{r,i}\}$ relative to mode strengths $\{\lambda_{r,i}\}$ determines finite-supervision resolvability. Thus, capability depends primarily on *how much* target-task signal is retained, while accessibility depends on *where* that signal resides in the representation geometry.

3.2 Retained Task Signal and Capability

Proposition 1 (Capability–Signal Identity). *Let $R_r^\star = \mathbb{E}[(y - \mathbf{z}_r^\top \boldsymbol{\beta}_r^\star)^2] = \sigma_r^2$ denote the population residual risk. For the population-optimal predictor in Eq. 3,*

$$\mathrm{Var}(y) = \Psi_r + R_r^\star. \tag{5}$$

Consequently, for source-pretrained and target-matched representations of the same target task,

$$R_s^\star - R_t^\star = \Psi_t - \Psi_s. \tag{6}$$

Hence $\Psi_t - \Psi_s$ exactly quantifies the population capability lost under source pretraining; see Appendix D.2.2.

3.3 Finite-Supervision Mode Resolution

To isolate mode-wise resolvability under a fixed TGP, we introduce a population-eigenmode spectral ridge surrogate. For each mode $\mathbf{q}_{r,i}$, we define the finite-supervision spectral score as

$$\begin{aligned} g_{r,i}(n) &= n\lambda_{r,i} b_{r,i} + \zeta_{r,i}(n), \\ \mathbb{E}[\zeta_{r,i}(n)] = 0, \qquad & \mathbb{E}[\zeta_{r,i}(n)\zeta_{r,j}(n)] = n\sigma_r^2 \lambda_{r,i}\delta_{ij}. \end{aligned} \tag{7}$$

where $\zeta_{r,i}(n)$ denotes the finite-supervision score noise and δ_{ij} is the Kronecker delta. The corresponding ridge estimate with regularization parameter $\eta > 0$ satisfies

$$\widehat{b}_{r,i}(n) = \frac{g_{r,i}(n)}{n\lambda_{r,i} + \eta}, \qquad \mathbb{E}[\widehat{b}_{r,i}(n)] = a_{r,i}(n) b_{r,i}, \qquad a_{r,i}(n) = \frac{n\lambda_{r,i}}{n\lambda_{r,i} + \eta}. \tag{8}$$

For $\lambda_{r,i} > 0$, $a_{r,i}(n)$ increases with both n and $\lambda_{r,i}$ and approaches one as $n \to \infty$; hence signal in weaker modes requires more supervision to resolve. The surrogate construction and derivations are detailed in Appendix D.2.3.

3.4 Task–Geometry Transfer Decomposition

Accordingly, we define

$$\mathcal{U}_r(n) = \sum_{i=1}^{d} \psi_{r,i}[1 - a_{r,i}(n)]^2, \qquad \mathcal{V}_r(n) = \frac{\sigma_r^2}{n} \sum_{i=1}^{d} a_{r,i}^2(n), \qquad \mathcal{A}_r(n) = \mathcal{U}_r(n) + \mathcal{V}_r(n), \tag{9}$$

where $\mathcal{U}_r(n)$ is the unresolved task signal, $\mathcal{V}_r(n)$ is the finite-sample estimation variance, and $\mathcal{A}_r(n)$ is their total finite-supervision accessibility cost. Let $\widehat{\boldsymbol{\beta}}_{r,n} = \sum_i \widehat{b}_{r,i}(n)\mathbf{q}_{r,i}$ and define the expected risk on an independent target sample $(\mathbf{z}_{r,0}, y_0)$ as $R_{r,n} = \mathbb{E}[(y_0 - \mathbf{z}_{r,0}^\top \widehat{\boldsymbol{\beta}}_{r,n})^2]$.

Theorem 1 (Task–Geometry Transfer Decomposition). *Under the spectral ridge model in Eqs. 7–8,*

$$R_{r,n} = R_r^\star + \mathcal{A}_r(n). \tag{10}$$

Consequently,

$$R_{s,n} - R_{t,n} = \underbrace{\Psi_t - \Psi_s}_{\textit{capability loss}} + \underbrace{\mathcal{A}_s(n) - \mathcal{A}_t(n)}_{\textit{finite-supervision accessibility loss}}. \tag{11}$$

Theorem 1 separates lost population capability from the finite-label cost induced by task–geometry organization; see Appendix D.2.4 for the proof.

Corollary 1 (Large-Sample Separation). *For fixed representation dimension d and ridge parameter $\eta > 0$, $\lim_{n\to\infty} \mathcal{A}_r(n) = 0$. Consequently,*

$$\lim_{n\to\infty} (R_{s,n} - R_{t,n}) = \Psi_t - \Psi_s. \tag{12}$$

Thus, finite-supervision effects vanish asymptotically, isolating retained task signal as the remaining source of transfer loss. Proofs of all corollaries are in Appendix D.2.5.

Corollary 2 (Pure Accessibility Degradation)**.** *If $\Psi_s = \Psi_t$, then*

$$R_{s,n} - R_{t,n} = \mathcal{A}_s(n) - \mathcal{A}_t(n). \tag{13}$$

Hence, equal population capability does not imply equal finite-supervision transfer: representations retaining the same total task signal can differ solely in how accessible that signal is.

Corollary 3 (Signal Relocation)**.** *For fixed spectrum and total signal Ψ, let $\mathcal{U}$ and $\mathcal{U}'$ denote the unresolved-signal costs before and after relocating signal $\delta > 0$ from mode λ_i to a weaker mode $\lambda_j < \lambda_i$. Then, at every finite budget,*

$$\mathcal{U}'(n) - \mathcal{U}(n) = \delta \left[\left(\frac{\eta}{n\lambda_j + \eta} \right)^2 - \left(\frac{\eta}{n\lambda_i + \eta} \right)^2 \right] > 0. \tag{14}$$

Thus, relocating fixed task signal toward weaker representation modes preserves population capability while degrading finite-supervision accessibility. Collectively, the corollaries formalize the central distinction of TGP: capability is a question of how much task-relevant signal survives transfer, while accessibility is a question of where that surviving signal resides in representation geometry. This task-dependent nature of accessibility also limits what can be guaranteed by task-agnostic geometric improvements.

Proposition 2 (Task-Agnostic Geometry Ceiling)**.** *Consider two geometries with eigenvalues $\{\lambda_i\}_{i=1}^d$ and $\{\lambda_i^*\}_{i=1}^d$, and denote their unresolved-signal costs by $\mathcal{U}$ and $\mathcal{U}^*$, respectively. For a fixed total task signal $\Psi > 0$, suppose there exists a mode j such that $0 < \lambda_j^* < \lambda_j$. Then there exists a valid task-signal allocation with $\psi_j = \Psi$ and $\psi_i = 0$ $(i \neq j)$, for which, at any finite n,*

$$\mathcal{U}^*(n) > \mathcal{U}(n). \tag{15}$$

Therefore, improving representation geometry without task information cannot guarantee improved accessibility across tasks. See Appendix D.3 for the proof.

3.5 From Graph Aggregation to Profile Distortion

We connect TGP distortion to two canonical graph aggregation mechanisms: normalized diffusion in GCN and unnormalized degree-dependent accumulation in GIN. Graph spectral modes and TGP representation modes are linked through aggregation-induced covariance distortion: spectral filtering reshapes the feature covariance that defines the TGP.

Consider a message-passing layer with target-node states $\mathbf{H}_r^{(\ell)} \in \mathbb{R}^{N \times d_\ell}$ and aggregation operator $\mathbf{A}_{\text{agg}} \in \mathbb{R}^{N \times N}$. Its linear aggregation is $\mathbf{M}_r^{(\ell)} = \mathbf{A}_{\text{agg}} \mathbf{H}_r^{(\ell)}$. With the centering matrix $\mathbf{J}_N = \mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$, the centered covariance after aggregation is

$$\mathbf{\Sigma}_{\text{agg},r}^{(\ell)} = \frac{1}{N} (\mathbf{H}_r^{(\ell)})^\top \mathbf{A}_{\text{agg}}^\top \mathbf{J}_N \mathbf{A}_{\text{agg}} \mathbf{H}_r^{(\ell)}. \tag{16}$$

Thus $\mathbf{A}_{\text{agg}}$ directly shapes representation geometry before role-specific transformations; derivations are given in Appendix D.4.

Normalized diffusion. Let $\mathbf{P} = \mathbf{V}\mathbf{\Omega}\mathbf{V}^\top$ be the normalized propagation operator of the target graph, with $\mathbf{\Omega} = \text{diag}(\omega_1, \ldots, \omega_N)$. For a linearized L-layer GCN, $\mathbf{Z}_r = \mathbf{P}^L \mathbf{X} \mathbf{W}_r$. Because $\mathbf{P}^L = \mathbf{V}\mathbf{\Omega}^L\mathbf{V}^\top$, graph Fourier mode k is scaled by ω_k^L, so propagation selectively attenuates modes with $|\omega_k| < 1$. The propagated feature geometry can be written as

$$\mathbf{K}_L = \frac{1}{N} \mathbf{X}^\top (\mathbf{P}^L)^\top \mathbf{J}_N \mathbf{P}^L \mathbf{X}, \qquad \mathbf{\Sigma}_r = \mathbf{W}_r^\top \mathbf{K}_L \mathbf{W}_r. \tag{17}$$

Normalized diffusion therefore creates an anisotropic substrate before the role-specific projection $\mathbf{W}_r$; source–target mismatch can relocate retained task signal toward weaker representation modes.

Degree-dependent gain. GIN uses unnormalized neighborhood accumulation, $\mathbf{m}_{r,v} = (1 + \epsilon_{\text{GIN}})\mathbf{h}_{r,v} + \sum_{u \in \mathcal{N}(v)} \mathbf{h}_{r,u}$. Define the normalized local content as

$$\bar{\mathbf{h}}_{r,v} = \frac{(1 + \epsilon_{\text{GIN}})\mathbf{h}_{r,v} + \sum_{u \in \mathcal{N}(v)} \mathbf{h}_{r,u}}{d_v + 1}. \tag{18}$$

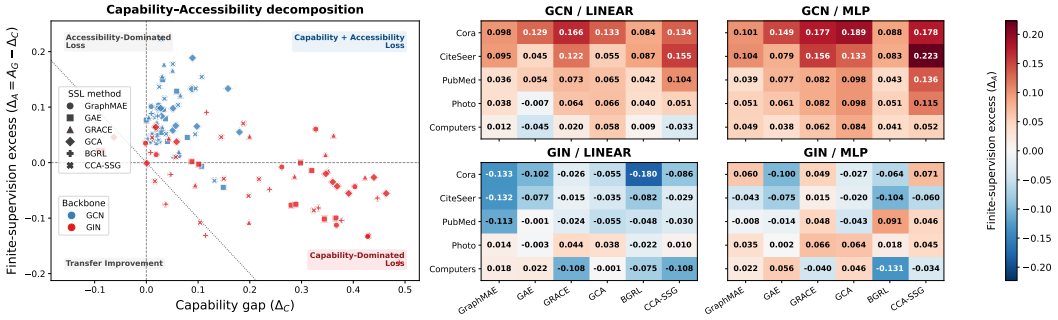

Figure 2: **Empirical landscape of utility transfer.** *Left:* each point is one SSL–backbone–target setting, positioned by capability loss Δ_C and finite-supervision accessibility loss Δ_A. Positive values indicate degradation relative to target-matched SSL, while the diagonal $\Delta_C + \Delta_A = 0$ separates net utility loss from net gain over the supervision range. *Right:* Δ_A across targets, SSL objectives, backbones, and probe families, highlighting the distinct accessibility regimes of GCN and GIN.

