CONTRASTIVE NEURAL ALGORITHMIC REASONING FOR GRAPH COLORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph coloring seeks to assign colors to a graph's nodes so that adjacent nodes receive different colors, using as few colors as possible. Here, we study approximate k-coloring, where the goal is to use at most k colors while minimizing the number of monochromatic edges. This problem is central to graph theory and has applications in areas such as scheduling and resource allocation. Recent unsupervised GNN approaches optimize each instance directly, without amortizing optimization across test graphs. We instead propose a contrastive learning framework that encourages transferable coloring geometry where the embeddings of same-color nodes align up to sign, while adjacent nodes' representations are pushed toward orthogonal directions. With proper labels and sufficient dimension, every optimum of our contrastive loss variant over unit embeddings collapses each color class containing a non-isolated vertex onto a common line, with adjacent vertices lying on orthogonal lines. Under these assumptions, optimizing our InfoNCE variant over unit-norm embeddings approaches the geometry of a global hard-margin solution. On finite graph collections, we prove global convergence for projected subgradient descent in the Gram parameterization. Under suitable conditions, shared ReLU encoders with frozen features converge almost surely to the same global objective value through a smooth surrogate. A signed spectral Lovász theta bound strengthens the local edge-error certificate and complements direct conflict and properness certificates. Experiments on synthetic and real-world graphs evaluate transfer across graph sizes and datasets and yield low-conflict assignments, at the cost of allowing a small number of monochromatic edges.

1 INTRODUCTION

Graph coloring is a canonical problem in discrete optimization. Given a graph $G = (V, E)$, the goal is to assign colors to vertices such that adjacent vertices receive distinct colors, while minimizing the total number of colors used. This minimum is known as the chromatic number. Classical hardness results show that, for any $\epsilon > 0$, approximating the chromatic number within a factor of $O(n^{1-\epsilon})$ is NP-hard (Zuckerman, 2007). Despite this worst-case intractability, graph coloring has received sustained interest because of its broad range of applications, such as resource allocation and scheduling, where colors correspond to resources and edges to conflicts in resource usage. In practice, heuristics such as greedy algorithms (Brélaz, 1979) and SAT-based solvers (Selman et al., 1992; Van Gelder, 2008) are used to solve real-world instances. More recently, unsupervised machine learning approaches have been introduced (Schuetz et al., 2022; Vanderbush & Weber, 2026), typically learning color logits through a physics-inspired loss.

While these methods apply to arbitrary graphs, their combinatorial search rules or per-instance neural objectives do not directly characterize the representation geometry induced by coloring supervision. For the learned approach studied here, it is difficult *a priori* to characterize the graph classes on which it is likely to perform well, making it challenging for practitioners working with specific graph families to assess its effectiveness. Moreover, SAT-based solvers and the cited instance-optimized unsupervised learning methods optimize each graph instance independently, which can be computationally expensive. We propose a pipeline that aims to address these limitations along two dimensions: *interpretability*, by leveraging the geometry of node representations, and *scalability and domain-knowledge incorporation*, by adopting a supervised learning framework. These goals are closely connected. Since optimal coloring is computationally intractable on unrestricted inputs, one

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.

should not expect a single method to efficiently solve the problem in full generality. Instead, it is natural to exploit the simplifying geometric structure present in node embeddings for restricted graph families of interest.

Our approach leverages contrastive learning (van den Oord et al., 2018; Khosla et al., 2020) to provide a simple representation-level view of graph coloring. Given a proper coloring, vertices assigned the same color are treated as positive pairs, while adjacent vertices are treated as negative pairs. We propose a contrastive loss that is a variant of InfoNCE and analyze their geometry and optimization. These objectives align unoriented lines. Under this geometry, each color is represented by a *prototype*, corresponding to a one-dimensional subspace, and vertices of that color may be mapped to either direction along the same line. Decoding then consists of canonicalizing the embedding directions up to sign to points on the sphere, followed by clustering the resulting embeddings to determine vertex colors. Our proposed algorithm is schematically shown in Fig. 1.

1.1 Contribution of this paper

We propose a supervised neural graph-coloring approach with a representation-level certificate of properness.

More formally, we study graph coloring with at most $k_{\max}$ colors via a population objective over a distribution $\mathcal{D}_{\mathcal{G}}$ supported on graphs of bounded size. For each graph G, a coloring rule c_G using at most $k_{\max}$ colors supplies the positive pairs, and a node embedding function $f(G) \in (\mathcal{S}^{d-1})^{n_G}$ maps vertices to the unit sphere. Our main theoretical question is as follows:

> Given a proper (or optimal) coloring rule, what embedding geometry is enforced by a contrastive objective, and how does optimization select among the resulting prototype configurations?

Our answer has three parts:

1. **Geometry and loss calibration.** We prove a sharp lower bound for our contrastive loss, attainable when the embedding dimension $d \geq k_{\max}$, and characterize its equality cases (Thms. 1 and 2). Equality forces each color class containing a nonisolated vertex onto a common line and adjacent vertices onto orthogonal lines; with optimal labels on a graph containing an edge, all color lines are pairwise orthogonal (Cor. 1). When shared encoders are trained on the whole dataset, we decomposes the population risk gap into optimization and realizability gaps (Lem. 1).
2. **Convergence and margins.** We prove global convergence for projected Gram optimization with nonnegative same-color entries and almost surely for preconditioned gradient descent on shared ReLU (Glorot et al., 2011) readouts (Thm. 3). Both allow unequal class sizes and uncollapsed initialization. The shared result allows arbitrary depth with frozen full-rank features and a smooth surrogate, under explicit width, output-dimension, initialization, and step-size conditions. Extending the Gaussian tangent argument of Mei et al. (2017) to nonlinear Gram objectives gives output-dimension bounds on loss at approximate second-order points of the surrogate (Thm. 9). For unnormalized free embeddings, we solve the hard-margin problem and identify the exponential rate at which (1) approaches its infimum along positive-margin rays (Thm. 4).
3. **Coloring certificates.** We derive a signed spectral upper bound on Lovász $\vartheta(\overline{G})$ that sharpens the local edge-error bound (Lem. 2). Edge overlaps and affinities to orthogonal prototypes give a checkable properness certificate for arbitrary decoded assignments, without assuming k-colorability (Thm. 12). Under the stated affinity and coverage conditions, excess loss also bounds monochromatic edges and the additional colors needed for repair (Cor. 5).

We complement our theoretical analysis with experiments on synthetic combinatorial instances and real-world datasets, evaluating size generalization and transfer across graph families and citation datasets (Sec. 6). Well-tuned message-passing (Gilmer et al., 2017) and GPS-style (Rampášek et al., 2022) encoders produce low-conflict assignments, in some settings with smaller tested palettes than the greedy proper-coloring count.

Together, these results provide a theoretical account of what coloring-aware contrastive training can certify, through a precise geometric mechanism by which color partitions become line prototypes

and graph edges become orthogonality constraints. The experiments evaluate downstream coloring performance; they do not directly measure learned line collapse.

1.2 Related works

Graph coloring has long served as a central test case for the limits of efficient optimization. This history motivates relaxations and certificates, not exact recovery alone. Our work is closest in spirit to geometric relaxations of coloring, including vector colorings, semidefinite programming relaxations, and Lovász's theta function (Karger et al., 1998; Lovász, 1979). Our learning pipeline differs in the starting point: instead of solving an SDP, we ask when a learned contrastive representation produces the relevant orthogonality geometry.

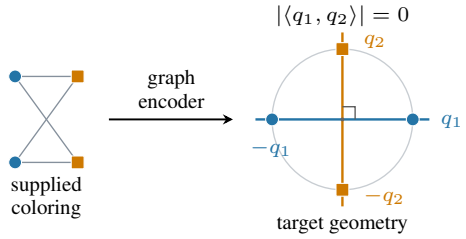


Figure 1: Supervised contrastive coloring framework and its target geometry. Vertices of one color may map to either end of the same prototype line, $\pm q_i$. Adjacent color classes target orthogonal lines, illustrated for two colors by $|\langle q_1, q_2\rangle| = 0$ on the unit circle. At testing, embedding directions are canonicalized up to sign and clustered.

Although our focus is on supervised learning of graph coloring, there exists a prolific line of SAT solvers (Selman et al., 1992) and neural algorithmic reasoning to tackle unsupervised coloring (Schuetz et al., 2022; Vanderbush & Weber, 2026). These per-instance approaches do not require a separate supervised training corpus, but their test-time cost includes optimization on each target graph.

The line-collapse phenomenon is related to neural collapse in supervised classification, where within-class features collapse and class means often approach a simplex ETF geometry (Papyan et al., 2020). Our geometry depends on the color quotient: its edges force orthogonality at lower-bound-attaining optima of (1). Convex Gram formulations underpin supervised contrastive landscape results (Behnia & Thrampoulidis, 2024) and characterizations of within-class collapse (Vock et al., 2026). Those analyses average individual log-softmax terms, whereas (2) pools positives before the logarithm, as in Khosla et al. (2020, Eq. (3)). Our absolute scores, graph-restricted negatives, and probability aggregation in (1) lead us to prove radial domination directly for the two objectives (Sec. D).

2 Setup and learning objectives

To connect the contrastive geometry with training, we first distinguish free embeddings from those produced by a shared encoder. Let the graph distribution $\mathcal{D}_\mathcal{G}$ have finite support on nonempty simple graphs $G = ([n_G], E_G)$ of bounded size. A fixed proper coloring c_G partitions each graph into nonempty classes $V_i(G)$, with at most $k_{\max}$ classes after relabeling. Properness means $c_G(u) \neq c_G(v)$ for every edge. Optimality, when stated, means that the number of classes is $\chi(G)$. The free representation is $h = f(G) \in (\mathcal{S}^{d-1})^{n_G}$, where $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ and $\tau > 0$ is a temperature. A shared encoder f_θ is a restricted family of such maps. Write $\mathcal{N}_G(v)$ for the neighbors of v. The color quotient $Q_c(G)$ has one vertex per color and an edge between colors that occur at opposite ends of an edge of G.

Contrastive objective. We measure similarity by the absolute inner product $s_{vw} = |\langle h_v, h_w\rangle|$ and define

$$\ell^{(G)}_{\mathrm{abs},\tau}(h, c_G) = -\frac{1}{n_G}\sum_v \frac{1}{|V_{c_G(v)}|}\sum_{w\in V_{c_G(v)}} \frac{e^{s_{vw}/\tau}}{e^{s_{vw}/\tau} + \sum_{u\in\mathcal{N}_G(v)} e^{s_{vu}/\tau}}. \tag{1}$$

It includes the anchor itself among the positives. Its population risk is $\ell_{\mathrm{abs},\tau}(f, c) = \mathbb{E}_G \ell^{(G)}_{\mathrm{abs},\tau}(f(G), c_G)$. Replacing s_{vw} in (1) with the signed inner product defines its signed counterpart $\ell^{(G)}_{\mathrm{InfoNCE},\tau}$.

Remark 1 (Relation to usual InfoNCE). *Usual InfoNCE averages negative log-softmax scores, whereas* (1) *averages negative softmax probabilities without a logarithm (van den Oord et al., 2018;*

Khosla et al., 2020). Here the scores are absolute inner products, negatives are graph neighbors, and positives include the anchor itself. Its line geometry will yield theta bounds and, with sufficient prototype affinity, proper-decoding certificates. For training stability, our contrastive experiments use (2), with signed or absolute similarities and nonself positives pooled before the logarithm.

Training objective. The absolute-loss implementation normalizes embeddings, removes self-positives, and omits anchors without another vertex of the same color. For a single graph, let P_v be the nonself positive set, p_v its size, and $\mathcal{A}$ the eligible anchors. Their positive and negative exponential sums, A_v and B_v, define the training objective

$$P_v = V_{c_G(v)} \setminus \{v\}, \quad p_v = |P_v|, \quad \mathcal{A} = \{v : p_v > 0\},$$
$$A_v = \sum_{w \in P_v} e^{s_{vw}/\tau}, \quad B_v = \sum_{u \in \mathcal{N}_G(v)} e^{s_{vu}/\tau}, \qquad J_\tau^{(G)}(h) = \frac{1}{|\mathcal{A}|} \sum_{v \in \mathcal{A}} \log\left(1 + \frac{B_v}{A_v}\right). \tag{2}$$

Set $J_\tau^{(G)} = 0$ when $\mathcal{A}$ is empty. Implementation details, including minibatch weighting and numerical stabilization, appear in Sec. I.1.