Then $\mathbf{m}_{r,v} = (d_v + 1)\bar{\mathbf{h}}_{r,v}$, which separates local content from the degree-dependent gain. Let $\mathbf{D} = \mathrm{diag}(d_1 + 1, \ldots, d_N + 1)$ and let $\bar{\mathbf{H}}_r$ stack $\bar{\mathbf{h}}_{r,v}$ row-wise. The centered pre-MLP covariance is

$$\mathbf{\Sigma}_{\mathrm{pre},r} = \frac{1}{N}\bar{\mathbf{H}}_r^\top \mathbf{D}\mathbf{J}_N\mathbf{D}\bar{\mathbf{H}}_r. \tag{19}$$

Hence degree-dependent gain enters the pre-MLP geometry and can alter both signal retention and modal placement; see Appendix D.4.3. These two aggregation mechanisms induce distinct TGP shifts, which we test next through profile diagnostics and controlled interventions.

4 Experiments

4.1 Experimental Setup

We evaluate six representative graph SSL objectives spanning distinct pretraining paradigms: reconstruction-based GraphMAE (Hou et al., 2022) and GAE (Kipf & Welling, 2016), contrastive GRACE (Zhu et al., 2020) and GCA (Zhu et al., 2021), bootstrap-based BGRL (Thakoor et al., 2022), and correlation-based CCA-SSG (Zhang et al., 2021). This selection tests whether utility-transfer failures persist across qualitatively different self-supervision signals.

Each objective is instantiated with both GCN and GIN and evaluated on Cora, CiteSeer, PubMed, Amazon Photo, and Amazon Computers. For every target, source-pretrained and target-matched encoders share the objective, backbone, split, and label budget, with the target excluded from source pretraining. Frozen representations are evaluated by Linear and MLP probes using Δ_C, Δ_A, and A_G. Full protocols are given in Appendix C.

Figure 2 summarizes all 120 settings. $A_G > 0$ in $110/120$ cases and $\Delta_C > 0$ in $113/120$, showing widespread incomplete retention of target-matched utility. The accompanying Δ_A varies substantially, separating capability-dominated, accessibility-limited, and mixed transfer regimes. Taken together, these results establish cross-domain transfer failure as a systematic but heterogeneous phenomenon, motivating a mechanism-level analysis of how capability retention and finite-supervision accessibility arise from representation geometry.

4.2 Task–Geometry Theory Audit

We test the two channels predicted by Theorem 1 using vanilla source- and target-pretrained representations. Capability is audited by comparing $\Delta_\Psi = \Psi_t - \Psi_s$ with empirical Δ_C. For accessibility, we define

$$\Delta_A^{\mathrm{TG}} = \mathrm{Avg}_{\log n}\left[\mathcal{A}_s^{\mathrm{TG}}(n) - \mathcal{A}_t^{\mathrm{TG}}(n)\right] - \left[\mathcal{A}_s^{\mathrm{TG}}(n_{\max}) - \mathcal{A}_t^{\mathrm{TG}}(n_{\max})\right], \tag{20}$$

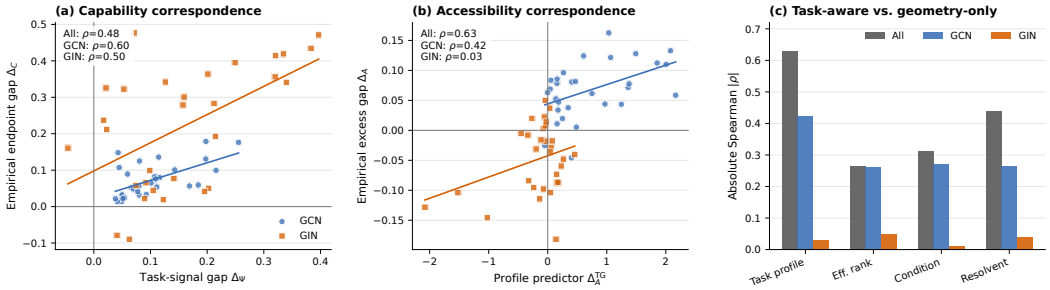

Figure 3: **Task–Geometry theory audit.** (a) Retained-signal gap Δ_Ψ versus empirical capability loss Δ_C, testing the predicted capability–signal relation. (b) Task–Geometry accessibility predictor Δ_A^{TG} versus empirical Δ_A, testing whether modal signal placement predicts finite-supervision degradation. (c) Spearman correspondence of the task-aware predictor and task-agnostic geometry statistics with empirical accessibility loss. Positive correspondence in (a)–(b), together with the stronger task-aware association in (c), supports the two channels of the Task–Geometry decomposition.

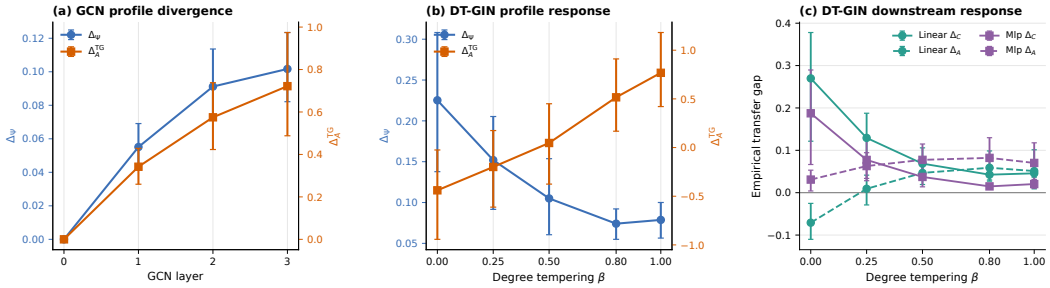

Figure 4: **Aggregation mechanisms and Task–Geometry responses.** (a) Layer-wise GCN propagation progressively enlarges source–target profile separation. (b) Increasing the GIN degree-tempering strength β suppresses degree-dependent gain and shifts the Task–Geometry response. (c) The corresponding empirical decomposition shows the resulting capability–accessibility trade-off: capability improves while finite-supervision accessibility can worsen. Error bars show variation across audited settings.

where $\mathcal{A}_r^{\mathrm{TG}}(n) = \sum_i \psi_{r,i}[1 - a_{r,i}(n)]^2 + \frac{\sigma_r^2}{n}\sum_i a_{r,i}(n)^2$, and $\mathrm{Avg}_{\log n}$ averages uniformly over the log-budget scale. The endpoint correction matches the empirical definition of Δ_A. Because the spectral ridge model is a representation-level surrogate for classification learning curves, we evaluate correspondence by sign agreement and Spearman correlation. Linear probing is the closest audit of the theory; MLP probing tests robustness to nonlinear readout.

Figure 3(a) validates the signal-quantity channel: Δ_Ψ correlates positively with Linear-probe Δ_C overall ($\rho = 0.481$), within GCN (0.602), and within GIN (0.496), with sign agreements of 0.950, 1.000, and 0.900, respectively. This supports the capability–signal relation in Proposition 1. Figure 3(b) evaluates signal–geometry resolution. Δ_A^{TG} correlates with empirical Δ_A overall ($\rho = 0.629$) and within GCN (0.422), but weakly within GIN (0.031), marking a boundary of the linear surrogate under GIN's nonlinear degree-dependent transformations. The MLP landscape remains consistent overall ($\rho = 0.407$ for capability and 0.622 for accessibility); see Appendix F for the complete nonlinear-probe analysis. Figure 3(c) further shows stronger correspondence for Δ_A^{TG} than for effective rank, condition number, or resolvent trace, highlighting the role of task-aware signal placement.

4.3 Aggregation Mechanisms and Capability–Accessibility Trade-offs

For GCN, we use three task-agnostic geometry controls. Resolvent Regularization (RR), Barlow Redundancy Objective (BRO), and Random-Direction Distortion (RDR) respectively target weak

Table 1: **Controlled aggregation interventions.** *Left:* GCN task-agnostic geometry controls and their induced Task–Geometry and empirical transfer responses. *Right:* DT-GIN at $\beta = 0.8$; positive $\delta^{C/A}_{\text{close}}$ denotes transfer-gap closure. Values in parentheses are 95% confidence intervals.

Method	Δ_A^{TG}	Linear		MLP	
		Δ_C	Δ_A	Δ_C	Δ_A
Vanilla	0.722	0.0655	0.0651	0.0279	0.0973
RR	0.990	0.0487	0.0647	0.0220	0.0869
BRO	0.930	0.0571	0.0601	0.0243	0.0868
RDR	0.732	0.0649	0.0634	0.0273	0.0963

Metric	Linear	MLP
δ_s^C	0.253 (0.212–0.298)	0.236 (0.195–0.279)
δ_t^C	0.123 (0.088–0.161)	0.103 (0.075–0.133)
δ^C_{close}	0.130 (0.075–0.184)	0.133 (0.081–0.185)
δ^A_{close}	-0.092 (-0.119–-0.066)	-0.082 (-0.106–-0.056)

covariance modes, channel redundancy, and directional second-order distortion, and are added as $\mathcal{L}_{\text{SSL}}+\lambda_{\text{geo}}\mathcal{R}_{\text{geo}}$; see Appendix E for details. Let $\{\mu_i\}$ be the trace-normalized covariance spectrum, $\mathbf{C}_H$ the standardized channel correlation matrix, and $v_k = \mathbf{u}_k^\top \mathbf{\Sigma}_H \mathbf{u}_k$, where $\{\mathbf{u}_k\}_{k=1}^K$ are random unit directions and v_k^{sub} is defined analogously on a random node subset. Here γ is the resolvent shift, ϵ_{num} a numerical stabilizer, τ_{floor} the variance floor, and $\alpha_{\text{RR}}, \alpha_{\text{floor}}$ weighting coefficients:

$$\mathcal{R}_{\text{RR}} = \frac{1+\gamma}{d}\sum_i (\mu_i+\gamma)^{-1} - 1 + \alpha_{\text{RR}}\left[\frac{(1+\gamma)^2}{d}\sum_i(\mu_i+\gamma)^{-2} - 1\right]. \tag{21a}$$

$$\mathcal{R}_{\text{BRO}} = d^{-2}\sum_{i\neq j}(\mathbf{C}_H)_{ij}^2, \quad \mathcal{R}_{\text{RDR}}^{\text{total}} = K^{-1}\sum_k \left|\log\frac{v_k^{\text{sub}}+\epsilon_{\text{num}}}{v_k+\epsilon_{\text{num}}}\right| + \alpha_{\text{floor}}K^{-1}\sum_k[\tau_{\text{floor}} - v_k]_+^2. \tag{21b}$$

For GIN, we directly control degree-dependent amplification through

$$\mathbf{m}_{r,v}^{(\beta)} = (d_v+1)^{1-\beta}\bar{\mathbf{h}}_{r,v}, \qquad \beta \in [0,1], \tag{22}$$

where $\beta = 0$ recovers standard GIN and $\beta = 1$ removes the degree-dependent gain; we call this variant *Degree-Tempered GIN* (DT-GIN).

GCN: normalized propagation amplifies profile separation. Figure 4(a) shows cumulative profile distortion, with $\Delta_\Psi = 0.102$ and $\Delta_A^{\text{TG}} = 0.722$ by layer 3. Table 1 shows that RR and BRO modestly reduce empirical gaps despite increasing the task-aware predictor, whereas RDR changes little, consistent with Proposition 2.

GIN: degree tempering exposes a capability–accessibility trade-off. At $\beta = 0.8$, degree–amplitude dependence decreases in all $9/9$ audited settings. Figures 4(b)–(c) show that DT-GIN reduces Δ_Ψ and empirical Δ_C, while Δ_A^{TG} and Δ_A increase. Correspondingly, A_G decreases from 0.199 to 0.102 for Linear and from 0.218 to 0.097 for MLP. Table 1 therefore shows substantial capability-gap recovery without parallel accessibility improvement. The net reduction in A_G demonstrates that architecture-level control can improve transferable utility even when the two transfer components move in opposite directions.

4.4 Large-Scale Out-of-Suite Validation

We further evaluate whether the transfer mechanisms identified on the classical benchmark suite persist at larger graph scale. We use OGBN-ARXIV (Hu et al., 2020a) as an out-of-suite target and consider three representative SSL objectives: GraphMAE, BGRL, and GRACE. We retain the same source-pretrained and target-matched evaluation protocol as in the main experiments and evaluate both Linear and MLP probes. For GCN, we compare vanilla propagation with the RR geometry control; for GIN, we compare standard aggregation with DT-GIN degree tempering.

GCN: RR primarily recovers source capability. Across the three SSL objectives, RR improves the source-pretrained representation while leaving target-matched performance nearly unchanged. Accordingly, the mean capability gap Δ_C decreases from 0.042 to 0.002 for Linear and from 0.039 to 0.001 for MLP, with corresponding reductions in A_G. The Task–Geometry response separates

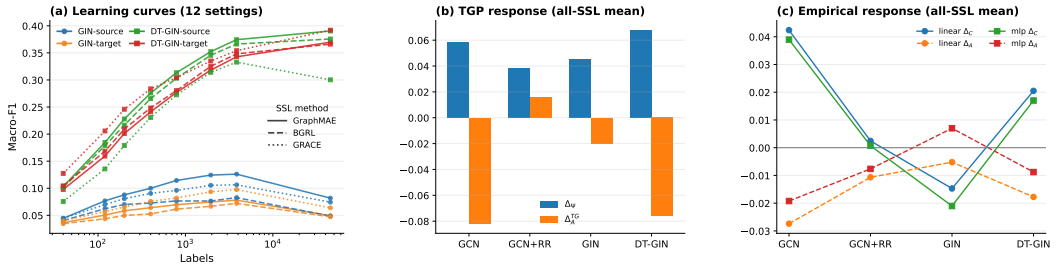

Figure 5: **Large-scale out-of-suite validation on OGBN-ARXIV.** (a)–(c) Linear-probe learning curves for GraphMAE, BGRL, and GRACE, comparing vanilla GIN with degree-tempered DT-GIN. (d) Mean Task–Geometry response to RR for GCN and DT-GIN for GIN. (e) Corresponding empirical capability and finite-supervision components under Linear and MLP probes.

this gain from accessibility: RR reduces the retained-signal gap, whereas the accessibility-side profile does not improve consistently. Thus, its large-scale transfer benefit is dominated by capability recovery.

GIN: degree tempering strongly improves absolute utility. DT-GIN yields substantial gains across the supervision range and across reconstruction, bootstrap, and contrastive objectives. Averaged over the three SSL methods, source/target Macro-F1 increases by $0.287/0.322$ for Linear and by $0.253/0.291$ for MLP. These gains show that controlling degree-dependent amplification remains effective beyond the classical benchmark suite.

Importantly, absolute utility improvement does not imply uniform source–target gap closure. DT-GIN can improve both source and target representations while changing their relative advantage in an SSL-dependent manner, producing a capability–accessibility redistribution rather than a uniform reduction of every transfer component. The large-scale results therefore reinforce the distinction between absolute representation utility and cross-domain transfer retention, and motivate the joint use of Δ_C, Δ_A, and A_G. Complete transfer decompositions, Task–Geometry responses, and absolute downstream results are provided in Appendix G.

5 Conclusion

In this work, we re-evaluate cross-domain graph self-supervised learning through the lens of *attainable utility* established by target-matched pretraining. This perspective reveals that graph transfer failure is not a monolithic bottleneck, but rather decomposes into two fundamentally distinct mechanisms: *capability loss* (loss of total target-relevant information) and *finite-supervision accessibility loss* (geometric misallocation of retained information into low-variance modes). To unify these failure modes, we introduced the Task–Geometry Profile (TGP), a representation-level framework that models both signal retention and modal placement.

Through extensive empirical audits across diverse SSL objectives, graph domains, backbones, and probe families, we connected TGP diagnostics directly to observed transfer outcomes. Crucially, we demonstrated how message-passing architectures distort this profile—showing that GCN's normalized diffusion and GIN's degree-dependent accumulation alter signal retention and modal allocation through complementary inductive biases. Furthermore, controlled interventions reveal that capability and accessibility can decouple: targeted degree tempering substantially recovers GIN's representation capability while significantly mitigating the overall transfer deficit.

Our findings establish that effective transfer requires moving beyond simple information preservation toward *usable signal geometry*—retaining target-relevant task signals while ensuring they remain easily resolvable under limited target supervision. For graph foundation models, this implies that pretraining objectives, aggregation layers, and adaptation mechanisms must be jointly designed and evaluated through both capability ceilings and accessibility efficiency. The capability–accessibility paradigm offers a principled theoretical and practical blueprint for engineering graph architectures that achieve both strong and sample-efficient cross-domain transfer.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and for assistance in writing portions of the experimental code. We have reviewed all AI-assisted work. In particular, AI-assisted code was verified and tested by the authors, and AI-assisted text was manually reviewed and revised. We take responsibility for the final content of this work, including text, claims, experimental results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available graph benchmarks and does not involve human subjects, personally identifiable information, or sensitive data. We are not aware of additional ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

Complete theoretical proofs are provided in the Appendix. Experimental protocols, datasets, label budgets, evaluation procedures, hyperparameters, and controlled interventions are documented in the main paper and Appendix to support reproducibility.

REFERENCES

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Moshe Eliasof, Krishna Sri Ipsit Mantri, Beatrice Bevilacqua, Bruno Ribeiro, and Carola-Bibiane Schönlieb. Bridging input feature spaces towards graph foundation models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xiaodong He, Haolan He, Ruiyi Fang, Ming Sun, and Zhao Kang. Structure-centric graph foundation model via geometric bases. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 594–604, 2022.

Zhenyu Hou, Yufei He, Yukuo Cen, Xiao Liu, Yuxiao Dong, Evgeny Kharlamov, and Jie Tang. GraphMAE2: A decoding-enhanced masked self-supervised graph learner. In *Proceedings of the ACM Web Conference 2023*, pp. 737–746. Association for Computing Machinery, 2023.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, 2020a.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020b.

Thomas N. Kipf and Max Welling. Variational graph auto-encoders. In *Bayesian Deep Learning Workshop at NIPS 2016*, 2016.

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

Lecheng Kong, Jiarui Feng, Hao Liu, Chengsong Huang, Jiaxin Huang, Yixin Chen, and Muhan Zhang. GOFA: A generative one-for-all model for joint graph language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1): 3538–3545, 2018.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In *International Conference on Learning Representations*, 2024.

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S. Yu, and Chuan Shi. Towards graph foundation models: A survey and beyond. *arXiv preprint arXiv:2310.11829*, 2023a.

Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and Philip S. Yu. Graph self-supervised learning: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 5879–5900, 2023b.

Haitao Mao, Zhikai Chen, Wenzhuo Tang, Jianan Zhao, Yao Ma, Tong Zhao, Neil Shah, Mikhail Galkin, and Jiliang Tang. Position: Graph foundation models are already here. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 34670–34692. PMLR, 2024.

Sohir Maskey, Raffaele Paolino, Fabian Jogl, Gitta Kutyniok, and Johannes F. Lutzeyer. Graph representational learning: When does more expressivity hurt generalization? In *International Conference on Learning Representations*, 2026.

Dimitri Meunier, Zikai Shen, Mattes Mollenhauer, Arthur Gretton, and Zhu Li. Optimal rates for vector-valued spectral regularization learning algorithms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 82514–82559, 2024.

Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael Schaub. Residual connections and normalization can provably prevent oversmoothing in GNNs. In *The Thirteenth International Conference on Learning Representations*, 2025.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *Relational Representation Learning Workshop*, 2018.

Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L. Dyer, Rémi Munos, Petar Veličković, and Michal Valko. Large-scale representation learning on graphs via bootstrapping. In *International Conference on Learning Representations*, 2022.

Qin Tian, Chen Zhao, Minglai Shao, Wenjun Wang, Yujie Lin, and Dong Li. MLDGG: Meta-learning for domain generalization on graphs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 1361–1372. Association for Computing Machinery, 2025.

Yancheng Wang, Chengshuai Zhao, Dongfang Sun, Huan Liu, and Yingzhen Yang. Low-rank few-shot node classification by node-level graph diffusion. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zehong Wang, Zheyuan Zhang, Chuxu Zhang, and Yanfang Ye. Subgraph pooling: Tackling negative transfer on graphs. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5153–5161, 2024.

Xinxu Wei, Kanhao Zhao, Yong Jiao, Lifang He, and Yu Zhang. A brain graph foundation model: Pre-training and prompt-tuning across broad atlases and disorders. In *The Fourteenth International Conference on Learning Representations*, 2026.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *The Seventh International Conference on Learning Representations*, 2019.

Rui Xue and Tianfu Wu. HarmonyGNNs: Harmonizing heterophily and homophily in GNNs via self-supervised node encoding. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 40–48. PMLR, 2016.

Hengrui Zhang, Qitian Wu, Junchi Yan, David P. Wipf, and Philip S. Yu. From canonical correlation analysis to self-supervised graph neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. In *ICML Workshop on Graph Representation Learning and Beyond*, 2020.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *Proceedings of The Web Conference 2021*, pp. 2069–2080, 2021.

A Additional Related Work

Graph Self-Supervised Learning and Foundation Models. Graph self-supervised learning learns reusable representations through contrastive, predictive, and reconstruction-based objectives. Recent methods further advance masked reconstruction and teacher–student learning (Hou et al., 2023; Xue & Wu, 2026). Graph foundation models extend this objective toward reuse across heterogeneous graphs. GOFA integrates graph-aware message passing with generative language modeling (Kong et al., 2025), SCGFM constructs shared structural coordinates through geometric bases (He et al., 2026), and ALL-IN addresses incompatible feature spaces through random projections and covariance-based operators (Eliasof et al., 2026). Our work instead asks how much target-matched SSL utility survives cross-domain pretraining and why the remaining utility can be difficult to exploit.

Graph Transfer under Distribution Shift. Prior work addresses graph transfer under source–target mismatch through structural pooling (Wang et al., 2024), domain generalization (Tian et al., 2025), low-rank few-shot diffusion (Wang et al., 2026), and prompting with meta-learning (Wei et al., 2026). Our focus is complementary: we separate transfer degradation into capability and finite-supervision accessibility and study their representation-level origins.