Each free objective is continuous on a finite product of unit spheres and therefore attains a minimum. To transfer this optimum to a shared encoder, we use random node features (Abboud et al., 2021) to interpolate arbitrary embeddings.

Proposition 1 (Exact realizability at arbitrary depth)**.** *Let the finite graph support contain $N = \sum_G n_G$ vertices, and freeze independent scalar features $\xi_{G,v}$ from an atomless distribution. Fix any hidden depth $L \geq 1$. A shared nodewise ReLU MLP with biases, widths $W_\ell \geq \max\{1, N-1\}$ for $\ell = 1, \ldots, L$, and d outputs realizes every prescribed family of unit embeddings almost surely after output normalization. Its finite-support population minima for (1), its signed counterpart, and (2) therefore equal the free minima and are attained. Admissible parameters have nonzero raw outputs at every support vertex.*

The construction in Sec. F chooses the parameters from the frozen features; it establishes expressivity, while the dynamics analysis below specifies updates that converge.

Training and decoding. We train a GNN on supplied proper colorings, including DSATUR (Brélaz, 1979) pseudo-labels, and decode its frozen embeddings by sign canonicalization and clustering. Palette selection uses the observed monochromatic-edge fraction (Sec. 6). We first characterize the representation geometry, identifying explicitly where optimal labels are needed.

3 Exact Representation Geometry

We now determine which embeddings attain the lowest value of the absolute InfoNCE variant (1). Write $D_v = |\mathcal{N}_G(v)|$ for the vertex degree, $n = n_G$ for the graph size, and $a_\tau = e^{1/\tau}$ for the largest positive exponential score. Proper labels suffice; the class sizes can be arbitrary.

Theorem 1 (Lower bound and attainment)**.** *For every unit embedding h and proper coloring c_G,*

$$\ell_{\mathrm{abs},\tau}^{(G)}(h, c_G) \geq b_G^{\mathrm{R}} := -\frac{1}{n} \sum_v \frac{a_\tau}{a_\tau + D_v}. \tag{3}$$

If $d \geq |\mathrm{Im}(c_G)|$, assigning orthonormal vectors to the supplied colors attains equality. Averaging (3) gives the corresponding population bound.

The bound is sharp because it simultaneously maximizes positive scores and minimizes edge scores. Its equality conditions identify the resulting geometry.

Theorem 2 (Equality geometry)**.** *Equality in (3) holds exactly when every edge joins orthogonal embeddings and every color class containing a vertex of positive degree lies on a single unoriented line. A class consisting entirely of isolated vertices is unrestricted.*

A nonisolated anchor attains its bound only at positive scores one and edge scores zero; an isolated anchor has constant loss -1 (Sec. B).

For optimal labels, every pair of classes has an edge between them, since otherwise the classes could be merged. The preceding characterization then forces all class lines to be orthogonal.

Corollary 1 (Optimal labels). *Suppose G has an edge, c_G uses $\chi(G)$ colors, and $d \geq \chi(G)$. Every unrestricted global minimizer has the form $h_v = \sigma_v q_{c_G(v)}$, where $\sigma_v \in \{-1, 1\}$ and the $\chi(G)$ unit prototypes are pairwise orthogonal.*

The baseline b_G^{R} is independent of the supplied proper coloring, so minimizing the absolute InfoNCE variant (1) fits the given partition without selecting the smallest number of colors. When G is edgeless, every embedding minimizes it. For trained representations, the remaining question is how an achieved loss gap controls departure from this geometry.

3.1 From trained loss to representation geometry

To apply the free optimum to a trained encoder, we separate optimization error from restrictions imposed by its architecture. For an encoder class $\mathcal{F}$, let $L(f) = \mathbb{E}_G \ell^{(G)}_{\mathrm{abs},\tau}(f(G), c_G)$ be its population risk and $B = \mathbb{E}_G b_G^{\mathrm{R}}$ the free baseline.

Lemma 1 (Realizability gap). *Assume $d \geq k_{\max}$, $\mathcal{F}$ is nonempty, and $f \in \mathcal{F}$. Its approximation gap $\mathcal{A}_{\mathcal{F}}$ satisfies*

$$L(f) - B = \underbrace{L(f) - \inf_{g \in \mathcal{F}} L(g)}_{\text{optimization gap}} + \underbrace{\inf_{g \in \mathcal{F}} L(g) - B}_{\mathcal{A}_{\mathcal{F}} \geq 0}. \tag{4}$$

The unrestricted graphwise product has $\mathcal{A}_{\mathcal{F}} = 0$. An encoder class also has zero gap if it simultaneously approximates one minimizing embedding on every graph in the finite support (in particular, under Prop. 1).

Geometry control via risk gap. The total achieved gap controls geometry regardless of its two sources. This is possible because a small excess loss forces same-color embeddings to align up to sign and adjacent embeddings to become nearly orthogonal, on average over the pairs covered by the objective (Thm. 5). For the contrastive loss variant (1), let $\epsilon_{\mathrm{R}} = \ell^{(G)}_{\mathrm{abs},\tau} - b_G^{\mathrm{R}}$ be the excess loss and abbreviate the anchor's color class as $V_v^c = V_{c_G(v)}$. Then, for a constant C_{R} specified in Thm. 5,

$$\frac{1}{n} \sum_v \left[\frac{D_v}{|V_v^c|} \sum_{w \in V_v^c} (1 - |\langle h_v, h_w \rangle|) + \sum_{u \in \mathcal{N}_G(v)} |\langle h_v, h_u \rangle| \right] \leq C_{\mathrm{R}} \epsilon_{\mathrm{R}}. \tag{5}$$

For the practical training loss variant (2), Thm. 5 gives an analogous bound over nonself positives and eligible anchors. Thus near-optimal loss bounds same-color dispersion and edge overlap for the embeddings produced by a GNN. Since $\min_{\sigma \in \{\pm 1\}} \|h_v - \sigma h_w\|^2 = 2(1 - |\langle h_v, h_w \rangle|)$, the first term measures squared distance between lines. If all classes have size at least two, every edge is counted twice. The constants can be large at low temperature. Proofs are in Sec. C. We next study updates that approach the unrestricted baseline.

4 Dynamics and implicit geometric bias

The loss-gap bounds raise an optimization question: which updates approach the geometric minimum? We answer this for free embeddings in the space of Gram matrices and a shared ReLU encoder, then study output dimension and large-norm behavior.

Convergence of objective values. To approach the geometric baseline, we use the class Gram matrix $X^{c_G}_{vw} = \mathbf{1}\{c_G(v) = c_G(w)\}$ as a common descent target. Let F_G denote the absolute InfoNCE variant (1) or its practical variant (2) in Gram coordinates, $\mathcal{L} = \mathbb{E}_G F_G$ its population risk, and $\mathcal{L}_* = \mathbb{E}_G F_G(X^{c_G})$ its free baseline.

Theorem 3 (Free and shared-encoder convergence; informal). *Assume the supplied colorings are proper.*

(i) ***Free representations.*** *Let $\mathcal{K}_c$ be the positive semidefinite, unit-diagonal Gram matrices with nonnegative same-color entries. From every $X_0 \in \mathcal{K}_c$, a projected flow and discrete scheme (details in Thm. 6), with its stated step-size conditions, converge to a global minimizer, with*

$$\lim_{t \to \infty} F_G(X_t) = F_G(X^{c_G}). \tag{6}$$

These Gram iterates can be represented by free embeddings when $d \geq n_G$.

(ii) ***Shared encoder.*** *At any depth, freeze some constructed features (details in Sec. F) and train only the selected readout, with $d \geq \max_G n_G$. For $\delta > 0$, let $\mathcal{L}_\delta$ be the smooth surrogate with linear positive scores and negative scores $\sqrt{x^2+\delta^2} - \delta$. Under a preconditioned gradient updates and absolutely continuous initialization (details in Thm. 11), the readouts B_t converge almost surely, and*

$$\lim_{t\to\infty} \mathcal{L}_\delta(B_t) = \lim_{t\to\infty} \mathcal{L}(B_t) = \mathcal{L}_*. \tag{7}$$

Both cases allow unequal class sizes and initially uncollapsed embeddings. The shared case reaches the original baseline by optimizing the smooth surrogate, while the geometry of (2) retains its coverage exceptions. The cited appendix results specify step sizes and prove convergence.

Landscape and output dimension. The exact objectives also admit descent directions under explicit factor-rank and coherent-sign conditions (Sec. E). For shared readouts with smaller output dimension, a nonlinear extension of Mei et al. (2017) gives quantitative approximation bounds and a finite negative-curvature optimizer (Sec. F).

Full-embedding margins. To study large-norm behavior, we compare positive and negative scores before normalization. For raw embeddings $U = (u_v)_{v\in[n]} \in \mathbb{R}^{n\times d}$, define the smallest positive–negative margin

$$\mathfrak{m}(U) = \min_{\substack{v\in[n],\, w\in V_{c_G(v)} \\ z\in\mathcal{N}_G(v)}} \big(|\langle u_v, u_w\rangle| - |\langle u_v, u_z\rangle|\big). \tag{8}$$

The positive set includes the anchor. Write $\ell_{\mathrm{R}}(U)$ for the absolute InfoNCE variant (1) without row normalization; the next theorem links its decay along a ray to this margin.

Theorem 4 (Full-embedding margin and loss decay)**.** *Let G have no isolated vertices, let c_G be any proper coloring with k classes, and assume $d \geq k$. Then*

$$\max_{\|U\|_F^2=1} \mathfrak{m}(U) = \frac{1}{n}. \tag{9}$$

Equality holds exactly when $u_v = \sigma_v q_{c_G(v)}/\sqrt{n}$, where $\sigma_v \in \{\pm 1\}$, every q_i is unit norm, and $q_i \perp q_j$ on quotient edges. For every U with $\mathfrak{m}(U) > 0$,

$$\lim_{r\to\infty} -\frac{\tau}{r^2} \log\big(1 + \ell_{\mathrm{R}}(rU)\big) = \mathfrak{m}(U). \tag{10}$$

Thus the optimal margin selects the same class lines as the normalized minimum. Sec. G proves the theorem, extends it to isolated vertices, and gives semidefinite certificates.

Max-margin consequence. Under the convergence assumptions of Thm. 3, the embeddings obtained for (1) satisfy $\mathfrak{m}(H_t/\sqrt{n_G}) \to 1/n_G$ on each training-support graph without isolated vertices. Thus these normalized dynamics approach the global hard margin characterized in Thm. 4. We next use the resulting geometry to certify relaxation values and decoded assignments.

5 Certificates from Approximate Orthogonality

Approximate orthogonality can certify a relaxation value even when training falls short of the geometric optimum. The Lovász sandwich inequality $\omega(G) \leq \vartheta(\overline{G}) \leq \chi(G)$ relates the clique number ω, theta function ϑ, and chromatic number χ (Lovász, 1979). We refine the classical orthogonal-representation argument to upper-bound the semidefinite relaxation $\vartheta(\overline{G})$ from arbitrary unit embeddings; certifying a chromatic upper bound will additionally require a proper assignment.

Lemma 2 (Spectral certificate)**.** *Let $G = ([n], E)$ be a finite simple graph with $n \geq 1$, $v_i \in \mathcal{S}^{d-1}$, and a unit handle $r \in \mathcal{S}^{d-1}$ with positive absolute overlaps $h_i := |\langle r, v_i\rangle|$. Let $H := \mathrm{diag}(h_i)$ be their diagonal matrix. Define the signed edge-overlap matrix by $S_{ij} := \langle v_i, v_j\rangle$ on edges and zero elsewhere, its entrywise magnitude by $W_{ij} := |S_{ij}|$, and its row sums by $D_i := \sum_j W_{ij}$. Then*

$$\vartheta(\overline{G}) \leq \underbrace{\lambda_{\max}\big(H^{-1}(I+S)H^{-1}\big)}_{\tau(S,h)} \leq \lambda_{\max}\big(H^{-1}(I+W)H^{-1}\big) \leq \max_i \frac{1+D_i}{h_i^2}. \tag{11}$$

If every edge overlap is at most ε and $\alpha := \min_i h_i$, then, writing A_G for the adjacency matrix and ρ for spectral radius,

$$\vartheta(\overline{G}) \le \frac{1+\rho(W)}{\alpha^2} \le \frac{1+\varepsilon\rho(A_G)}{\alpha^2} \le \frac{1+\varepsilon\Delta(G)}{\alpha^2}. \tag{12}$$

Proofs and optimized corrections appear in Sec. H.

Prototype affinity certifies decoding. To certify a decoded assignment $a : [n] \to [k]$, suppose its prototypes $q_1, \dots, q_k$ are orthonormal. Let $s = |\mathrm{Im}(a)|$ count its occupied colors, $\beta = \min_i |\langle v_i, q_{a(i)} \rangle|$ be its minimum prototype affinity, $\gamma(\beta) = 2\beta^2 - 1$ the induced overlap threshold, and $E_{\mathrm{mono}}(a)$ its monochromatic edges. If $\beta > 1/\sqrt{2}$, then

$$\gamma(\beta)|E_{\mathrm{mono}}(a)| \le \sum_{\{i,j\}\in E} |\langle v_i, v_j \rangle|. \tag{13}$$

If every edge overlap is strictly below $\gamma(\beta)$, the assignment is proper and $\chi(G) \le s$, without assuming colorability or class-size balance. Thm. 12 gives the complete statement, including a theta handle from prototype affinities and a sharper nonuniform conflict bound. The strict decoding threshold is sharp. Weighted handles accommodate heterogeneous affinities (Prop. 5). Certifying optimality requires an independent chromatic lower bound matching s.

Achieved loss gives a conflict bound. For the absolute InfoNCE variant (1), set $Q = nC_{\mathrm{R}}\epsilon_{\mathrm{R}}/2$. Combining Thm. 5 with the decoding theorem gives

$$\sum_{\{i,j\}\in E} |\langle v_i, v_j \rangle| \le Q, \qquad |E_{\mathrm{mono}}(a)| \le \lfloor Q/\gamma(\beta) \rfloor \quad (\beta > 1/\sqrt{2}). \tag{14}$$

For the practical training variant (2), the same statement holds with $Q = nC_{\log}(J_G - b_G^{\log})/2$ when every supplied class has at least two vertices. Thus $Q < \gamma(\beta)$ certifies a proper assignment, and otherwise recoloring a cover of the conflict edges gives at most $s + \lfloor Q/\gamma(\beta) \rfloor$ colors (Cor. 5).

6 Experiments

The above theoretical analysis are complete, but they impose too many restrictions to be practically tested. In our experiments, we opt to apply our loss functions more generally to large graph networks and even transformers.

To evaluate the coloring assignments and their transfer across graphs, we train a GNN encoder, normalize its node embeddings, and decode them by sweeping k in k-medoids (Kaufman & Rousseeuw, 1990) until the monochromatic-edge fraction meets a prescribed threshold. The v1/v1abs models use the signed/absolute versions of the practical training loss (2); v2/v2abs add a soft edge-conflict head (Sec. I.1). Citation labels come from DSATUR proper colorings, whose optimality is not assumed. We denote the selected palette by $k^\star$ and its monochromatic-edge fraction by Mono, counting a run as a 'Hit' when Mono≤ 0.05. When exact chromatic numbers are not uniformly available, we normalize by the greedy-coloring proxy χ_{greedy} and report $\rho = k^\star/\chi_{\mathrm{greedy}}$. The search returns the smallest successful tested palette or the recorded fallback, retaining failures and timeouts in the appendices. Because tied embeddings can leave clusters empty, $k^\star$ counts requested colors. Threshold hits can contain conflicts, so $\rho < 1$ does not establish a better proper coloring.

Ablations and transfer. We perform ablations across loss functions, feature constructions, and architectures spanning MPNNs and attention-based models on the Cora citation graph. We subsample 200 subgraphs for training and a separate set of 50 for testing. Although the center pools are disjoint, neighborhoods can share vertices; the records also lack a separate untouched model-selection split. In the controlled GCN (Kipf & Welling, 2017) comparison, absolute similarity reduces mean Mono from .0327 to .0251 while increasing mean k from 6.20 to 6.40. The full feature/loss, encoder, gate, residual, normalization, canonicalization, and SignNet (Lim et al., 2023) ablations are in Sec. I.

The GPS variants use GCN or GraphSAGE (Hamilton et al., 2017) local layers. The COLOR benchmark studies in-family generalization from bounded-size training graphs, with results summarized in Tab. 1. The evaluation contains 2 Book, 2 Myciel, and 7 Queen graphs; independent repeat summaries vary, so these small evaluation sets do not establish a best configuration per family (Sec. L).

Table 1: Selected COLOR configurations (ten-seed mean ± standard deviation). Selection uses OOD scores. Source split definitions and unresolved discrepancies between repeated aggregates are documented in Sec. L.

Family	Feature mode	Pipeline	Encoder	ρ_{ID}	$\mathrm{MONO}_{\mathrm{ID}}$	$\mathrm{Hit}_{\mathrm{ID}}$	ρ_{OOD}	$\mathrm{MONO}_{\mathrm{OOD}}$	$\mathrm{Hit}_{\mathrm{OOD}}$
Book	Random	v1	gps_sage	1.41±0.18	0.0451±0.0035	1.00±0.00	1.08±0.08	0.0452±0.0034	1.00±0.00
Myciel	Struct	v1	gps_gcn	1.45±0.28	0.0395±0.0041	1.00±0.00	1.54±0.40	0.0438±0.0045	1.00±0.00
Queen	Random	v2	gps_sage	0.90±0.03	0.0474±0.0018	1.00±0.00	0.76±0.02	0.0471±0.0008	1.00±0.00

With these GPS variants and a GatedGCN-style encoder (Bresson & Laurent, 2017), we train on Cora using 16 shared structural features and evaluate without retraining on CiteSeer and PubMed. Tabs. 2 and 22 report transfer to target subgraphs and full graphs, respectively.

Table 2: Cora-trained structural-feature transfer on held-out target subgraphs. Mean ± standard deviation over ten seeds, except CiteSeer v1/GatedGCN (nine).

Train → Test	Encoder	Pipeline	$k^\star$ avg.	MONO avg.	ρ avg.	Hit-rate
Cora → CiteSeer	gated_gcn	v1	6.06±0.41	0.0488±0.0049	1.36±0.08	0.74±0.11
Cora → CiteSeer	gated_gcn	v2	6.32±0.39	0.0478±0.0035	1.42±0.07	0.74±0.11
Cora → CiteSeer	gps_gcn	v1	6.72±0.64	0.0493±0.0033	1.49±0.10	0.74±0.10
Cora → PubMed	gated_gcn	v2	5.09±0.48	0.0310±0.0026	1.20±0.08	0.98±0.02
Cora → PubMed	gps_sage	v2	6.03±0.60	0.0331±0.0026	1.42±0.12	0.99±0.02
Cora → PubMed	gated_gcn	v1	5.06±0.44	0.0307±0.0033	1.20±0.06	0.99±0.02

We next consider node-feature adaptation schemes that use a shared cross-dataset input space. Hybrid concatenates normalized hashed bag-of-words (BOW) features (Weinberger et al., 2009) with structural features; LayerNorm (LN) (Ba et al., 2016) normalizes the concatenated vector. PCA/SVD (Jolliffe, 2002) reuse a Cora-fitted projection of hashed BOW on the target graphs. Tab. 3 reports the results on subgraph-split data preparation, run with our v2 pipeline. The v1 and full-graph transfer results, together with same-graph fitting, random-feature absolute transfer, full-graph tuning, and unitary-inspired (Kiani et al., 2024) GPS experiments, are in Sec. J. Full-graph fitting is transductive.

Table 3: Cora subgraph transfer with the v2 pipeline. Entries are mean ± standard deviation over ten random seeds.

Adaptation	Encoder	Cora → CiteSeer		Cora → PubMed	
		ρ	MONO	ρ	MONO
Hybrid	gated_gcn	1.26±0.13	0.0335±0.0036	1.27±0.12	0.0310±0.0027
	gps_gcn	1.29±0.10	0.0335±0.0043	1.38±0.12	0.0313±0.0032
	gps_sage	1.25±0.09	0.0328±0.0032	1.27±0.14	0.0305±0.0030
Hybrid+LN	gated_gcn	1.29±0.10	0.0336±0.0044	1.28±0.11	0.0317±0.0031
	gps_gcn	1.38±0.11	0.0334±0.0025	1.36±0.16	0.0310±0.0023
	gps_sage	1.28±0.08	0.0346±0.0042	1.32±0.11	0.0310±0.0023

Adaptation	Encoder	Cora → CiteSeer		Cora → PubMed	
		ρ	MONO	ρ	MONO
PCA	gated_gcn	1.53±0.21	0.0375±0.0027	2.27±0.43	0.0351±0.0020
	gps_gcn	1.48±0.18	0.0361±0.0026	1.60±0.18	0.0344±0.0033
	gps_sage	1.42±0.27	0.0369±0.0046	1.52±0.18	0.0342±0.0020
SVD	gated_gcn	1.47±0.11	0.0357±0.0038	1.43±0.11	0.0356±0.0017
	gps_gcn	1.60±0.24	0.0377±0.0024	1.43±0.12	0.0333±0.0025
	gps_sage	1.41±0.16	0.0359±0.0037	1.31±0.08	0.0323±0.0027

Size transfer and runtime. To test transfer across graph sizes, we train only on cycles C_n with $50 \leq n \leq 200$ using random node initialization. The test suite contains 40 small graphs and 20 large cycles, from C_{7000} to C_{7019}. On the small 40-graph subset, PI-GNN (Schuetz et al., 2022) and full-GCN (Vanderbush & Weber, 2026) are more color-efficient than the learned cycle-trained models. On the large-cycle stress test, both baselines time out under the 15-minute per-graph cap. GatedGCN v1abs requires $5.88 \pm .36$ seconds for training and $.0253 \pm .0075$ seconds per large cycle at test time (Sec. M); these timings exclude labeling and preprocessing. Tab. 4 reports representative instances, including threshold failures and baseline timeouts. Tabs. 44 and 45 give the full results.

Table 4: Representative instances after cycle training, reported as $k^\star$ / MONO. Learned entries are ten-seed mean $\pm$ standard deviation. Full-GCN is warm-start ColoringGNN (Vanderbush & Weber, 2026); both external baselines have a 900-second cap.

Graph	n	χ	v1abs + GatedGCN	v2abs + GPS-SAGE	PI-GNN	full-GCN
K_10	10	10	8.1±0.3 / 0.0422±0.0070	8.3±0.5 / 0.0378±0.0107	8 / 0.0444	8 / 0.0444
K_{4,8}	12	2	4.6±2.8 / 0.0063±0.0132	2.8±0.6 / 0.0000±0.0000	2 / 0.0000	2 / 0.0000
C_30	30	2	3.3±0.5 / 0.0067±0.0141	3.8±0.6 / 0.0133±0.0172	3 / 0.0000	3 / 0.0000
C_31	31	3	3.3±0.5 / 0.0032±0.0102	4.0±1.1 / 0.0032±0.0102	3 / 0.0000	3 / 0.0000
C_7000	7000	2	3.5±0.7 / 0.0133±0.0100	3.9±0.3 / 0.0228±0.0145	timeout	timeout
C_7019	7019	3	3.4±0.5 / 0.0137±0.0106	4.1±0.6 / 0.0161±0.0124	timeout	timeout
W_14	14	4	4.8±0.4 / 0.0231±0.0199	5.6±2.1 / 0.0308±0.0162	3 / 0.0385	3 / 0.0385
Petersen	10	3	4.6±1.3 / 0.0000±0.0000	5.1±1.3 / 0.0000±0.0000	3 / 0.0000	3 / 0.0000
Icosahedral	12	4	7.1±1.4 / 0.0267±0.0141	6.0±0.7 / 0.0300±0.0105	4 / 0.0333	5 / 0.0000
KG(9,3)	84	5	9.5±1.0 / 0.0931±0.0211	9.4±0.7 / 0.0677±0.0100	4 / 0.0119	4 / 0.0119
Mycielski(C5)^3	47	6	8.7±1.3 / 0.0653±0.0160	9.0±1.4 / 0.0458±0.0099	4 / 0.0169	4 / 0.0254

Established solvers and empirical limits. We additionally evaluate HEA (Galinier & Hao, 1999), ANYCSP (Tönshoff et al., 2023), and an independent GIN-based baseline by Zhang et al. (2024) (Zhang-GIN) in 1452 attempts over 121 instances, two seeds, and fixed-k/legal-search protocols (Sec. N). Zhang-GIN did not publish their code to the best of our knowledge, so our test uses a reimplementation. HEA returns proper colorings in every attempt and reaches the prescribed target in $240/242$ attempts under each protocol, making it the strongest proper-coloring comparator in this evaluation. All methods ran on the same cluster, with neural models using GPUs and the evaluated HEA implementation running on CPUs because it has no GPU support. Our method has lower recorded per-instance inference latency after shared training (Tabs. 42 and 46). Zhang-GIN records no improvement over its DSATUR initialization in successful legal-search attempts. ANYCSP completes $32/484$ attempts, with the remaining attempts timing out. Our ANYCSP evaluation uses an approximately 800-second training snapshot, charged against each 900-second attempt budget, whereas the original paper reports 24–48 hours of training. These results depend on the evaluated implementations and budgets. Our method amortizes training across graphs, using an encoder forward pass followed by clustering over candidate palettes at inference.