Graph Inductive Bias and Representation Geometry. GCN propagation induces Laplacian smoothing (Li et al., 2018), whereas GIN uses injective sum aggregation to preserve multiset information. Prior analyses study eigenspace convergence under residual connections and normalization (Scholkemper et al., 2025), finite-sample spectral estimation (Meunier et al., 2024), and the interaction between GNN expressivity and structural generalization (Maskey et al., 2026). We instead connect graph-induced geometry with target-task signal to explain both retained capability and finite-supervision resolvability.

B Notation

Table 2 summarizes the principal notation used throughout the paper.

C Experimental Protocol and Implementation Details

C.1 Hardware and execution environment

All experiments are executed in a fixed Conda environment using PyTorch and PyTorch Geometric. Training runs use one NVIDIA A800 GPU with 80 GB of memory. Data preprocessing, result aggregation, and CPU-side statistics use an Intel(R) Xeon(R) Gold 6348 CPU at 2.60 GHz. Each reported pretraining seed is an independent run; downstream probe initialization is paired across the source-pretrained and target-matched representations. Pretraining uses BF16 automatic mixed precision on the A800; covariance, profile, and statistical computations are performed in FP32 or FP64 as appropriate.

Table 2: Principal notation.

Symbol	Meaning
Graphs, representations, and evaluation	
$\mathcal{G} = (V, E, \mathbf{X})$	Graph with node set V, edge set E, and node features $\mathbf{X}$
$N,\ d_f,\ d$	Number of nodes, input-feature dimension, and representation dimension
$r \in \{0, s, t\}$	Encoder role: random initialization, source-pretrained, or target-matched
$f_r,\ \mathbf{Z}_r$	Encoder of role r and its target-node representation matrix
$n,\ n_{\min}, n_{\max}$	Total number of labeled target nodes and its evaluation range
$L_r(n)$	Downstream performance of representation role r at budget n
Utility-transfer quantities	
$G(n)$	Utility-transfer deficit, $L_t(n) - L_s(n)$
A_G	Average utility-transfer deficit over the log-budget scale
Δ_C	Endpoint capability loss, $G(n_{\max})$
Δ_A	Endpoint-relative finite-supervision accessibility loss, $A_G - \Delta_C$
Task–Geometry quantities	
$\mathbf{z}_r,\ y$	Centered target-node representation and centered target response
$\boldsymbol{\beta}_r^\star$	Population-optimal linear prediction coefficient
$\xi_r,\ \sigma_r^2$	Population residual and its variance
$\boldsymbol{\Sigma}_r = \mathbf{Q}_r \boldsymbol{\Lambda}_r \mathbf{Q}_r^\top$	Representation covariance and eigendecomposition
$\lambda_{r,i},\ \mathbf{q}_{r,i}$	Strength and eigenvector of representation mode i
$b_{r,i}$	Task coefficient along mode $\mathbf{q}_{r,i}$
$g_{r,i}(n)$	Finite-supervision spectral score along mode i
$\zeta_{r,i}(n)$	Zero-mean finite-supervision score noise
$\widehat{b}_{r,i}(n),\ \widehat{\boldsymbol{\beta}}_{r,n}$	Mode-wise ridge estimate and reconstructed predictor
$\psi_{r,i}$	Target-task signal carried by mode i, $\psi_{r,i} = \lambda_{r,i} b_{r,i}^2$
$\mathcal{P}_r$	Task–Geometry Profile $\{(\lambda_{r,i}, \psi_{r,i})\}_{i=1}^d$
Ψ_r	Total retained target-task signal, $\sum_i \psi_{r,i}$
η	Ridge regularization parameter
$a_{r,i}(n)$	Finite-supervision resolution of mode i
$\mathcal{U}_r(n)$	Unresolved target-task signal
$\mathcal{V}_r(n)$	Finite-sample estimation variance
$\mathcal{A}_r(n)$	Theoretical accessibility cost, $\mathcal{U}_r(n) + \mathcal{V}_r(n)$
$R_r^\star,\ R_{r,n}$	Population-optimal and finite-supervision prediction risks
Aggregation quantities	
$\mathbf{H}_r^{(\ell)}$	Target-node states entering message-passing layer ℓ
$\mathbf{A}_{\text{agg}}$	Backbone-specific aggregation operator
$\mathbf{J}_N$	Node-centering matrix, $\mathbf{I}_N - N^{-1}\mathbf{1}\mathbf{1}^\top$
$\mathbf{P} = \mathbf{V}\boldsymbol{\Omega}\mathbf{V}^\top$	Normalized GCN propagation operator
ω_k	Response of normalized propagation along graph mode k
d_v	Degree of target node v
$\bar{\mathbf{h}}_{r,v},\ \bar{\mathbf{H}}_r$	Degree-normalized local content and its stacked matrix
β	Degree-tempering strength in DT-GIN

C.2 Classical cross-domain protocol

The classical suite contains Cora, CiteSeer, PubMed, Amazon-Photo, and Amazon-Computers. For each target graph, source pretraining uses the remaining four graphs, whereas target-matched pretraining uses only the unlabeled target graph. All tasks are transductive node classification: the representation matrix is $\mathbf{Z}_r \in \mathbb{R}^{N \times d}$ for the N target nodes, and no graph-level pooling is applied. Input features are row-ℓ_1 normalized and mapped by a fixed Gaussian projection to $d = 256$. The same cached projection is reused by every SSL objective, backbone, role, and seed.

We use three-layer encoders with hidden dimension 256, dropout 0.2, and BatchNorm and PReLU after each message-passing layer. Models are pretrained for 200 epochs with seeds $\{0, 1, 2\}$. The

Table 3: SSL optimization settings used in the classical suite.

SSL	Optimizer	Learning rate	Weight decay
GraphMAE	Adam	10^{-3}	10^{-5}
GAE	Adam	10^{-2}	0
GRACE	Adam	10^{-3}	10^{-5}
GCA	Adam	10^{-2}	0
BGRL	AdamW	5×10^{-4}	10^{-5}
CCA-SSG	Adam	10^{-3}	0

label budgets are specified per class as $b \in \{1, 3, 5, 10, 20, 50, 100, \text{full}\}$. For a target with C classes, n denotes the corresponding total number of labeled training nodes; under the balanced finite-budget splits, $n = Cb$. Five deterministic class-stratified splits are used; each split reserves 10% of the nodes in every class for validation and 20% for testing. All budgets within a split are nested. Linear and one-hidden-layer MLP probes are trained on frozen representations, with paired initialization and identical train/validation/test nodes for the source and target roles. The MLP uses a 256-dimensional hidden layer, ReLU, dropout 0.5, Adam with learning rate 10^{-3}, and weight decay 10^{-4}. Early stopping selects the checkpoint with the best validation Macro-F1.

GraphMAE uses mask rate 0.5, replacement rate 0.1, and scaled-cosine exponent 2. GRACE uses edge-drop rates 0.2/0.4, feature-drop rates 0.3/0.2, and temperature 0.5. BGRL uses edge-drop rates 0.5/0.4, feature-drop rates 0.2/0.1, base momentum 0.99, and a 20-epoch learning-rate warmup. GCA uses degree-adaptive edge dropping with rates 0.3/0.5, feature-drop rates 0.1/0.2, and temperature 0.3. CCA-SSG uses edge-drop rate 0.4, feature-drop rate 0.2, and decorrelation weight 10^{-3}.

The Task–Geometry audit uses two-fold class-stratified cross-fitting inside the probe training pool. One half estimates the task signal and the other trains the probe; the roles are then exchanged and averaged. We use $\tau = 10^{-3}\bar{\lambda}$ when estimating the profile and $\eta = 0.1\bar{\lambda}$ for finite-supervision resolution, where $\bar{\lambda}$ is the mean representation-covariance eigenvalue. Bootstrap resampling is performed over the complete SSL–backbone–target setting, not over budgets, seeds, or cross-fit folds as if they were independent settings.

For controlled GCN interventions, RR uses weight 0.1, resolvent shift 0.1, and second-moment weight 0.1. BRO uses weight 0.1 and numerical stabilizer 10^{-6}. RDR uses eight random directions, a 0.25 node-sampling ratio, directional and floor weights 0.1, and floor threshold 0.01. DT-GIN uses the degree-tempered message $(d_v+1)^{-\beta}[(1+\epsilon_{\text{GIN}})\mathbf{h}_v + \sum_{u\in\mathcal{N}(v)} \mathbf{h}_u]$; the broad comparison fixes $\beta = 0.8$.

C.3 Datasets

We evaluate graph self-supervised transfer on six node-classification benchmarks, including citation networks, co-purchase networks, and a large-scale academic citation network. Cora, CiteSeer, and PubMed are citation networks from the Planetoid benchmark (Yang et al., 2016), where nodes represent papers and edges denote citation relationships. Amazon Photo and Amazon Computers are co-purchase networks (Shchur et al., 2018), where nodes represent products and edges denote co-purchase relationships. We additionally use OGBN-Arxiv from the Open Graph Benchmark (Hu et al., 2020a) for out-of-suite validation, which is a large-scale citation network where nodes represent arXiv papers and edges denote citation relationships.

Table 4 summarizes the statistics of all datasets used in our experiments.

D Proof

D.1 Proof of the Empirical Learning-Curve Identity

The main text defines the utility-transfer deficit

$$G(n) = L_t(n) - L_s(n), \tag{23}$$

Table 4: Statistics of datasets used in experiments.

Dataset	Nodes	Edges	Features	Classes
Cora	2,708	5,429	1,433	7
CiteSeer	3,327	4,732	3,703	6
PubMed	19,717	44,338	500	3
Amazon Photo	7,650	119,081	745	8
Amazon Computers	13,752	245,861	767	10
OGBN-Arxiv	169,343	1,166,243	128	40

the endpoint capability component

$$\Delta_C = G(n_{\max}), \tag{24}$$

the log-budget average deficit

$$A_G = \frac{1}{\ell}\int_{n_{\min}}^{n_{\max}} G(n)\, d(logn), \qquad \ell = \log n_{\max} - \log n_{\min}, \tag{25}$$

and the endpoint-relative finite-supervision component

$$\Delta_A = \frac{1}{\ell}\int_{n_{\min}}^{n_{\max}} \big[G(n) - G(n_{\max})\big]\, d(logn). \tag{26}$$

For the quantities in Eqs. 24, 25 and 26,

$$A_G = \Delta_C + \Delta_A. \tag{27}$$

Proof. Because $G(n_{\max})$ is constant with respect to n,

$$\Delta_A = \frac{1}{\ell}\int_{n_{\min}}^{n_{\max}} G(n)\, d(logn) - \frac{G(n_{\max})}{\ell}\int_{n_{\min}}^{n_{\max}} d(logn) \tag{28}$$

$$= A_G - G(n_{\max}) = A_G - \Delta_C. \tag{29}$$

Rearranging proves Eq. 27. The argument holds for any scalar performance metric, including accuracy and Macro-F1. □

D.2 Proofs for the Task–Geometry Theory

D.2.1 Population projection and profile notation

Fix a representation role $r \in \{s, t\}$. Let $\mathbf{z}_r \in \mathbb{R}^d$ be a centered target-node representation and let $y \in \mathbb{R}$ be the centered target. The population-optimal linear coefficient is

$$\boldsymbol{\beta}_r^\star \in \arg\min_{\boldsymbol{\beta}} \mathbb{E}[(y - \mathbf{z}_r^\top \boldsymbol{\beta})^2]. \tag{30}$$