7 DISCUSSION

We introduced an absolute-value contrastive framework for supervised graph coloring. Under proper labels, lower-bound optima of the contrastive loss variant (1) collapse each color class containing a nonisolated vertex onto an unoriented line, with adjacent classes on orthogonal lines.

Our analysis connects excess loss with line geometry and computable decoding certificates. We prove global convergence for Gram projection and for a shared ReLU readout with full-rank features, a smooth surrogate, and specified updates. Extending these guarantees to end-to-end GNN training and coloring supervision with conflicts remains open.

Empirically, contrastive GNN encoders provide fast reusable colorings and show promising size and OOD generalization. Matched solver timing that includes label generation, geometry measurements throughout training, and decoder comparisons across thresholds would further test the proposed mechanism.

AI USE STATEMENT

In this work, we used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, providing feedback on research methodology or experiments, implementing methods, and interpreting results. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, or cleaning and reformatting datasets, and translation and qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for literature review, scientific figures, manuscript editing, and reference formatting. We have reviewed all AI-assisted work. Our checks included comparison with primary references, numerical checks of selected identities and bounds, comparison of reported tables with experiment records, and compilation and visual inspection of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs of our theoretical claims appear in Appendices B–H. Sec. I.1 specifies the implemented objectives and their relation to the analyzed losses. Secs. I, J, L and M describe graph construction, feature processing, data splits, architectures, training and decoding settings, and aggregation across seeds. Sec. N gives the solver budgets and timing conventions, and Sec. O records software versions, hardware, and dataset sources. The reporting conventions in Sec. I identify unavailable predictions and checkpoints and discrepancies between recorded aggregates, which limit independent reconstruction of the reported neural results.

REFERENCES

Ralph Abboud, İsmail İlkan Ceylan, Martin Grohe, and Thomas Lukasiewicz. The surprising power of graph neural networks with random node initialization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

P.-A. Absil, Robert Mahony, and Ben Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005. doi: 10.1137/040605266.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL https://arxiv.org/abs/1607.06450.

Tina Behnia and Christos Thrampoulidis. Supervised contrastive representation learning: Landscape analysis with unconstrained features. In *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024. URL https://arxiv.org/abs/2402.18884.

Nicolas Boumal, Vladislav Voroninski, and Afonso S. Bandeira. The non-convex Burer–Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/3de2334a314a7a72721f1f74a6cb4cee-Abstract.html.

Daniel Brélaz. New methods to color the vertices of a graph. *Commun. ACM*, 22(4):251–256, April 1979. ISSN 0001-0782. doi: 10.1145/359094.359101. URL https://doi.org/10.1145/359094.359101.

Xavier Bresson and Thomas Laurent. Residual gated graph convnets. *arXiv preprint arXiv:1711.07553*, 2017.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998. doi: 10.1016/S0169-7552(98)00110-X. URL https://research.google/pubs/the-anatomy-of-a-large-scale-hypertextual-web-search-engine/.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.07289.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428*, 2019. URL https://arxiv.org/abs/1903.02428.

Philippe Galinier and Jin-Kao Hao. Hybrid evolutionary algorithms for graph coloring. *Journal of Combinatorial Optimization*, 3:379–397, 1999. doi: 10.1023/A:1009823419804.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 315–323, 2011. URL https://proceedings.mlr.press/v15/glorot11a.html.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Alain Hertz and Dominique de Werra. Using tabu search techniques for graph coloring. *Computing*, 39(4):345–351, 1987. doi: 10.1007/BF02239976.

David S. Johnson and Michael A. Trick (eds.). *Cliques, Coloring, and Satisfiability: Second DIMACS Implementation Challenge*, volume 26 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, Providence, RI, 1996.

I. T. Jolliffe. *Principal Component Analysis*. Springer, 2 edition, 2002. doi: 10.1007/b98835.

David R. Karger, Rajeev Motwani, and Madhu Sudan. Approximate graph coloring by semidefinite programming. *Journal of the ACM*, 45(2):246–265, 1998.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990. doi: 10.1002/9780470316801.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Bobak Kiani, Lukas Fesser, and Melanie Weber. Unitary convolutions for learning on graphs and groups. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=lG1VEQJvUH.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176:311–337, 2019. doi: 10.1007/s10107-019-01374-3. URL https://arxiv.org/abs/1710.07406.

Jiajin Li, Anthony Man-Cho So, and Wing-Kin Ma. Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions. *IEEE Signal Processing Magazine*, 37(5):18–31, 2020. doi: 10.1109/MSP.2020.3003845.

Derek Lim, Joshua Robinson, Lingxiao Zhao, Tess Smidt, Suvrit Sra, Haggai Maron, and Stefanie Jegelka. Sign and basis invariant networks for spectral graph representation learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Q-UHqMorzil.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

László Lovász. On the Shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25 (1):1–7, 1979.

Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto Imbuzeiro Oliveira. Solving SDPs for synchronization and MaxCut problems via the Grothendieck inequality. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1476–1515, 2017. URL https://proceedings.mlr.press/v65/mei17a.html.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 1998. doi: 10.1007/978-3-642-02431-3.

Martin J. A. Schuetz, J. Kyle Brubaker, Zhihuai Zhu, and Helmut G. Katzgraber. Graph coloring with physics-inspired graph neural networks. *Phys. Rev. Res.*, 4:043131, Nov 2022. doi: 10.1103/PhysRevResearch.4.043131. URL https://link.aps.org/doi/10.1103/PhysRevResearch.4.043131.

Bart Selman, Hector J. Levesque, and David G. Mitchell. A new method for solving hard satisfiability problems. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92)*, pp. 440–446, San Jose, California, 1992. AAAI Press.

Hamed Shirzad, Ameya Velingker, Balaji Venkatachalam, Danica J. Sutherland, and Ali Kemal Sinop. Exphormer: Sparse transformers for graphs. In *International Conference on Machine Learning*, 2023.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL https://jmlr.org/papers/v15/srivastava14a.html.

Jan Tönshoff, Berke Kisin, Jakob Lindner, and Martin Grohe. One model, any CSP: Graph neural networks as fast global search heuristics for constraint satisfaction. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4280–4288, 2023. doi: 10.24963/ijcai.2023/476. URL https://www.ijcai.org/proceedings/2023/476.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Allen Van Gelder. Another look at graph coloring via propositional satisfiability. *Discrete Applied Mathematics*, 156(2):230–243, 2008. doi: 10.1016/j.dam.2006.07.016.

Knut Vanderbush and Melanie Weber. Neural algorithmic reasoning for approximate k-coloring with recursive warm starts. *arXiv preprint arXiv:2601.05137*, 2026. URL https://arxiv.org/abs/2601.05137.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Raphaël Vock, Edouard Duchesnay, and Benoit Dufumier. A unified geometric framework for weighted contrastive learning. *arXiv preprint arXiv:2605.13943*, 2026. URL https://arxiv.org/abs/2605.13943.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1113–1120, 2009. doi: 10.1145/1553374.1553516. URL https://doi.org/10.1145/1553374.1553516.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 40–48. PMLR, 2016. URL https://proceedings.mlr.press/v48/yanga16.html.

Yicheng Zhang, Kai Zhang, and Ni Wu. Using graph isomorphism network to solve graph coloring problem. In *2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pp. 209–215, 2024. doi: 10.1109/ICCECE61317.2024.10504151.

David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, 3(6):103–128, 2007. doi: 10.4086/toc.2007.v003a006. URL https://www.theoryofcomputing.org/articles/v003a006/.

Appendix Table of Contents

A Nonsmooth stationarity used in the proofs 15

B Proofs for the absolute-value InfoNCE variant 15

C Realizability and loss calibration 16

D Global dynamics for the exact objectives in Gram coordinates 18

E Landscape of normalized free embeddings 22

F Shared ReLU encoders and a nonlinear extension of Mei et al. 24

G Full-embedding margins and semidefinite certificates 29

H Proofs and refinements of the geometric certificates 33

I Cora ablations 39

J Experiments on citation graphs 47

K Extra citation network experiments 56

L COLOR experiments 58

M Cycle experiments 61

N Comparisons with per-instance coloring solvers 63

O Details on Experimental Setups 65

A NONSMOOTH STATIONARITY USED IN THE PROOFS

Absolute inner products introduce nonsmooth points when overlaps vanish. To define projected Gram dynamics at those points, we use the Clarke directional derivative (Rockafellar & Wets, 1998; Li et al., 2020) $f^\circ(x; v)$ of a locally Lipschitz function $f : \mathbb{R}^m \to \mathbb{R}$ at x in direction v, and its Clarke subdifferential $\partial_C f(x)$:

$$f^\circ(x; v) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y + tv) - f(y)}{t}, \tag{15}$$

$$\partial_C f(x) = \{z : \langle z, v\rangle \le f^\circ(x; v) \text{ for every } v\}. \tag{16}$$

In finite dimensions, this subdifferential is the closed convex hull of limits of gradients at nearby differentiability points and is nonempty, compact, and convex. It reduces to $\{\nabla f(x)\}$ when f is continuously differentiable. Every local minimum is Clarke stationary, meaning $0 \in \partial_C f(x)$, although a stationary point need not minimize the function (Rockafellar & Wets, 1998).

Our embeddings lie on spheres, so we also need stationarity for the restriction to a smooth embedded manifold. We express its Clarke subdifferential in local coordinates and transport it to the tangent space using the Riemannian metric; for a smooth function on a unit sphere, this gives the gradient $(I - qq^\top)\nabla f(q)$. For a finite-maximum objective $F = \max_{s \in S} F_s$, where S indexes smooth branches F_s, the subdifferential has an explicit representation. Its directional derivative is the maximum of the derivatives of the active branches, and its Clarke subdifferential is their convex hull. Restriction to a smooth local chart preserves this finite-maximum formula and projects each branch gradient onto the tangent space. This argument supplies the equality that a general ambient Clarke chain rule, by itself, would not guarantee.

Applied to $F(q) = |\langle q_i, q_j\rangle|$, the branch formula multiplies the gradient of the inner product by its sign when the overlap is nonzero and by a scalar in $[-1, 1]$ when it vanishes. The two gradient blocks for an unordered pair share that scalar. Accordingly, we require one joint selection to satisfy every block of a stationary system; separate membership in the partial subdifferentials would not suffice. The resulting equations identify first-order candidates, whose optimality must be established by a further argument.

B PROOFS FOR THE ABSOLUTE-VALUE INFONCE VARIANT

The geometric characterization follows by examining when each anchor-positive term reaches its lower bound. We work throughout with unit embeddings, first establishing attainment over free representations and then identifying the equality conditions. Independence of the graphwise variables applies to this unrestricted class; the interpolation argument in Sec. F supplies the corresponding shared-encoder construction.

Free-representation attainment used in Prop. 1. We obtain a free minimizer from compactness. For a finite support $\{G_1, \dots, G_m\}$, the unrestricted domain is the compact product $\prod_{r=1}^{m} (\mathcal{S}^{d-1})^{n_{G_r}}$. Strictly positive denominators make the signed and absolute InfoNCE variants continuous on this product, so their finite weighted sums attain minima. The aggregate log objective is continuous for each fixed nonempty active-anchor set and is constant zero when that set is empty; its population minimum is therefore attained as well. Interpolating these free minimizers, as proved in Sec. F, establishes the shared-encoder part of the proposition. □

Proof of Thm. 1. We bound each loss term by maximizing its positive score and minimizing its negative scores; orthonormal color vectors attain both choices simultaneously. For an anchor v and positive $w \in V_{c_G(v)}$, let $A = e^{|\langle h_v, h_w\rangle|/\tau}$ be the positive exponential score and $B = \sum_{u \in \mathcal{N}_G(v)} e^{|\langle h_v, h_u\rangle|/\tau}$ the total negative score. Unit norms give $A \le a_\tau$ and $B \ge D_v$. Since $-A/(A+B)$ is nonincreasing in A and nondecreasing in B, it is at least $-a_\tau/(a_\tau + D_v)$; averaging over positives and anchors gives (3). If $d \ge |\mathrm{Im}(c_G)|$, we can assign an orthonormal vector to each used color, making $A = a_\tau$ and $B = D_v$ for every pair. This construction attains the bound, whose population version follows by averaging over graphs. □

Proof of Thm. 2. Because the loss is an average of nonnegative gaps above the pointwise bound, we can characterize equality one anchor at a time. For $D_v > 0$, the preceding monotonicities are strict: its gap vanishes if and only if $A = a_\tau$ and every summand in B equals one. In geometric terms, $|\langle h_v, h_w \rangle| = 1$ for every positive w and $\langle h_v, h_u \rangle = 0$ for every neighbor u. The positive equality gives $h_w = \pm h_v$, so one nonisolated vertex forces its entire color class, including any isolated members, onto a single line. The negative equality gives orthogonality on every edge, since both endpoints are nonisolated.

Isolated anchors account for the remaining freedom. When $D_v = 0$, the pointwise loss is identically -1 and imposes no restriction on positives. A class consisting entirely of isolated vertices never occurs among the negatives either, leaving its embeddings arbitrary. Substitution shows that these conditions suffice term by term. With positive mass on every graph in a finite support, the population gap vanishes precisely when they hold on every support graph. □

Proof of Cor. 1. Optimality of the labels turns orthogonality on quotient edges into orthogonality between every pair of color lines. Indeed, if two used classes had no edge between them, merging them would produce a proper coloring with fewer colors; the quotient of a chromatic-optimal coloring must therefore be complete. Since G has an edge, $\chi(G) \geq 2$, and each class contains a nonisolated vertex. The dimension hypothesis gives attainment, while the equality characterization supplies a line for every class and orthogonality between all such lines. Choosing a unit representative of each line yields the asserted signs and prototypes. The edge hypothesis excludes the edgeless case, where the objective is constant and all embeddings minimize it, even for an optimal one-color labeling. □

C Realizability and loss calibration

The free optimum provides a reference for learned representations, but using it requires two further comparisons. We first separate the encoder's realizability gap from its optimization gap, then bound geometric defects by the achieved loss.

Proof of Lem. 1. We obtain the decomposition by inserting the best loss available to the encoder class, and use simultaneous approximation of free minimizers to identify when the realizability gap vanishes. The lower bound gives $L(f) \geq B$ for every $f \in \mathcal{F}$, hence $\mathcal{A}_\mathcal{F} \geq 0$; adding and subtracting $\inf_{g \in \mathcal{F}} L(g)$ yields (4). In the unrestricted class, the feasible domain is a product over support graphs, so the orthonormal construction attains each b_G^{R} simultaneously and has population loss B. More generally, suppose a sequence of encoders $f_m \in \mathcal{F}$ converges to one such minimizing family on every support graph. Continuity of the losses and finiteness of the support give $L(f_m) \to B$, establishing equality of infima without requiring the encoder class to attain its own infimum.

Input symmetry shows why the approximation hypothesis matters. Let P be the permutation exchanging the two vertices of K_2. With identical node features, this automorphism fixes the input, and deterministic permutation equivariance gives $Pf(K_2) = f(K_2)$. The resulting embeddings coincide and have absolute inner product one, although lower-bound equality requires zero on this edge. Such an encoder class therefore has a strict realizability gap on K_2, regardless of its width. □

An explicit approximation estimate. To quantify the preceding continuity argument, suppose f_0 is a minimizing family and $f \in \mathcal{F}$ approximates it with uniform error $\max_{G,v} \|[f(G)]_v - [f_0(G)]_v\| \leq \eta$, with both families unit norm. The resulting realizability gap satisfies

$$0 \leq \mathcal{A}_\mathcal{F} \leq L(f) - B \leq \eta/\tau. \tag{17}$$

To verify this estimate, we pass the embedding error through the scores and then through the scalar loss. Each absolute inner product changes by at most 2η; at a nonisolated anchor, both the log-sum-exp of negative scores and the positive score divided by temperature therefore change by at most $2\eta/\tau$. Their difference enters the loss through a signed sigmoid with derivative magnitude at most $1/4$, giving a loss change of at most η/τ. Isolated terms contribute no change because their loss is constant.

For the aggregate log objective, the same score bounds control the change in $\log B_v - \log A_v$ by $4\eta/\tau$. Composing with the 1-Lipschitz function $z \mapsto \log(1 + e^z)$ bounds the log-loss approximation

gap by $4\eta/\tau$, with zero-negative anchors again contributing constant terms. These estimates turn an assumed representation approximation into a loss bound; obtaining that approximation from a particular architecture remains a separate requirement.

The converse comparison starts from the achieved loss and controls departures from the target geometry. For an anchor v and a nonempty positive set P, define its geometric defect $T_v(P)$ as the sum of degree-weighted positive dispersion and total edge overlap:

$$T_v(P) := \frac{D_v}{|P|} \sum_{w \in P} (1 - |\langle h_v, h_w \rangle|) + \sum_{u \in \mathcal{N}_G(v)} |\langle h_v, h_u \rangle|. \tag{18}$$

Theorem 5 (Excess loss controls geometric defects)**.** *Let h be a unit embedding and c_G a proper coloring. Write $\Delta_G = \max_v D_v$ for the maximum degree and $\kappa_\tau = 1 - a_\tau^{-1}$ for the temperature factor. For the InfoNCE variant, define the excess loss $\epsilon_{\mathrm{R}} = \ell^{(G)}_{\mathrm{abs},\tau} - b^{\mathrm{R}}_G$. Then*

$$\frac{1}{n} \sum_v T_v(V_{c_G(v)}) \le C_{\mathrm{R}} \epsilon_{\mathrm{R}}, \qquad C_{\mathrm{R}} = \tau a_\tau (1 + \Delta_G a_\tau)(1 + \Delta_G / a_\tau). \tag{19}$$

Write $J_G = J^{(G)}_\tau(h)$ for the aggregate log objective in the setup. If its active-anchor set satisfies $\mathcal{A} \neq \varnothing$, define the log baseline $b^{\log}_G$ and defect constant $C_{\log}$ by

$$b^{\log}_G = \frac{1}{|\mathcal{A}|} \sum_{v \in \mathcal{A}} \log\left(1 + \frac{D_v}{p_v a_\tau}\right), \qquad C_{\log} = \frac{a_\tau}{\kappa_\tau} \max_{v \in \mathcal{A}} (p_v + D_v). \tag{20}$$

Then $J_G \ge b^{\log}_G$ and

$$\frac{1}{|\mathcal{A}|} \sum_{v \in \mathcal{A}} T_v(P_v) \le C_{\log}(J_G - b^{\log}_G). \tag{21}$$

Both lower bounds are attained when $d \ge |\mathrm{Im}(c_G)|$.

Proof of Thm. 5: InfoNCE variant. We lower-bound each term's excess by its positive deficit and negative overlaps, then average to recover the geometric defect. Fix an anchor v and positive w. Let x denote their positive deficit, let y_u be the overlap with neighbor u, and abbreviate the anchor degree and temperature exponential by D and a:

$$x = 1 - |\langle h_v, h_w \rangle|, \quad y_u = |\langle h_v, h_u \rangle|, \quad D = D_v, \quad a = a_\tau.$$

Factoring out the positive exponential score writes the pointwise loss as $-1/(1+S)$, where S is the ratio of total negative score to positive score and S_0 its baseline value:

$$S = a^{-1} \sum_{u \in \mathcal{N}_G(v)} e^{(x + y_u)/\tau}, \qquad S_0 = D/a.$$

The inequality $e^t \ge 1 + t$ for $t \ge 0$ turns the score-ratio gap into the defect bound $S - S_0 \ge (Dx + \sum_u y_u)/(\tau a)$. To control the denominator in the loss difference, we also use $S \le Da$, since the original exponent $(|\langle h_v, h_u \rangle| - |\langle h_v, h_w \rangle|)/\tau$ is at most $1/\tau$. Together these estimates give

$$-\frac{1}{1+S} + \frac{1}{1+S_0} = \frac{S - S_0}{(1+S)(1+S_0)} \ge \frac{Dx + \sum_u y_u}{\tau a (1 + Da)(1 + D/a)}. \tag{22}$$

Replacing D by Δ_G in the positive denominator, averaging over $w \in V_{c_G(v)}$, and then over v proves (19). When $D = 0$, both sides of the pointwise bound are zero. □

Proof of Thm. 5: aggregate log objective. We separate the anchor's excess into a negative-overlap term and a positive-dispersion term, bounding each through the exponential scores. Fix an active anchor v, with positive count $p = p_v \ge 1$, degree $D = D_v$, temperature exponential $a = a_\tau$, and factor $\kappa = 1 - a^{-1}$. Denote the positive deficits by x_w, their mean by $\bar{x}$, and the negative overlaps by y_u:

$$x_w = 1 - |\langle h_v, h_w \rangle|, \quad \bar{x} = p^{-1} \sum_{w \in P_v} x_w, \quad y_u = |\langle h_v, h_u \rangle|.$$

Let $A = \sum_{w \in P_v} e^{(1-x_w)/\tau}$ and $B = \sum_{u \in \mathcal{N}_G(v)} e^{y_u/\tau}$ be the total positive and negative scores. The bounds $A \le pa$ and $B \ge D$ place the anchor loss $\log(1+B/A)$ above its baseline $\log(1+D/(pa))$. To quantify this excess, we separate the change in B from the change in A:

$$\log(1+B/A) - \log(1+D/(pa)) = \log(A+B) - \log(A+D) \tag{23}$$

$$+ \log(1+D/A) - \log(1+D/(pa)). \tag{24}$$

The first difference controls negative overlaps. Integrating $1/t$ over $[A+D, A+B]$ bounds it below by $(B-D)/(A+B)$; using $B-D \ge \tau^{-1}\sum_u y_u$ and $A+B \le a(p+D)$ then gives the contribution $\sum_u y_u/[\tau a(p+D)]$.

For the positive-dispersion term, convexity of $x \mapsto e^{-x/\tau}$ gives the secant bound $e^{-x/\tau} \le 1-\kappa x$ on $[0,1]$, hence $A \le pa(1-\kappa\bar{x})$. Write $r = D/(pa)$ for the baseline negative-to-positive ratio and $z = \kappa\bar{x} < 1$ for the scaled mean deficit. The scalar loss $\psi(z) = \log(1 + r/(1-z))$ satisfies $\psi'(z) = r/[(1-z)(1-z+r)] \ge r/(1+r)$ on $[0,1)$. Integrating this lower bound makes the second difference at least $D\kappa\bar{x}/(pa+D) \ge D\kappa\bar{x}/[a(p+D)]$. Combining both contributions and using $\kappa \le 1/\tau$ yields

$$\log(1+B/A) - \log(1+D/(pa)) \ge \frac{\kappa}{a(p+D)} \left(D\bar{x} + \sum_u y_u \right). \tag{25}$$

The formula holds also for $D = 0$, when both sides vanish. Averaging over active anchors proves (21). Orthonormal color embeddings simultaneously have $A = pa$ and $B = D$, proving attainment under the dimension hypothesis. □

Coverage and numerical implementation. The active-anchor set determines which geometric constraints the log objective can enforce. An empty set gives the constant-zero objective and no geometric information. For a nonempty set, strict monotonicity in A and B at every anchor with $D_v > 0$ forces absolute alignment with all nonself positives and orthogonality to every neighbor at equality; substitution establishes sufficiency. Edges between singleton classes have no active endpoint and remain unconstrained. If all classes have at least two vertices, every anchor is active and the equality geometry agrees with Thm. 2.

To apply the ideal bound to the numerical loss, we account for the 10^{-12} stabilizer added to its numerator and denominator. For a stabilizer $\delta > 0$, the anchor value is $\log((A+B+\delta)/(A+\delta))$; its absolute difference from $\log((A+B)/A)$ is at most $\log(1+\delta/A) \le \delta/A \le \delta$, because unit-norm embeddings give $A \ge p \ge 1$. If $\widetilde{J}_G$ denotes the reported stabilized loss, replacing $J_G - b_G^{\log}$ by $\widetilde{J}_G + \delta - b_G^{\log}$ therefore recovers the ideal excess-risk bound under exact evaluation of the stabilized formula. Floating-point overflow and roundoff require separate error control. The code also excludes same-label neighbors from the negative set; with imperfect labels, the theorem applies to the graph of cross-label edges and leaves the omitted monochromatic edges uncertified.

These estimates constrain the outputs of any encoder. To determine when an optimizer approaches the geometric baseline, we next work in Gram coordinates.