Define

$$\xi_r = y - \mathbf{z}_r^\top \boldsymbol{\beta}_r^\star, \qquad \sigma_r^2 = \mathbb{E}[\xi_r^2], \qquad R_r^\star = \sigma_r^2. \tag{31}$$

The first-order condition for Eq. 30 is

$$\mathbb{E}[\mathbf{z}_r \xi_r] = \mathbf{0}. \tag{32}$$

Let

$$\boldsymbol{\Sigma}_r = \mathbb{E}[\mathbf{z}_r \mathbf{z}_r^\top] = \mathbf{Q}_r \boldsymbol{\Lambda}_r \mathbf{Q}_r^\top, \qquad \boldsymbol{\Lambda}_r = \mathrm{diag}(\lambda_{r,1}, \ldots, \lambda_{r,d}), \tag{33}$$

and write

$$\boldsymbol{\beta}_r^\star = \sum_{i=1}^{d} b_{r,i} \mathbf{q}_{r,i}. \tag{34}$$

The Task–Geometry Profile and its total signal are

$$\psi_{r,i} = \lambda_{r,i} b_{r,i}^2, \qquad \mathcal{P}_r = \{(\lambda_{r,i}, \psi_{r,i})\}_{i=1}^d, \qquad \Psi_r = \sum_{i=1}^{d} \psi_{r,i}. \tag{35}$$

D.2.2 CAPABILITY–SIGNAL IDENTITY

Proposition 1 (Capability–Signal Identity)**.** *For the population-optimal predictor in Eq. 30,*

$$\mathrm{Var}(y) = \Psi_r + R_r^\star. \tag{36}$$

Consequently, for source-pretrained and target-matched representations of the same target task,

$$R_s^\star - R_t^\star = \Psi_t - \Psi_s. \tag{37}$$

Proof. Using $y = \mathbf{z}_r^\top \boldsymbol{\beta}_r^\star + \xi_r$ and the orthogonality condition in Eq. 32,

$$\begin{aligned}
\mathrm{Var}(y) &= \mathbb{E}[(\mathbf{z}_r^\top \boldsymbol{\beta}_r^\star + \xi_r)^2] && (38)\\
&= \boldsymbol{\beta}_r^{\star\top} \boldsymbol{\Sigma}_r \boldsymbol{\beta}_r^\star + 2\boldsymbol{\beta}_r^{\star\top} \mathbb{E}[\mathbf{z}_r \xi_r] + \mathbb{E}[\xi_r^2] && (39)\\
&= \boldsymbol{\beta}_r^{\star\top} \boldsymbol{\Sigma}_r \boldsymbol{\beta}_r^\star + \sigma_r^2. && (40)
\end{aligned}$$

Expanding the quadratic form in the covariance eigenbasis gives

$$\boldsymbol{\beta}_r^{\star\top} \boldsymbol{\Sigma}_r \boldsymbol{\beta}_r^\star = \sum_{i=1}^{d} \lambda_{r,i} b_{r,i}^2 = \sum_{i=1}^{d} \psi_{r,i} = \Psi_r. \tag{41}$$

Since $R_r^\star = \sigma_r^2$, this proves Eq. 36. The target-label variance is the same for $r = s$ and $r = t$. Subtracting the two identities proves Eq. 37. □

D.2.3 FINITE-SUPERVISION MODE RESOLUTION

We formalize finite-supervision resolution directly in the population covariance eigenbasis. This construction isolates the effect of mode strength from finite-sample perturbations of the empirical covariance eigenspace.

For representation role r, let

$$g_{r,i}(n) = n\lambda_{r,i} b_{r,i} + \zeta_{r,i}(n), \tag{42}$$

where

$$\mathbb{E}[\zeta_{r,i}(n)] = 0, \qquad \mathbb{E}[\zeta_{r,i}(n)\zeta_{r,j}(n)] = n\sigma_r^2 \lambda_{r,i} \delta_{ij}. \tag{43}$$

The $n\lambda_{r,i}$ factor represents the supervision accumulated along population mode $\mathbf{q}_{r,i}$, while the second moment in Eq. 43 gives the corresponding homoskedastic residual-score variance.

Define

$$\widehat{b}_{r,i}(n) = \frac{g_{r,i}(n)}{n\lambda_{r,i} + \eta}, \qquad a_{r,i}(n) = \frac{n\lambda_{r,i}}{n\lambda_{r,i} + \eta}, \qquad \eta > 0. \tag{44}$$

Then

$$\widehat{b}_{r,i}(n) = a_{r,i}(n) b_{r,i} + \frac{\zeta_{r,i}(n)}{n\lambda_{r,i} + \eta}, \tag{45}$$

and therefore

$$\mathbb{E}[\widehat{b}_{r,i}(n)] = a_{r,i}(n) b_{r,i}. \tag{46}$$

Moreover,

$$\mathrm{Var}[\widehat{b}_{r,i}(n)] = \frac{n\sigma_r^2 \lambda_{r,i}}{(n\lambda_{r,i} + \eta)^2}. \tag{47}$$

For $\lambda_{r,i} > 0$,

$$\frac{\partial a_{r,i}(n)}{\partial n} = \frac{\lambda_{r,i}\eta}{(n\lambda_{r,i} + \eta)^2} > 0, \qquad \frac{\partial a_{r,i}(n)}{\partial \lambda_{r,i}} = \frac{n\eta}{(n\lambda_{r,i} + \eta)^2} > 0. \tag{48}$$

Thus stronger representation modes are resolved more rapidly as target supervision increases.

D.2.4 TASK–GEOMETRY TRANSFER DECOMPOSITION

We now derive the finite-supervision prediction risk under the spectral ridge surrogate introduced in Appendix D.2.3.

Let $(\mathbf{z}_{r,0}, y_0)$ be an independent target test sample satisfying

$$y_0 = \mathbf{z}_{r,0}^\top \boldsymbol{\beta}_r^\star + \xi_{r,0}, \qquad \mathbb{E}[\mathbf{z}_{r,0}\mathbf{z}_{r,0}^\top] = \boldsymbol{\Sigma}_r, \tag{49}$$

with

$$\mathbb{E}[\mathbf{z}_{r,0}\xi_{r,0}] = \mathbf{0}, \qquad \mathbb{E}[\xi_{r,0}^2] = \sigma_r^2. \tag{50}$$

The test sample is independent of the finite-supervision score noise $\{\zeta_{r,i}(n)\}_{i=1}^d$.

Define

$$\widehat{\boldsymbol{\beta}}_{r,n} = \sum_{i=1}^d \widehat{b}_{r,i}(n)\mathbf{q}_{r,i}, \tag{51}$$

and

$$R_{r,n} = \mathbb{E}\left[\left(y_0 - \mathbf{z}_{r,0}^\top \widehat{\boldsymbol{\beta}}_{r,n}\right)^2\right]. \tag{52}$$

As in the main text, define

$$\mathcal{U}_r(n) = \sum_{i=1}^d \psi_{r,i}[1 - a_{r,i}(n)]^2, \tag{53}$$

$$\mathcal{V}_r(n) = \frac{\sigma_r^2}{n}\sum_{i=1}^d a_{r,i}(n)^2, \tag{54}$$

$$\mathcal{A}_r(n) = \mathcal{U}_r(n) + \mathcal{V}_r(n). \tag{55}$$

Theorem 1 (Task–Geometry Transfer Decomposition)**.** *Under the spectral score model in Eqs. 42–43 and the independent-test convention in Eqs. 49–50, the expected target prediction risk satisfies*

$$R_{r,n} = R_r^\star + \mathcal{A}_r(n). \tag{56}$$

Consequently,

$$R_{s,n} - R_{t,n} = \underbrace{\Psi_t - \Psi_s}_{\textit{capability loss}} + \underbrace{\mathcal{A}_s(n) - \mathcal{A}_t(n)}_{\textit{finite-supervision accessibility loss}}. \tag{57}$$

Proof. Fix a representation role r and suppress the role subscript. Using $y_0 = \mathbf{z}_0^\top \boldsymbol{\beta}^\star + \xi_0$,

$$y_0 - \mathbf{z}_0^\top \widehat{\boldsymbol{\beta}}_n = \xi_0 - \mathbf{z}_0^\top(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star). \tag{58}$$

Because the test residual is orthogonal to $\mathbf{z}_0$ and the test sample is independent of the finite-supervision score noise,

$$R_n - R^\star = \mathbb{E}\left[(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star)^\top \boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star)\right]. \tag{59}$$

Since $\boldsymbol{\Sigma} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top$ and $\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star = \sum_i (\widehat{b}_i - b_i)\mathbf{q}_i$,

$$R_n - R^\star = \sum_{i=1}^d \lambda_i \mathbb{E}\left[(\widehat{b}_i - b_i)^2\right]. \tag{60}$$

From Eq. 45,

$$\widehat{b}_i - b_i = -[1 - a_i(n)]b_i + \frac{\zeta_i(n)}{n\lambda_i + \eta}. \tag{61}$$

The score noise has zero mean, so the bias–noise cross term vanishes. The squared prediction bias contributed by mode i is

$$\lambda_i b_i^2[1 - a_i(n)]^2 = \psi_i[1 - a_i(n)]^2. \tag{62}$$

Summing over modes yields $\mathcal{U}(n)$.

For the estimation-variance term, Eq. 43 gives

$$\mathrm{Var}[\zeta_i(n)] = n\sigma^2\lambda_i. \tag{63}$$

Therefore the prediction variance contributed by mode i is

$$\lambda_i \mathrm{Var}[\widehat{b}_i(n)] = \lambda_i \frac{n\sigma^2\lambda_i}{(n\lambda_i+\eta)^2} \tag{64}$$

$$= \frac{\sigma^2}{n}\left(\frac{n\lambda_i}{n\lambda_i+\eta}\right)^2 = \frac{\sigma^2}{n}a_i(n)^2. \tag{65}$$

Summing over modes yields $\mathcal{V}(n)$. Hence

$$R_n = R^\star + \mathcal{U}(n) + \mathcal{V}(n) = R^\star + \mathcal{A}(n), \tag{66}$$

which proves Eq. 56.