D Global dynamics for the exact objectives in Gram coordinates

The minima characterized earlier identify the desired geometry; we now ask whether a specified optimizer reaches them. We work with Gram matrices in the Euclidean space $\mathbb{S}^n$ of real symmetric matrices, equipped with the Frobenius inner product. Fix a proper coloring c_G of a nonempty simple graph. Vertex signs $\sigma_v \in \{-1, 1\}$ replace h_v by $\sigma_v h_v$ without changing either absolute-similarity objective. We analyze configurations for which these signs make all same-color inner products nonnegative. Their feasible Gram matrices form the chamber

$$\mathcal{K}_c = \{X \succeq 0 : \; X_{vv} = 1, \; X_{vw} \ge 0 \text{ if } c_G(v) = c_G(w)\}. \tag{26}$$

This chamber includes the uncollapsed identity initialization $X_0 = I$. Every feasible matrix factors in at most n embedding coordinates; the Gram optimizer imposes no rank constraint. Let $F(X)$ be either (1) or (2), evaluated from X, and let X^c be the coloring Gram matrix with entries $X^c_{vw} = \mathbf{1}\{c_G(v) = c_G(w)\}$. Write $F_* = F(X^c)$ for its objective value and $\Delta_G = \max_v D_v$ for the maximum degree. The key estimate controls the loss excess by the derivative toward X^c.

Theorem 6 (Global Gram dynamics). *For every $X \in \mathcal{K}_c$ and every subgradient $g \in \partial_C F(X)$ of the extension that uses linear within-color scores,*

$$\langle g, X - X^c \rangle_{\mathrm{F}} \geq \mu_F (F(X) - F_*), \qquad \mu_{\mathrm{R}} = \frac{1 + \Delta_G / a_\tau}{1 + \Delta_G a_\tau}, \quad \mu_{\log} = \frac{1}{\tau(a_\tau - 1)}. \tag{27}$$

Projected subgradient flow on $\mathcal{K}_c$ from any feasible initialization converges to a global minimizer and, for $t > 0$, satisfies

$$F(X(t)) - F_* \leq \frac{\|X(0) - X^c\|_{\mathrm{F}}^2}{2\mu_F t}. \tag{28}$$

Projected subgradient descent has a best-iterate $O(T^{-1/2})$ guarantee with explicit constants, and converges with positive steps satisfying $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$.

The extension at zero positive scores. To define the projected dynamics at the chamber boundary, we use the linear score X_{vw} for every positive pair and retain $|X_{vu}|$ on negative edges. The resulting ambient extension agrees with the original objective throughout $\mathcal{K}_c$, where positive overlaps are nonnegative. At a zero positive overlap it selects the derivative toward the chamber. Negative zero-overlap subgradients remain arbitrary valid joint selections. This choice fixes the boundary derivative used in the theorem.

The positive sets determine the baseline in this chamber. Equation (1) includes the anchor among its same-color positives, whereas (2) uses the nonself positives P_v and the fixed active-anchor set $\mathcal{A}$. If $\mathcal{A}$ is empty, the log objective is constant zero and its excess-loss bounds are vacuous. Otherwise its minimum is $b_G^{\log}$; the minimum of (1) is b_G^{R}. The coloring Gram matrix X^c attains both baselines because it is positive semidefinite with unit diagonal, positive scores one, and edge scores zero. We next show that motion toward this matrix controls the full loss excess.

Lemma 3 (Radial domination for the two exact objectives). *For F just defined, (27) holds for every $X \in \mathcal{K}_c$ and every $g \in \partial_C F(X)$. It also holds with X^c replaced by any global minimizer $Y \in \mathcal{K}_c$.*

Proof. We compare each anchor term with its baseline along the segment to X^c, then average. Along $X(t) = (1-t)X + tX^c$, a positive score $a \in [0,1]$ becomes $a + t(1-a)$, and a negative absolute score $b \in [0,1]$ becomes $(1-t)b$. Put $a_\tau = e^{1/\tau}$. At a zero negative score, every allowed sign selection gives the same directional derivative along this segment, because the corresponding coordinate of $X - X^c$ is zero. The extension makes all positive scores smooth. The calculation therefore evaluates $\langle g, X - X^c \rangle_{\mathrm{F}}$ for every joint subgradient.

For an anchor-positive pair in (1) with degree $D > 0$, let $x = 1 - a$ be its positive deficit and let b_u be its negative scores. Define the negative-to-positive exponential sum and its baseline by

$$S = a_\tau^{-1} \sum_u e^{(x + b_u)/\tau}, \qquad S_0 = D / a_\tau.$$

Its contribution to the InfoNCE variant is $-1/(1+S)$, with baseline $-1/(1+S_0)$. The negative directional derivative toward X^c is

$$H_{vw} = \frac{a_\tau^{-1}}{(1+S)^2} \sum_u \frac{x + b_u}{\tau} e^{(x+b_u)/\tau}.$$

For $z \geq 0$, $ze^z \geq e^z - 1$, so

$$H_{vw} \geq \frac{S - S_0}{(1+S)^2} \tag{29}$$

$$= \frac{1 + S_0}{1 + S} \left[-\frac{1}{1+S} + \frac{1}{1+S_0} \right] \tag{30}$$

$$\geq \frac{1 + D/a_\tau}{1 + Da_\tau} \left[-\frac{1}{1+S} + \frac{1}{1+S_0} \right]. \tag{31}$$

Here $S \leq Da_\tau$ because $b_u - a \leq 1$. The coefficient $(1 + D/a_\tau)/(1 + Da_\tau)$ decreases in $D \geq 0$, and is therefore at least μ_{R}. For $D = 0$, the anchor term and its baseline both equal -1, and both directional quantities vanish. Averaging over positives and anchors proves the inequality for the InfoNCE variant.

For an anchor in (2) with $p \geq 1$ positives and $D > 0$ negatives, let $x_w = 1 - a_w \in [0, 1]$ be its positive deficits. We use the positive and negative exponential sums A, B, and their log odds z with baseline z_*:

$$A = a_\tau \sum_w e^{-x_w/\tau}, \quad B = \sum_u e^{b_u/\tau}, \quad z = \log(B/A), \quad z_* = \log(D/(pa_\tau)).$$

Let $s(z) = \log(1 + e^z)$ and $s'(z) = e^z/(1 + e^z)$. Convexity gives

$$s(z) - s(z_*) \leq s'(z)(z - z_*). \tag{32}$$

The negative directional derivative is

$$H_v = \frac{s'(z)}{\tau} \left(\frac{\sum_u b_u e^{b_u/\tau}}{\sum_u e^{b_u/\tau}} + \frac{\sum_w x_w e^{-x_w/\tau}}{\sum_w e^{-x_w/\tau}} \right). \tag{33}$$

We bound the two tilted expectations separately. The convex function $t \mapsto \log(D^{-1} \sum_u e^{tb_u/\tau})$ vanishes at zero; its derivative at one is at least its value at one. Hence the first expectation, divided by τ, is at least $\log(B/D)$. For the positive part, put $T = 1/\tau$, $y_w = x_w/\tau \in [0, T]$, and $m = p^{-1} \sum_w e^{-y_w} = A/(pa_\tau)$. The function $y/(e^y - 1)$ is decreasing for $y > 0$ and has limit one at zero. Thus, with

$$c = \frac{T}{e^T - 1} = \frac{1}{\tau(a_\tau - 1)} \in (0, 1),$$

we have $ye^{-y} \geq c(1 - e^{-y})$ for every $y \in [0, T]$. Averaging and dividing by m gives

$$\frac{\sum_w x_w e^{-x_w/\tau}}{\tau \sum_w e^{-x_w/\tau}} \geq c \frac{1 - m}{m} \geq c(-\log m) = c \log(pa_\tau/A), \tag{34}$$

where the second inequality follows from $\log z \leq z - 1$ with $z = 1/m \geq 1$. Since $c \leq 1$ and both logarithmic differences are nonnegative, the two estimates and (32) give

$$H_v \geq c\, s'(z) \big[\log(B/D) + \log(pa_\tau/A)\big] \geq \mu_{\log} \big[s(z) - s(z_*)\big].$$

An active anchor with $D = 0$ contributes the constant zero. Averaging over active anchors proves the log inequality.

To extend the estimate to every minimizer, recall the equality conditions: each positive pair at a nonisolated anchor has score one, and each covered negative pair has score zero. These are the coordinates used in the nonconstant terms above. Replacing X^c by another minimizer Y leaves their derivatives unchanged, while every remaining term is constant. The same radial bound thus holds toward Y. □

Subgradient bounds and regularity. The flow needs existence and a chain rule; the discrete update also needs a uniform bound on subgradients. Both follow from a finite smooth-branch representation: choose one sign for each unordered negative edge, substitute its signed linear score, and maximize over all sign choices. Monotonicity in each negative score ensures equality with the original extension. The smooth branches have Hessians bounded below on a convex neighborhood of the compact set $\mathcal{K}_c$. Thus F is weakly convex there: for some finite λ, $F(X) + (\lambda/2)\|X\|_{\mathrm{F}}^2$ is convex on that neighborhood. The constrained extended-valued function $F + \iota_{\mathcal{K}_c}$ has the same weak-convexity property, where ι is zero on its set and infinity elsewhere.

Explicit uniform subgradient bounds are

$$\|g\|_{\mathrm{F}} \leq G_F, \qquad G_{\mathrm{R}} = \frac{1}{2\tau}, \quad G_{\log} = \frac{2}{\tau}. \tag{35}$$

Indeed, for one summand of the InfoNCE variant the absolute positive-score derivative and the sum of the absolute negative-score derivatives both equal $AB/[\tau(A + B)^2] \leq 1/(4\tau)$. For one log anchor those two quantities both equal $B/[\tau(A+B)] \leq 1/\tau$. The Frobenius norm of the coordinate gradient of a symmetric entry X_{ij} is at most one, so the triangle inequality, averaging, and convexity of the Clarke subdifferential give (35). These constants do not require a minimum graph degree or a minimum class size.

Proof of the flow statement in Thm. 3. We construct the constrained flow from implicit steps, then combine its energy identity with radial domination. The outward normal cone records the constraint:

$$N_{\mathcal{K}_c}(X) = \{n : \langle n, Y - X\rangle_{\mathrm{F}} \le 0 \text{ for every } Y \in \mathcal{K}_c\}.$$

The projected subgradient flow is the differential inclusion

$$\dot{X}(t) + g(t) + n(t) = 0, \qquad g(t) \in \partial_C F(X(t)), \quad n(t) \in N_{\mathcal{K}_c}(X(t)), \quad X(t) \in \mathcal{K}_c, \tag{36}$$

holding almost everywhere for an absolutely continuous trajectory. For this weakly convex objective with a convex constraint, we obtain a solution from every feasible initial condition by minimizing

$$F(X) + \frac{\|X - X_k\|_{\mathrm{F}}^2}{2h} \quad \text{over } X \in \mathcal{K}_c$$

at each implicit time step $h\lambda < 1$. Adding the quadratic makes the subproblem strictly convex, so its minimizer is unique. Comparing with X_k gives $\sum_k \|X_{k+1} - X_k\|_{\mathrm{F}}^2/h \le 2(F(X_0) - F_*)$. Compactness and this bound give convergent absolutely continuous interpolants on bounded time intervals. The subgradient inequalities for the convex function $F + \iota_{\mathcal{K}_c} + (\lambda/2)\|\cdot\|^2$ pass to the limit and give (36). The same inequalities give uniqueness through $\frac{d}{dt}\|X - Z\|^2 \le 2\lambda\|X - Z\|^2$ and Gronwall's inequality.