Applying this identity to $r = s$ and $r = t$ gives

$$R_{s,n} - R_{t,n} = (R_s^\star - R_t^\star) + \mathcal{A}_s(n) - \mathcal{A}_t(n). \tag{67}$$

By Proposition 1, $R_s^\star - R_t^\star = \Psi_t - \Psi_s$. Substitution proves Eq. 57. □

D.2.5 PROOFS OF COROLLARIES

Corollary 1 (Large-Sample Separation). *For fixed representation dimension d and fixed $\eta > 0$,*

$$\lim_{n\to\infty} \mathcal{A}_r(n) = 0, \qquad \lim_{n\to\infty}(R_{s,n} - R_{t,n}) = \Psi_t - \Psi_s. \tag{68}$$

Proof. If $\lambda_{r,i} > 0$, then

$$1 - a_{r,i}(n) = \frac{\eta}{n\lambda_{r,i}+\eta} \longrightarrow 0. \tag{69}$$

If $\lambda_{r,i} = 0$, then $\psi_{r,i} = \lambda_{r,i}b_{r,i}^2 = 0$, so that mode contributes zero to $\mathcal{U}_r(n)$. Because d is fixed, $\mathcal{U}_r(n) \to 0$. Also $0 \le a_{r,i}(n) \le 1$, so

$$0 \le \mathcal{V}_r(n) \le \frac{\sigma_r^2 d}{n} \longrightarrow 0. \tag{70}$$

Therefore $\mathcal{A}_r(n) \to 0$. Taking limits in Eq. 57 proves the second claim. □

Corollary 2 (Pure Accessibility Degradation; main-text corollary). *If $\Psi_s = \Psi_t$, then $R_s^\star = R_t^\star$ and*

$$R_{s,n} - R_{t,n} = \mathcal{A}_s(n) - \mathcal{A}_t(n). \tag{71}$$

In particular, $\mathcal{A}_s(n) > \mathcal{A}_t(n)$ implies $R_{s,n} > R_{t,n}$.

Proof. By Proposition 1, $\Psi_s = \Psi_t$ implies $R_s^\star = R_t^\star$. Setting the capability term to zero in Eq. 57 proves Eq. 71 and its stated implication. □

Corollary 3 (Signal Relocation; main-text corollary). *Consider a fixed spectrum and move an amount δ of task signal from mode i to a weaker positive mode j:*

$$\psi_i' = \psi_i - \delta, \qquad \psi_j' = \psi_j + \delta, \qquad 0 < \delta \le \psi_i, \qquad 0 < \lambda_j < \lambda_i. \tag{72}$$

All other ψ_k are unchanged. Then $\Psi' = \Psi$, $R^{\star\prime} = R^\star$, and for every finite n,

$$\mathcal{U}'(n) - \mathcal{U}(n) = \delta\left[\left(\frac{\eta}{n\lambda_j+\eta}\right)^2 - \left(\frac{\eta}{n\lambda_i+\eta}\right)^2\right] > 0. \tag{73}$$

Proof. The relocation preserves total signal because the decrease in mode i equals the increase in mode j. Therefore $R^{\star\prime} = R^{\star}$ by Proposition 1. The spectrum and residual variance are unchanged, so $\mathcal{V}'(n) = \mathcal{V}(n)$. Only the two displayed signal terms in $\mathcal{U}(n)$ change, yielding the equality in Eq. 73. The function

$$f_n(\lambda) = \left(\frac{\eta}{n\lambda + \eta}\right)^2 \tag{74}$$

is strictly decreasing because

$$f_n'(\lambda) = -\frac{2n\eta^2}{(n\lambda + \eta)^3} < 0. \tag{75}$$

Since $\lambda_j < \lambda_i$, the difference in Eq. 73 is positive. Since $\mathcal{V}'(n) = \mathcal{V}(n)$, the same increase also holds for $\mathcal{A}'(n) - \mathcal{A}(n)$. □

D.3 Task-Agnostic Geometry Ceiling

Proposition 2 (Task-Agnostic Geometry Ceiling; main-text proposition)**.** *Consider two geometries with eigenvalues $\{\lambda_i\}_{i=1}^d$ and $\{\lambda_i^*\}_{i=1}^d$, and denote their unresolved-signal costs by $\mathcal{U}$ and $\mathcal{U}^*$, respectively. For a fixed total task signal $\Psi > 0$, suppose there exists a mode j such that $0 < \lambda_j^* < \lambda_j$. Then there exists a valid task-signal allocation with $\psi_j = \Psi$ and $\psi_i = 0$ for $i \neq j$, for which, at every finite n,*

$$\mathcal{U}^*(n) > \mathcal{U}(n). \tag{76}$$

Proof. Allocate all signal to the weakened mode:

$$\psi_j = \Psi, \qquad \psi_i = 0 \quad (i \neq j). \tag{77}$$

This allocation is valid under both positive eigenvalues by taking $b_j = \sqrt{\Psi/\lambda_j}$ and $b_j^* = \sqrt{\Psi/\lambda_j^*}$. Using the decreasing function f_n from Eq. 74,

$$\mathcal{U}^*(n) - \mathcal{U}(n) = \Psi\big[f_n(\lambda_j^*) - f_n(\lambda_j)\big] > 0. \tag{78}$$

This proves Eq. 76. □

Zero-eigenvalue boundary The positivity condition makes the fixed-Ψ construction well defined. If $\lambda_j^* = 0$, the transformed mode cannot carry positive signal at all; that case produces capability loss rather than a pure fixed-capability comparison.

D.4 Proofs of the Aggregation Covariance Identities

D.4.1 Common aggregation covariance

Lemma 1 (Common aggregation covariance identity)**.** *Let $\mathbf{H}_r^{(\ell)} \in \mathbb{R}^{N \times d_\ell}$ be the input node states and let*

$$\mathbf{M}_r^{(\ell)} = \mathbf{A}_{\text{agg}} \mathbf{H}_r^{(\ell)}. \tag{79}$$

With $\mathbf{J}_N = \mathbf{I}_N - N^{-1}\mathbf{1}\mathbf{1}^\top$, the centered covariance is

$$\mathbf{\Sigma}_{\text{agg},r}^{(\ell)} = \frac{1}{N}(\mathbf{H}_r^{(\ell)})^\top \mathbf{A}_{\text{agg}}^\top \mathbf{J}_N \mathbf{A}_{\text{agg}} \mathbf{H}_r^{(\ell)}. \tag{80}$$

Proof. The centered node states are $\mathbf{J}_N \mathbf{M}_r^{(\ell)}$. Since $\mathbf{J}_N^\top = \mathbf{J}_N$ and $\mathbf{J}_N^2 = \mathbf{J}_N$,

$$\mathbf{\Sigma}_{\text{agg},r}^{(\ell)} = \frac{1}{N}(\mathbf{J}_N \mathbf{M}_r^{(\ell)})^\top (\mathbf{J}_N \mathbf{M}_r^{(\ell)}) \tag{81}$$

$$= \frac{1}{N}(\mathbf{M}_r^{(\ell)})^\top \mathbf{J}_N \mathbf{M}_r^{(\ell)} \tag{82}$$

$$= \frac{1}{N}(\mathbf{H}_r^{(\ell)})^\top \mathbf{A}_{\text{agg}}^\top \mathbf{J}_N \mathbf{A}_{\text{agg}} \mathbf{H}_r^{(\ell)}. \tag{83}$$

□

D.4.2 NORMALIZED DIFFUSION IN GCN

Proposition 3 (Linearized GCN covariance identity)**.** *Let* $\mathbf{P} = \mathbf{V}\mathbf{\Omega}\mathbf{V}^\top$ *be a symmetric normalized propagation operator and let*

$$\mathbf{Z}_r = \mathbf{P}^L \mathbf{X}\mathbf{W}_r. \tag{84}$$

Then

$$\mathbf{\Sigma}_r = \frac{1}{N}\mathbf{W}_r^\top \mathbf{X}^\top (\mathbf{P}^L)^\top \mathbf{J}_N \mathbf{P}^L \mathbf{X}\mathbf{W}_r. \tag{85}$$

Moreover, graph Fourier mode k *is multiplied by* ω_k^L *before the role-specific projection* $\mathbf{W}_r$*.*

Proof. The centered representation covariance is

$$\mathbf{\Sigma}_r = \frac{1}{N}(\mathbf{J}_N \mathbf{Z}_r)^\top (\mathbf{J}_N \mathbf{Z}_r) \tag{86}$$

$$= \frac{1}{N}\mathbf{Z}_r^\top \mathbf{J}_N \mathbf{Z}_r \tag{87}$$

$$= \frac{1}{N}\mathbf{W}_r^\top \mathbf{X}^\top (\mathbf{P}^L)^\top \mathbf{J}_N \mathbf{P}^L \mathbf{X}\mathbf{W}_r, \tag{88}$$

which proves Eq. 85. Since

$$\mathbf{P}^L = \mathbf{V}\mathbf{\Omega}^L\mathbf{V}^\top, \tag{89}$$

the coefficient of graph Fourier mode k is multiplied by ω_k^L. Therefore modes with $|\omega_k| < 1$ are attenuated in magnitude, with unequal $|\omega_k|$ producing anisotropic attenuation. □

D.4.3 DEGREE-DEPENDENT GAIN IN GIN

Proposition 4 (GIN pre-MLP covariance identity)**.** *For*

$$\mathbf{m}_{r,v} = (1 + \epsilon_{\text{GIN}})\mathbf{h}_{r,v} + \sum_{u \in \mathcal{N}(v)} \mathbf{h}_{r,u}, \tag{90}$$

define

$$\bar{\mathbf{h}}_{r,v} = \frac{(1 + \epsilon_{\text{GIN}})\mathbf{h}_{r,v} + \sum_{u \in \mathcal{N}(v)} \mathbf{h}_{r,u}}{d_v + 1}. \tag{91}$$

Then

$$\mathbf{m}_{r,v} = (d_v + 1)\bar{\mathbf{h}}_{r,v}. \tag{92}$$

Let $\bar{\mathbf{H}}_r$ *stack* $\bar{\mathbf{h}}_{r,v}$ *row-wise and define*

$$\mathbf{D} = \text{diag}(d_1 + 1, \ldots, d_N + 1). \tag{93}$$

Then

$$\mathbf{M}_r = \mathbf{D}\bar{\mathbf{H}}_r, \tag{94}$$

and its centered covariance is

$$\mathbf{\Sigma}_{\text{pre},r} = \frac{1}{N}\bar{\mathbf{H}}_r^\top \mathbf{D}\mathbf{J}_N \mathbf{D}\bar{\mathbf{H}}_r. \tag{95}$$

Proof. Stacking Eq. 92 over all nodes gives $\mathbf{M}_r = \mathbf{D}\bar{\mathbf{H}}_r$. Therefore,

$$\mathbf{\Sigma}_{\text{pre},r} = \frac{1}{N}(\mathbf{J}_N \mathbf{M}_r)^\top (\mathbf{J}_N \mathbf{M}_r) \tag{96}$$

$$= \frac{1}{N}\mathbf{M}_r^\top \mathbf{J}_N \mathbf{M}_r \tag{97}$$

$$= \frac{1}{N}\bar{\mathbf{H}}_r^\top \mathbf{D}\mathbf{J}_N \mathbf{D}\bar{\mathbf{H}}_r, \tag{98}$$

where $\mathbf{J}_N^\top \mathbf{J}_N = \mathbf{J}_N$ and $\mathbf{D}^\top = \mathbf{D}$. □

E AUXILIARY GEOMETRY CONTROLS

Section 4.3 gives the compact forms of the three auxiliary geometry controls. Here we provide their precise normalization, implementation, and geometric properties. All controls depend only on unlabeled source-representation statistics and are applied during source-domain SSL pretraining.

Let $\mathbf{H} \in \mathbb{R}^{m\times d}$ denote the node representations used to compute a geometry control, where m is the number of visible source nodes and d is the representation dimension. With $\mathbf{C}_m = \mathbf{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$, we define the centered covariance

$$\mathbf{\Sigma}_H = \frac{1}{m}\mathbf{H}^\top \mathbf{C}_m \mathbf{H}. \tag{99}$$

Each control below modifies a different aspect of this unlabeled second-order geometry.