To control the objective along this flow, use the chain rule supplied by its finite-maximum representation: $\frac{d}{dt}F(X(t)) = \langle g(t), \dot{X}(t)\rangle_{\mathrm{F}}$ almost everywhere. To see why any active-branch subgradient gives the same value, note that two differentiable branch values that agree along an absolutely continuous path have equal derivatives at almost every time in their coincidence set. At a point with only one active branch the claim is the ordinary chain rule; taking convex combinations handles multiple active branches. Similarly, feasibility at both earlier and later times implies $\langle n(t), \dot{X}(t)\rangle_{\mathrm{F}} = 0$ almost everywhere. Substituting the differential inclusion gives

$$\frac{d}{dt}F(X(t)) = -\|\dot{X}(t)\|_{\mathrm{F}}^2 \le 0. \tag{37}$$

For any global minimizer Y, radial domination and the normal-cone inequality give

$$\frac{d}{dt}\frac{1}{2}\|X(t) - Y\|_{\mathrm{F}}^2 = -\langle g(t) + n(t), X(t) - Y\rangle_{\mathrm{F}} \tag{38}$$

$$\le -\mu_F(F(X(t)) - F_*). \tag{39}$$

Take $Y = X^c$ and integrate. Since the loss is nonincreasing, the integrated excess is at least $t(F(X(t)) - F_*)$, proving (28). Compactness gives a subsequential limit, and the rate and continuity make it a global minimizer Y_∞. Equation (39) says distance to this particular minimizer is nonincreasing, while along the convergent subsequence it tends to zero. Therefore the whole trajectory converges to Y_∞. □

Discrete guarantees in Thm. 3. For the discrete updates, projection converts radial domination into a distance recurrence. We sum it to obtain the rate and use its summable error term to prove convergence. Let $\Pi_{\mathcal{K}_c}$ denote exact Euclidean projection in the Frobenius norm and set

$$X_{t+1} = \Pi_{\mathcal{K}_c}(X_t - \eta_t g_t), \qquad g_t \in \partial_C F(X_t), \quad \eta_t > 0. \tag{40}$$

Projection is nonexpansive and fixes every feasible Y. For any global minimizer Y, Lem. 3 therefore gives

$$\|X_{t+1} - Y\|_{\mathrm{F}}^2 \le \|X_t - Y\|_{\mathrm{F}}^2 - 2\mu_F\eta_t(F(X_t) - F_*) + \eta_t^2 G_F^2. \tag{41}$$

Summing with $Y = X^c$ proves the explicit bound

$$\min_{0 \le t < T}\big(F(X_t) - F_*\big) \le \frac{\|X_0 - X^c\|_{\mathrm{F}}^2 + G_F^2\sum_{t<T}\eta_t^2}{2\mu_F\sum_{t<T}\eta_t}. \tag{42}$$

For a known $R > 0$ with $\|X_0 - X^c\|_{\mathrm{F}} \le R$, the constant step $\eta_t = R/(G_F\sqrt{T})$ gives the best-iterate bound $RG_F/(\mu_F\sqrt{T})$. One may always take $R = n$: every off-diagonal entry of $X_0 - X^c$ has absolute value at most one and every diagonal entry vanishes. This guarantee permits transient increases of the discrete loss.

If $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$, the recursion shows $\sum_t \eta_t(F(X_t) - F_*) < \infty$, hence $\liminf_t F(X_t) = F_*$. It also implies that $\|X_t - Y\|^2$ converges for each fixed global minimizer Y: adding the tail $G_F^2 \sum_{j \geq t} \eta_j^2$ makes this sequence nonincreasing and bounded below. Compactness supplies a subsequence converging to a global minimizer Y_∞, and the convergence of the distance to that minimizer forces the entire sequence to converge to it. For example, $\eta_t = \eta_0(t+1)^{-a}$ with $1/2 < a \leq 1$ has the required sums. □

From Gram dynamics to embedding coordinates. Radial domination gives global convergence despite nonconvexity of the objectives, for every initial matrix in the chosen chamber, including $X_0 = I$. Exact projection onto $\mathcal{K}_c$ is a convex semidefinite problem and may increase rank; these updates therefore specify an optimizer for the full Gram representation. The limiting geometry retains the isolated-class and positive-coverage exceptions of each objective. To understand optimization in embedding coordinates, we next use the same radial estimate to construct descent curves. That argument will retain the coherent-sign condition, which arbitrary frustrated within-color configurations need not satisfy; neither result identifies these updates with AdamW (Loshchilov & Hutter, 2019) on a shared GNN.

E LANDSCAPE OF NORMALIZED FREE EMBEDDINGS

The Gram dynamics reach a minimum by updates in matrix space. To understand the landscape in embedding coordinates, we instead turn their radial direction into feasible curves of unit vectors. Let $H \in \mathbb{R}^{n \times d}$ be the matrix of unit embedding rows, and let $F(HH^\top)$ denote either normalized objective. Write F_* for the minimum of the full Gram problem, attained by free embeddings when $d \geq k$, and let $\mu_F > 0$ be the radial constant in (27). We work with configurations whose within-color inner products become nonnegative after vertex sign changes, which leave both objectives invariant. An unused embedding direction will provide a curve that improves the loss by moving one color class at a time. This use of unused directions is related to Behnia & Thrampoulidis (2024); our classwise curves quantify the decrease through the graph-specific radial bound.

Theorem 7 (Descent without an equitable partition)**.** *Suppose signs $\sigma_v \in \{-1, 1\}$ exist such that $\sigma_v \sigma_w \langle h_v, h_w \rangle \geq 0$ whenever $c(v) = c(w)$. No class balance, neighbor-count regularity, or initial collapse is required.*

(i) *If $\operatorname{rank}(H) < d$ and $F(HH^\top) > F_*$, a smooth unit-row curve $H(t)$ through H has $\|H'(0)\|_F = 1$, zero first loss derivative, and*

$$\frac{d^2}{dt^2} F(H(t)H(t)^\top)\Big|_{t=0} \leq -\frac{2\mu_F}{n}\big(F(HH^\top) - F_*\big) < 0. \tag{43}$$

(ii) *If $\operatorname{rank}(H) = n$, a nonglobal point admits a feasible first-order descent curve.*

Thus, if $d \geq n$, every local minimum satisfying the sign condition attains F_*. This conclusion applies to the original absolute edge scores, including zero overlaps where the ambient objective need not be differentiable.

Proof. We treat deficient and full row rank separately. In the first case, we move one class along an unused direction and average over classes; in the second, we lift the Gram descent direction directly. Apply the row signs first and take $\sigma_v = 1$ without loss of generality. Let $X = HH^\top$ be the current Gram matrix and let X_* be the coloring Gram matrix, with entries one within each class and zero between classes. The loss depends only on within-color entries and graph-edge entries. The radial inequality is

$$\langle g, X_* - X \rangle \leq -\mu_F(F(X) - F_*).$$

Here g is a valid joint edge subgradient of the oriented Gram objective. Its inner product with this direction is independent of the absolute-value choices at zero edges, since the corresponding direction entries are zero.

For (i), rank deficiency supplies a unit vector e orthogonal to every row of H. For a color class V_i of size n_i, define the classwise curve

$$h_v^{(i)}(t) = \begin{cases} (h_v + te)/\sqrt{1+t^2}, & v \in V_i, \\ h_v, & v \notin V_i. \end{cases} \tag{44}$$

These rows remain unit and the initial squared speed is n_i. For a positive pair in V_i the inner product becomes $(X_{vw}+t^2)/(1+t^2)$, increasing toward one. For any edge incident to V_i, properness means that exactly one endpoint is in V_i, and its absolute inner product becomes $|X_{vu}|/\sqrt{1+t^2}$. All other entries used by the loss remain unchanged. Thus no term of either objective increases along this curve. The scalar loss is a smooth even function of t near zero along the curve, including when a positive score or an edge score is zero. Derivatives with respect to t^2 below are right derivatives.

To find a class with sufficiently negative curvature, let $b_i = dF(H^{(i)}(t)H^{(i)}(t)^\top)/d(t^2)\big|_0$ be its derivative with respect to squared time. Summing the score derivatives across color classes gives $1 - X_{vw}$ on every positive entry and $-|X_{vu}|$ on every edge. Each edge receives its two half-contributions from its endpoint classes. Hence

$$\sum_i b_i = \langle g, X_* - X\rangle \le -\mu_F(F(X) - F_*).$$

Since $\sum_i n_i = n$, some class has $b_i/n_i \le -\mu_F(F(X) - F_*)/n$. Replacing t in its curve by $t/\sqrt{n_i}$ gives unit initial speed and second loss derivative $2b_i/n_i$, proving (43).

For (ii), X is positive definite. Set $D = X_* - X$ and

$$V = \tfrac{1}{2}DX^{-1}H.$$

Then $VH^\top + HV^\top = D$. Its diagonal is zero, so $\langle V_v, h_v\rangle = 0$ for each row. Row-normalizing $H + tV$ gives a feasible curve with Gram derivative D at zero. Positive entries at zero increase to first order, whereas zero edge entries remain zero to first order. The one-sided loss derivative is $\langle g, D\rangle < 0$, which rules out a local minimum.

If $d \ge n$, either $\operatorname{rank}(H) = n$ or $\operatorname{rank}(H) < d$. The two cases exhaust all matrices, so any nonglobal point satisfying the sign condition has a feasible descent curve. Undoing the row signs proves the stated form. □

The classwise curve changes only positive pairs in that class and edges incident to it. Its unused direction therefore needs to be orthogonal only to those rows, giving a weaker rank condition.

Corollary 2 (Local rank condition)**.** *Let $I_E = \{i : \exists v \in V_i,\ \deg(v) > 0\}$ index classes containing a nonisolated vertex, and let $n_E = \sum_{i\in I_E} |V_i|$ count their vertices. Under the sign condition of Thm. 7, replace its ambient rank-deficiency hypothesis by*

$$\operatorname{rank} H_{V_i \cup \mathcal{N}_G(V_i)} < d \qquad (i \in I_E), \qquad \mathcal{N}_G(V_i) := \bigcup_{v\in V_i} \mathcal{N}_G(v). \tag{45}$$

Then every point with $F > F_*$ has a unit-speed escape curve whose second loss derivative is at most $-2\mu_F(F - F_*)/n_E$. In particular, it cannot be a local minimum.

Proof. We localize the preceding construction and omit classes whose loss terms are constant. For each class $i \in I_E$, choose its own unit direction e_i orthogonal to the rows indexed by $V_i \cup \mathcal{N}_G(V_i)$. The curve (44) has the same within-class and incident-edge score formulas with this direction. Cross entries involving nonneighbors do not enter the objective. Thus each coefficient b_i and their summed radial identity are unchanged. Classes outside I_E have $b_i = 0$. Averaging only over I_E gives a class with $b_i/|V_i| \le -\mu_F(F - F_*)/n_E$, and the same unit-speed rescaling proves the result. A singleton class with incident edges must remain in I_E: it can affect the log loss through its appearance among another anchor's negatives. □

From free embeddings to a shared encoder. These curves apply to coherent-sign configurations with unequal class sizes, arbitrary degrees, and orthogonal within-color pairs; collapse is unnecessary. Rank deficiency supplies a descent curve whenever the relevant unused direction exists, while the conclusion for every local minimum uses $d \ge n$. Applying this landscape result to a training trajectory would also require control of its signs. A shared encoder creates a further question: does its Jacobian realize the embedding descent direction? A singular Jacobian can produce a parameter-space stationary point even when such a direction exists. The full-rank readout construction in Sec. F addresses this issue for the smooth surrogate.

F SHARED RELU ENCODERS AND A NONLINEAR EXTENSION OF MEI ET AL.

The free-embedding landscape leaves two questions for a shared encoder: whether its parameters realize the required descent directions, and how the output dimension affects attainable accuracy. We address them by combining exact interpolation on the finite support with a nonlinear extension of the landscape inequality of Mei et al. (2017, Theorem 2). Their Gaussian tangent perturbation controls approximate second-order points of a linear SDP factored over product spheres. For a nonlinear objective, the same argument adds a Gram-curvature term; radial domination then turns it into a loss bound. We apply this bound to a smooth oriented surrogate and specify the readout updates that reach its guarantees.

F.1 EXACT INTERPOLATION WITH FROZEN RANDOM NODE FEATURES

Proof of Prop. 1. We construct nonnegative hinge features that interpolate the prescribed outputs, then pass those features unchanged through every hidden layer. Let $L \geq 1$ be the number of hidden layers and let $W_\ell \geq \max\{1, N-1\}$ be the width of layer ℓ. The N independent atomless node features are pairwise distinct almost surely. Condition on such a realization and order them as $x_1 < \cdots < x_N$.

For $N \geq 2$, the first hidden layer computes $z_j(t) = (t - x_j)_+$ in its first $N-1$ coordinates and zero in every remaining coordinate. In each later layer, set the leading $(N-1) \times (N-1)$ weight block to the identity and all other weights and biases to zero. Since $z_j(t) \geq 0$, ReLU preserves these coordinates. Thus the final hidden layer contains the same hinge features at every depth. A selected affine readout has the form $g_\theta(t) = b + \sum_{j=1}^{N-1} a_j z_j(t)$, with $b, a_j \in \mathbb{R}^d$; unused output coefficients are zero.