E.1 RESOLVENT REGULARIZATION

Resolvent regularization (RR) acts directly on the covariance spectrum. To remove global scale, we first use the trace-normalized covariance

$$\mathbf{\Sigma}_H^{\mathrm{n}} = \frac{d\,\mathbf{\Sigma}_H}{\mathrm{tr}(\mathbf{\Sigma}_H) + \epsilon_{\mathrm{num}}}, \tag{100}$$

whose eigenvalues $\mu_1, \ldots, \mu_d$ satisfy $d^{-1}\sum_i \mu_i \approx 1$. Given a resolvent shift $\gamma > 0$, we define

$$r_1 = \frac{1+\gamma}{d}\sum_{i=1}^{d}\frac{1}{\mu_i + \gamma} - 1, \tag{101}$$

$$r_2 = \frac{(1+\gamma)^2}{d}\sum_{i=1}^{d}\frac{1}{(\mu_i + \gamma)^2} - 1, \tag{102}$$

and use

$$\mathcal{R}_{\mathrm{RR}} = r_1 + \alpha_{\mathrm{RR}} r_2, \qquad \alpha_{\mathrm{RR}} \geq 0. \tag{103}$$

The isotropic spectrum $\mu_i = 1$ for all i gives $r_1 = r_2 = 0$. Moreover,

$$\frac{\partial r_1}{\partial \mu_i} = -\frac{1+\gamma}{d(\mu_i+\gamma)^2}, \tag{104}$$

$$\frac{\partial r_2}{\partial \mu_i} = -\frac{2(1+\gamma)^2}{d(\mu_i+\gamma)^3}. \tag{105}$$

Hence RR places stronger pressure on small eigenvalues and preferentially acts on weak representation modes. Whether these modes carry target-task signal is determined separately by the Task–Geometry audit.

E.2 BARLOW REDUNDANCY OBJECTIVE

BRO controls redundancy between representation coordinates. Let $\widetilde{\mathbf{H}}$ denote the centered representation after channel-wise variance normalization, and define its correlation matrix

$$\mathbf{C}_H = \frac{1}{m}\widetilde{\mathbf{H}}^\top \widetilde{\mathbf{H}}. \tag{106}$$

BRO penalizes its off-diagonal entries:

$$\mathcal{R}_{\mathrm{BRO}} = \frac{1}{d^2}\sum_{i\neq j}(\mathbf{C}_H)_{ij}^2. \tag{107}$$

Ignoring the numerical stabilization used in channel normalization, $(\mathbf{C}_H)_{ii} = 1$, and therefore

$$\|\mathbf{C}_H - \mathbf{I}_d\|_F = d\sqrt{\mathcal{R}_{\mathrm{BRO}}}. \tag{108}$$

Let

$$\varepsilon_{\mathrm{B}} = d\sqrt{\mathcal{R}_{\mathrm{BRO}}}.$$

When $\varepsilon_{\mathrm{B}} < 1$, Weyl's inequality gives

$$1 - \varepsilon_{\mathrm{B}} \le \lambda_{\min}(\mathbf{C}_H) \le \lambda_{\max}(\mathbf{C}_H) \le 1 + \varepsilon_{\mathrm{B}}, \tag{109}$$

and consequently

$$\kappa(\mathbf{C}_H) \le \frac{1 + \varepsilon_{\mathrm{B}}}{1 - \varepsilon_{\mathrm{B}}}. \tag{110}$$

Thus, decreasing BRO improves the conditioning of the standardized representation geometry by suppressing cross-channel redundancy. It does not specify how target-task signal is distributed across the resulting modes.

E.3 Random-Direction Distortion

RDR serves as a task-agnostic perturbation of second-order geometry. Let $\mathbf{\Sigma}_H$ denote the covariance computed from all visible source nodes and let $\mathbf{\Sigma}_H^{\mathrm{sub}}$ denote the covariance computed from a random source-node subset. For K random unit directions $\mathbf{u}_k \in \mathbb{R}^d$, define

$$v_k = \mathbf{u}_k^\top \mathbf{\Sigma}_H \mathbf{u}_k, \qquad v_k^{\mathrm{sub}} = \mathbf{u}_k^\top \mathbf{\Sigma}_H^{\mathrm{sub}} \mathbf{u}_k. \tag{111}$$

The directional distortion is

$$\mathcal{R}_{\mathrm{RDR}} = \frac{1}{K} \sum_{k=1}^{K} \left| \log(v_k^{\mathrm{sub}} + \epsilon_{\mathrm{num}}) - \log(v_k + \epsilon_{\mathrm{num}}) \right|. \tag{112}$$

We additionally use the variance floor

$$\mathcal{R}_{\mathrm{floor}} = \frac{1}{K} \sum_{k=1}^{K} [\tau_{\mathrm{floor}} - v_k]_+^2, \tag{113}$$

which prevents the directional discrepancy from being reduced through global representation collapse. The complete RDR control is

$$\mathcal{R}_{\mathrm{RDR}}^{\mathrm{total}} = \mathcal{R}_{\mathrm{RDR}} + \alpha_{\mathrm{floor}} \mathcal{R}_{\mathrm{floor}}, \tag{114}$$

where $\tau > 0$ is the minimum directional variance and $\alpha_{\mathrm{floor}} \ge 0$ controls the floor term.

For a uniformly distributed random unit direction,

$$\mathbb{E}[\mathbf{u}_k \mathbf{u}_k^\top] = \frac{1}{d} \mathbf{I}_d,$$

and hence

$$\mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k^\top \mathbf{\Sigma}_H \mathbf{u}_k] = \frac{1}{d} \operatorname{tr}(\mathbf{\Sigma}_H). \tag{115}$$

RDR therefore probes unlabeled second-order energy across random directions. Its sampling and objective contain no target-task weights ψ_i, making it a suitable geometry-only control for testing whether geometry changes alone induce the predicted accessibility response.

E.4 Role in the Task–Geometry Analysis

The three controls act on complementary aspects of representation geometry: RR emphasizes weak covariance modes, BRO reduces coordinate redundancy, and RDR perturbs directional second-order structure. During pretraining, each is combined with the original SSL objective as

$$\mathcal{L} = \mathcal{L}_{\mathrm{SSL}} + \lambda_{\mathrm{geo}} \mathcal{R}_{\mathrm{geo}}, \tag{116}$$

where

$$\mathcal{R}_{\mathrm{geo}} \in \{\mathcal{R}_{\mathrm{RR}}, \mathcal{R}_{\mathrm{BRO}}, \mathcal{R}_{\mathrm{RDR}}^{\mathrm{total}}\}.$$

These controls modify unlabeled geometry but do not determine the target-task signal allocation $\{\psi_i\}$. Their effects are therefore evaluated through the induced Task–Geometry Profile and the empirical (Δ_C, Δ_A) decomposition in Sec. 4.3.

F Robustness to Nonlinear Probes

The Task–Geometry theory uses a population-eigenmode spectral ridge surrogate to characterize finite-supervision resolution. We use MLP probing as an empirical robustness test of whether the same representation-level organization remains informative under a nonlinear downstream readout. This experiment does not introduce a separate nonlinear theory: the TGP quantities are computed exactly as in the Linear-probe audit, while the empirical outcomes are recomputed with the MLP probe.

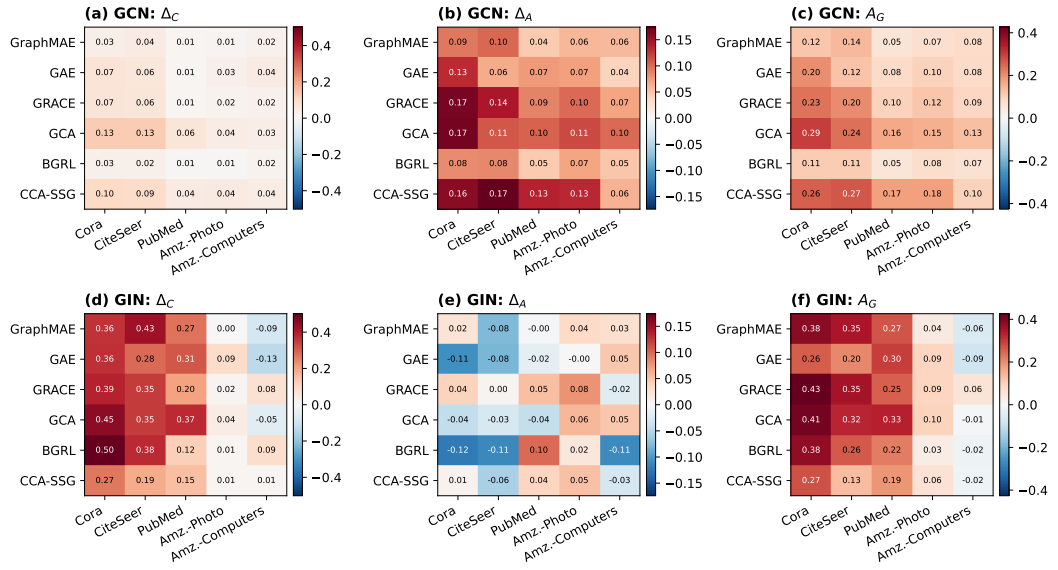

Figure 6: **Full MLP-probe utility-transfer landscape.** Each cell is one SSL–backbone–target setting and reports endpoint capability loss Δ_C, endpoint-relative finite-supervision loss Δ_A, or average trajectory gap A_G. The 60 settings complement the Linear-probe landscape in the main text and expose the complete nonlinear-readout evidence rather than only an aggregate statistic.

Figure 6 shows that GCN remains dominated by a positive finite-supervision component across SSL objectives, whereas GIN shows larger and more target-dependent endpoint gaps. The precise magnitudes change with the nonlinear probe, but the backbone-level contrast persists.

The capability correspondence remains positive for both GCN and GIN. The accessibility correspondence is clearer for GCN and heterogeneous for GIN, consistent with the main-text observation that GIN can jointly change signal retention and signal placement. Thus the MLP results support robustness of the representation-level organization without claiming that the ridge theorem directly characterizes nonlinear optimization.

G Detailed Large-Scale Results

Section 4.4 reports the large-scale OGBN-ARXIV results. Here we provide the complete transfer decomposition, Task–Geometry response, and absolute downstream performance.

RR primarily closes the GCN capability gap. Table 7 shows that RR consistently improves the source-pretrained GCN representation while leaving target performance nearly unchanged. Across the three SSL objectives, source Macro-F1 increases by 0.040 for Linear and 0.038 for MLP on average. Accordingly, Table 5 shows that mean Δ_C decreases from 0.042 to 0.002 for Linear and from 0.039 to 0.001 for MLP, with corresponding reductions in A_G.

The Task–Geometry response in Table 6 separates this capability gain from accessibility. RR reduces the mean signal gap from 0.059 to 0.038, whereas Δ_A^{TG} increases from -0.082 to 0.015. Empirical Δ_A likewise moves toward zero rather than becoming more favorable. Thus, RR's large-

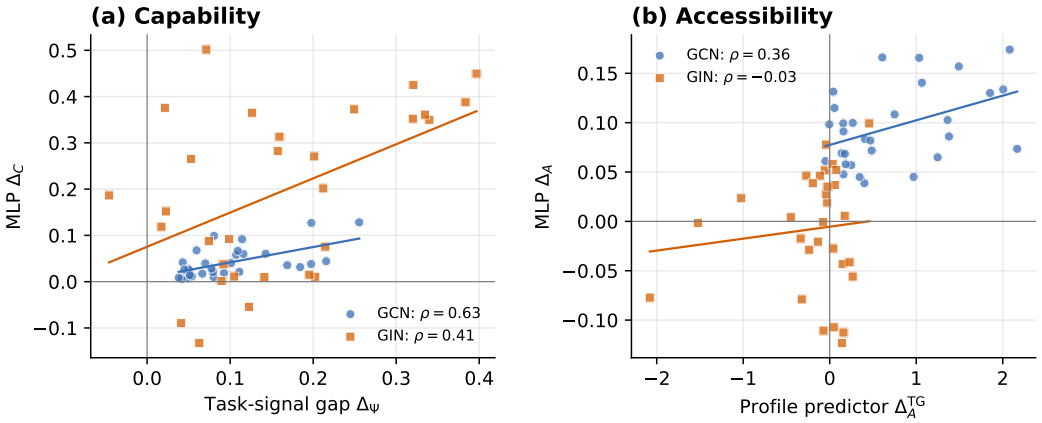

Figure 7: **Task–Geometry audit with an MLP probe.** The retained-signal gap is compared with empirical MLP endpoint loss, and the profile accessibility predictor is compared with empirical MLP finite-supervision loss. Points are all 60 SSL–backbone–target settings; correlations are reported separately by backbone.

Table 5: **Transfer decomposition on OGBN-ARXIV.** "Change" denotes Int.−Van. Negative A_G indicates that source-pretrained utility exceeds the target-matched reference on average over the evaluated label range.

			Δ_C			Δ_A			A_G		
SSL	Int.	Probe	Van.	Int.	Change	Van.	Int.	Change	Van.	Int.	Change
GraphMAE	RR	Linear	0.019	−0.003	−0.022	−0.023	−0.016	+0.007	−0.003	−0.018	−0.015
		MLP	0.012	−0.008	−0.020	−0.016	−0.009	+0.007	−0.004	−0.017	−0.013
	DT-GIN	Linear	−0.032	−0.021	+0.011	−0.004	−0.007	−0.003	−0.037	−0.028	+0.009
		MLP	−0.024	−0.032	−0.008	0.005	0.003	−0.002	−0.018	−0.029	−0.011
BGRL	RR	Linear	0.038	0.009	−0.029	−0.027	−0.015	+0.012	0.012	−0.006	−0.018
		MLP	0.036	0.008	−0.028	−0.020	−0.012	+0.008	0.016	−0.005	−0.021
	DT-GIN	Linear	−0.001	−0.010	−0.009	−0.010	−0.004	+0.006	−0.012	−0.014	−0.002
		MLP	−0.017	−0.024	−0.007	0.001	0.005	+0.004	−0.016	−0.019	−0.003
GRACE	RR	Linear	0.069	0.001	−0.068	−0.033	−0.001	+0.032	0.037	−0.000	−0.037
		MLP	0.068	0.002	−0.066	−0.022	−0.001	+0.021	0.047	0.001	−0.046
	DT-GIN	Linear	−0.011	0.092	+0.103	−0.001	−0.042	−0.041	−0.011	0.050	+0.061
		MLP	−0.022	0.107	+0.129	0.015	−0.035	−0.050	−0.008	0.073	+0.081

scale transfer benefit is dominated by capability recovery rather than improved finite-supervision accessibility.

Degree tempering strongly improves absolute utility. As shown in Table 7, DT-GIN produces substantially larger absolute gains than RR. Averaged across the three SSL objectives, source and target Macro-F1 increase by 0.287 and 0.322 for Linear, and by 0.253 and 0.291 for MLP, respectively.

Absolute improvement and transfer retention are distinct. Despite its large performance gain, DT-GIN does not uniformly close the source–target capability gap. Table 5 shows that vanilla GIN already has a small negative mean Δ_C on OGBN-ARXIV; after tempering, the mean becomes 0.020 for Linear and 0.017 for MLP. This effect is driven primarily by GRACE, for which target-matched pretraining benefits more strongly than source pretraining.

At the same time, accessibility becomes more favorable to the source representation: mean Δ_A changes from −0.005 to −0.018 for Linear and from 0.007 to −0.009 for MLP. Table 6 shows the same accessibility-side direction, with mean Δ_A^{TG} decreasing from −0.020 to −0.076. The

Table 6: **Task–Geometry response on OGBN-ARXIV.** RR is applied to GCN and DT-GIN to GIN. "Van." and "Int." denote vanilla and intervened models, and "Change" denotes Int.−Van.

		Δ_Ψ			Δ_A^{TG}		
SSL	Int.	Van.	Int.	Change	Van.	Int.	Change
GraphMAE	RR	0.044	0.028	−0.016	−0.103	−0.001	+0.102
	DT-GIN	0.041	0.018	−0.023	−0.008	−0.020	−0.012
BGRL	RR	0.050	0.036	−0.014	−0.097	−0.023	+0.074
	DT-GIN	0.017	0.063	+0.046	−0.008	−0.115	−0.107
GRACE	RR	0.082	0.050	−0.032	−0.045	0.070	+0.115
	DT-GIN	0.077	0.122	+0.045	−0.043	−0.092	−0.049

Table 7: **Absolute downstream performance on OGBN-ARXIV.** Entries report Macro-F1 before and after intervention. "Change" denotes Int.−Van.

			Source Macro-F1			Target Macro-F1		
SSL	Int.	Probe	Van.	Int.	Change	Van.	Int.	Change
GraphMAE	RR	Linear	0.409	0.431	+0.022	0.429	0.428	−0.000
		MLP	0.392	0.412	+0.020	0.404	0.404	−0.000
	DT-GIN	Linear	0.082	0.390	+0.309	0.049	0.370	+0.320
		MLP	0.073	0.356	+0.283	0.050	0.324	+0.274
BGRL	RR	Linear	0.392	0.421	+0.030	0.430	0.430	+0.000
		MLP	0.371	0.399	+0.028	0.407	0.407	−0.000
	DT-GIN	Linear	0.049	0.376	+0.327	0.047	0.366	+0.319
		MLP	0.052	0.348	+0.296	0.035	0.324	+0.289
GRACE	RR	Linear	0.350	0.418	+0.068	0.419	0.419	−0.000
		MLP	0.327	0.393	+0.066	0.395	0.395	+0.000
	DT-GIN	Linear	0.074	0.300	+0.226	0.064	0.392	+0.328
		MLP	0.082	0.263	+0.181	0.060	0.370	+0.310

large-scale regime therefore exhibits an architecture-level utility gain together with a capability–accessibility redistribution, rather than uniform transfer-gap closure.

Large-scale implication. The OGBN-ARXIV results reinforce the distinction between absolute representation utility and cross-domain transfer retention. RR mainly recovers source capability, whereas DT-GIN strongly improves both source and target representations but can alter their relative advantage in an SSL-dependent manner. Thus, an architectural improvement need not reduce every transfer component, further motivating the joint use of Δ_C, Δ_A, and A_G when evaluating graph foundation representations.

H Discussion

Implications for graph foundation model pretraining. Our results suggest that the objective of graph foundation pretraining should extend beyond producing representations with high downstream capacity. A transferable representation must both retain target-relevant utility across domains and make that utility accessible from limited target supervision. These requirements are distinct: representations with similar large-budget performance can exhibit different learning trajectories, while an architectural intervention can improve capability without improving accessibility. We therefore view the pair (Δ_C, Δ_A), together with the overall trajectory deficit A_G, as a more informative target for evaluating graph foundation representations than endpoint performance alone.

Graph inductive bias is part of the transfer objective. The GCN and GIN results indicate that transfer is shaped not only by the SSL objective but also by how the backbone organizes informa-

tion before it reaches the downstream task. Normalized propagation progressively induces mode-selective profile separation, whereas unnormalized accumulation introduces degree-dependent scaling that can alter retained signal and its modal allocation. This suggests that backbone design and self-supervised objectives should not be treated as independent components of a GFM. A broadly pretrained objective can still yield poor transfer if the aggregation operator systematically places useful information in modes that are weak, unstable, or strongly domain dependent.

From representation preservation to resolvability. A common goal in graph SSL is to preserve informative or diverse representations. The Task–Geometry view suggests a stronger criterion: useful information should also remain statistically resolvable under the target supervision regime in which the foundation model will be used. In particular, preserving total signal alone does not determine how quickly that signal can be recovered by a downstream learner. This motivates pretraining objectives that account for both signal retention and the distribution of representational strength across modes, especially when few-label adaptation is a primary deployment setting.

Limits of task-agnostic geometry shaping. The GCN interventions clarify that generic improvements in representation geometry do not necessarily translate into improved task-aware accessibility. RR and BRO modestly change the empirical transfer gaps while increasing the measured Δ_A^{TG} relative to vanilla GCN. Their downstream changes are therefore not mediated by an improvement in the measured task-aware accessibility profile. This behavior is consistent with Proposition 2: without knowing where future task signal will lie, a task-agnostic transformation of the spectrum cannot guarantee uniformly improved accessibility. For GFM design, this suggests that optimizing a single global statistic such as effective rank, conditioning, or isotropy is unlikely to be sufficient. A more promising direction is to preserve a broad set of usable modes or to exploit task-family and multi-domain structure during pretraining to obtain better proxies for future task-relevant directions.

Architecture-level control can improve transferable utility. Degree tempering illustrates a complementary design route. Suppressing GIN's degree-dependent gain substantially improves signal retention and capability, although finite-supervision accessibility does not improve in parallel. The resulting reduction in A_G shows that this trade-off still yields a net transfer benefit. Thus, DT-GIN is useful beyond its role as a mechanism probe: it shows that the Task–Geometry analysis can identify an architectural bias whose controlled modification improves cross-domain utility. More generally, this points toward GFM architectures in which structural amplification and attenuation are explicitly controlled instead of being fixed consequences of the aggregation rule.

Capability and accessibility should be optimized jointly. The DT-GIN results also show why maximizing either component alone can be misleading. An intervention may substantially recover capability while making the remaining signal harder to exploit with few labels. Conversely, improved finite-supervision behavior is of limited value if target-relevant capability has already been lost. For low-label transfer, the natural design problem is therefore multi-objective: *retain target-relevant capability* and *maintain finite-supervision accessibility*. The trajectory-level quantity A_G provides one summary of their net effect, while (Δ_C, Δ_A) preserves the distinction between the two failure modes.

Scope of the Task–Geometry surrogate. Our theoretical analysis uses a population-eigenmode spectral ridge surrogate to isolate finite-supervision resolution induced by a fixed representation geometry. The surrogate captures how mode strength and task-signal placement govern shrinkage and estimation variance, while abstracting away finite-sample rotation of the empirical covariance eigenspace and nonlinear downstream optimization. Accordingly, the TGP should be interpreted as a representation-level organizing principle rather than an exact predictor for every backbone and downstream learner. This distinction is visible within GIN, where the correspondence between Δ_A^{TG} and empirical Δ_A is weak despite clear landscape-level correspondence across the complete benchmark. At the same time, the MLP probe preserves the overall capability and accessibility trends, indicating that the identified organization is not restricted to linear downstream readout. Extending the theory to random empirical geometry and nonlinear prediction remains an important direction.