For $N \geq 2$ and unit targets $y_1, \ldots, y_N \in \mathbb{R}^d$, define the interpolation slopes $s_j = (y_{j+1} - y_j)/(x_{j+1} - x_j)$. Choose the readout to be

$$g(t) = y_1 + s_1(t - x_1)_+ + \sum_{j=2}^{N-1} (s_j - s_{j-1})(t - x_j)_+.$$

This piecewise-linear function satisfies $g(x_i) = y_i$: it starts at y_1 and has slope s_j on $[x_j, x_{j+1}]$. For $N = 1$, set every hidden layer to zero and choose the output bias $g(t) = y_1$. In both cases the sampled outputs are nonzero and normalization leaves them unchanged. Each objective has a free minimizer by continuity on the finite product of compact spheres, with the supplied colorings and eligible-anchor sets fixed. Interpolating that minimizer proves attainment and equality of minimum values for the shared encoder. □

For optimization, freeze the constructed hidden layers and retain the output bias together with the $N-1$ hinge coefficients as the selected readout (only the bias when $N = 1$). Its final feature vector and evaluation matrix are

$$\phi(t) = \big(1, (t - x_1)_+, \ldots, (t - x_{N-1})_+\big)^\top, \qquad \Phi_{ij} = \phi_j(x_i).$$

The $N \times N$ matrix Φ is lower triangular, with diagonal $1, x_2 - x_1, \ldots, x_N - x_{N-1}$, so it is invertible; for $N = 1$ it is the matrix (1). A selected readout $B \in \mathbb{R}^{N \times d}$ produces raw outputs ΦB. Additional hidden width and depth preserve these selected features by construction. All hidden parameters and unused readout coefficients remain fixed in the square-matrix gradient method below.

The complete final-layer feature matrix, including its bias column, has $W_L + 1$ columns and contains the invertible selected matrix as a submatrix. It therefore has full row rank. The readout certificate and lifted optimizer also permit any fixed feature matrix $\Phi \in \mathbb{R}^{N \times m}$ with this rank property, using its Moore–Penrose right inverse $\Phi^\dagger$. That general case allows all m readout coefficients to vary; the gradient convergence theorem uses the selected N coefficients and their square matrix. In either case, the rank condition holds jointly across all vertices in the finite support. The construction guarantees it by freezing the sampled node features and choosing the hidden weights as above; resampled features and arbitrary random hidden layers need a separate argument.

F.2 A NONLINEAR GRAM LANDSCAPE INEQUALITY

Full row rank lets us transfer embedding directions to readout parameters. We first bound the loss at embeddings with small negative curvature, using the elliptope $\mathcal{E}_N = \{X \succeq 0 : \operatorname{diag} X = \mathbf{1}\}$ of feasible Gram matrices and the product sphere $\mathcal{M}_d = (\mathbb{S}^{d-1})^N$, equipped with the Frobenius metric. For a symmetric matrix G, its linear objective oscillation over the elliptope is

$$\operatorname{osc}_{\mathcal{E}_N}(G) = \max_{Z \in \mathcal{E}_N} \langle G, Z \rangle_F - \min_{Z \in \mathcal{E}_N} \langle G, Z \rangle_F.$$

Theorem 8 (Nonlinear extension of the Mei et al. inequality)**.** *Let F be C^2 on a neighborhood of $\mathcal{E}_N$, and let $Y \in \mathcal{E}_N$ minimize F over $\mathcal{E}_N$, with value F_*. Suppose*

$$\langle \nabla F(X), X - Y \rangle_F \geq \mu\big(F(X) - F_*\big), \qquad X \in \mathcal{E}_N,$$

for some $\mu > 0$. Let $d \geq 2$, $H \in \mathcal{M}_d$, $X = HH^\top$, and $f(H) = F(HH^\top)$. Suppose its Riemannian Hessian satisfies

$$\operatorname{Hess} f(H)[U, U] \geq -\varepsilon \|U\|_F^2 \qquad (U \in T_H \mathcal{M}_d).$$

Factor the comparator as $Y = VV^\top$, with unit rows v_i, and let $Z \in \mathbb{R}^{d \times N}$ have independent standard normal entries. Define the random tangent rows u_i, their induced Gram perturbation D, and the comparison matrix Q by

$$u_i = (I - h_i h_i^\top) Z v_i, \quad D = HU^\top + UH^\top, \quad Q = Y \odot X \odot X.$$

Here $\odot$ denotes entrywise multiplication. Then

$$F(X) - F_* \leq \frac{\langle \nabla F(X), Q - Y \rangle_F + \frac{1}{2}\mathbb{E}\nabla^2 F(X)[D, D]}{\mu(d-1)} + \frac{N\varepsilon}{2\mu}. \tag{46}$$

The first numerator term is at most $\operatorname{osc}_{\mathcal{E}_N}(\nabla F(X))$. If $\nabla^2 F(X)[D, D] \leq L\|D\|_F^2$ for $L \geq 0$, the second is at most $2LN(N-1)$. No first-order stationarity assumption is needed.

Proof. We average the factor Hessian over Gaussian tangent directions coupled through the comparator Y. Radial domination will convert the resulting first-order term into the loss excess. Set $G = \nabla F(X)$ and $\Lambda = \operatorname{Diag}(\operatorname{diag}(GX))$. Differentiating along product-sphere geodesics gives

$$\operatorname{Hess} f(H)[U, U] = 2\langle G - \Lambda, UU^\top \rangle_F + \nabla^2 F(X)[D, D]. \tag{47}$$

The Gaussian perturbation used in Mei et al. (2017, Section 3.2), rescaled here to variance one, satisfies

$$\mathbb{E}UU^\top = (d-2)Y + Q, \qquad \mathbb{E}\|U\|_F^2 = N(d-1).$$

Indeed, $\mathbb{E}\langle u_i, u_j \rangle = Y_{ij} \operatorname{tr}((I - h_i h_i^\top)(I - h_j h_j^\top)) = Y_{ij}(d - 2 + X_{ij}^2)$. Since $\operatorname{tr}\Lambda = \langle G, X \rangle_F$, taking expectations in (47) yields

$$2(d-1)\langle G, X - Y \rangle_F \leq 2\langle G, Q - Y \rangle_F + \mathbb{E}\nabla^2 F(X)[D, D] + \varepsilon N(d-1).$$

Radial domination now yields (46). To bound its numerator, the Schur product theorem gives $Q \in \mathcal{E}_N$, so the first term is at most the linear objective oscillation. For the curvature term, a direct covariance calculation gives

$$\mathbb{E}D_{ij}^2 = 2(1 - X_{ij}^2)(1 - Y_{ij}X_{ij}) \leq 4, \qquad D_{ii} = 0. \tag{48}$$

Thus $\mathbb{E}\|D\|_F^2 \leq 4N(N-1)$. When F is linear, its Gram Hessian vanishes, $\mu = 1$ is valid, and the oscillation bound recovers the minimization form of the published Mei et al. inequality. □

F.3 APPLICATION TO THE TWO CONTRASTIVE OBJECTIVES

Applying the nonlinear inequality from arbitrary embeddings requires a smooth objective with radial domination on the entire elliptope. Absolute values fail to be C^2 at zero overlaps, and within-color signs need not be coherent. We therefore define a smooth *oriented surrogate* F_δ, for $\delta > 0$: in either objective use the signed linear score X_{vw} for every positive pair and

$$\psi_\delta(X_{vu}) = \sqrt{X_{vu}^2 + \delta^2} - \delta \tag{49}$$

for every negative edge. All averaging, self-positive conventions, and eligible-anchor sets remain those of the corresponding objective. We analyze optimization of this surrogate and bound the resulting original loss. The surrogate is not used in the reported experiments.

For the finite-support population, enumerate all N graph–vertex pairs, form one Gram matrix, and let F_δ be the probability-weighted average of the graph losses. Cross-graph entries are unused. Let Y be the block-diagonal coloring Gram matrix with blocks X^{c_G}, let $F_* = F_\delta(Y)$ be its loss, and let Δ be the maximum degree over the support. The radial constant $\widehat{\mu}_F$, score-derivative bound A_F, and score-curvature bound $B_{F,\delta}$ are

$$\widehat{\mu}_{\mathrm{R}} = \frac{1+\Delta/a_\tau}{1+\Delta a_\tau^2}, \qquad A_{\mathrm{R}} = \frac{1}{4\tau}, \qquad B_{\mathrm{R},\delta} = \frac{3}{2\tau^2} + \frac{1}{4\tau\delta}, \tag{50}$$

$$\widehat{\mu}_{\log} = \frac{2}{\tau(a_\tau^2-1)}, \qquad A_{\log} = \frac{1}{\tau}, \qquad B_{\log,\delta} = \frac{4}{\tau^2} + \frac{1}{\tau\delta}. \tag{51}$$

The subscripts R and log refer to (1) and (2), respectively. These bounds retain each objective's degree and positive-coverage conventions.

Theorem 9 (Output dimension and approximate optimality)**.** *For either objective and $d \geq 2$, suppose*

$$\mathrm{Hess}(F_\delta \circ \mathrm{Gram})(H) \succeq -\varepsilon I \quad \text{on } T_H\mathcal{M}_d.$$

Let $F_{\mathrm{abs}}(H)$ denote the corresponding original absolute-similarity objective. Then

$$0 \leq F_{\mathrm{abs}}(H) - F_* \leq \frac{A_F + 2B_{F,\delta}}{\widehat{\mu}_F(d-1)} + \frac{N\varepsilon}{2\widehat{\mu}_F} + A_F\delta. \tag{52}$$

The theorem allows arbitrary class sizes and does not require sign coherence, initial collapse, or $d \geq N$. Here F_* is the full-Gram minimum; it need not be attained in dimension d. Taking δ of order $(d-1)^{-1/2}$ gives a bound of order $(d-1)^{-1/2} + N\varepsilon$, with constants depending on τ and Δ.

Proof. We verify radial domination for the surrogate, bound the two nonlinear terms in Thm. 8, and then control the error in returning to absolute scores. The smoothing satisfies

$$\psi_\delta(0) = 0, \quad 0 \leq |x| - \psi_\delta(x) \leq \delta, \quad |\psi_\delta'(x)| \leq 1, \quad 0 \leq \psi_\delta''(x) \leq \delta^{-1}, \quad x\psi_\delta'(x) \geq \psi_\delta(x).$$

Every linear positive score is at most one and every smoothed negative score is nonnegative. The coloring matrix Y attains all termwise lower bounds, so it minimizes F_δ with the same baseline as the original absolute objective.

Along the segment from X to Y, a positive deficit $1 - X_{vw}$ lies in $[0,2]$. A negative score decreases at least as fast as its present value by the last displayed inequality. The proof of Lem. 3 therefore applies with two changes. For the InfoNCE variant, $S \leq Da_\tau^2$ since a negative score minus a positive score is at most two. For the log objective, $y_w = (1 - X_{vw})/\tau \in [0, 2/\tau]$, so the constant $T/(e^T - 1)$ uses $T = 2/\tau$. These changes give the two constants $\widehat{\mu}_F$ in (51) on the entire elliptope. Averaging graphwise inequalities preserves the bound with the maximum support degree.

It remains to control the nonlinear terms in Thm. 8. For Y just chosen, $Q - Y$ vanishes on every negative edge, and on a positive pair it equals $X_{vw}^2 - 1$. The loss is decreasing in each positive score. The sum of absolute positive-score derivatives, and separately the sum of negative-score derivatives, is at most A_F. Thus

$$\langle \nabla F_\delta(X), Q - Y\rangle_F \leq A_F.$$

Here and below derivatives can be indexed by individual score occurrences before identifying repeated Gram entries; averaging preserves all bounds.

To control the remaining curvature term, we bound the entrywise absolute sum of the score Hessian by $3/(2\tau^2)$ for one InfoNCE-variant summand and $4/\tau^2$ for one log anchor. For the former, write the summand as $-p$, where p is the positive-coordinate softmax probability. Differentiation bounds the Hessian sum by

$$\tau^{-2}\{4p(1-p)^2 + 2p(1-p^2)\} = 2\tau^{-2}p(1-p)(3-p) \leq \frac{3}{2\tau^2}.$$

For the latter, write the anchor as $\log(A+B) - \log A$. Each log-sum-exp Hessian is a scaled probability covariance with entrywise absolute sum at most $2/\tau^2$. Applying the triangle inequality

gives $4/\tau^2$. Composition with ψ_δ does not increase the first Hessian contribution because $|\psi'_\delta| \le 1$; its additional diagonal contribution is at most A_F/δ. Therefore the total Hessian sum is at most $B_{F,\delta}$. By (48) and Cauchy–Schwarz, every product of two score perturbations has expected absolute value at most four, giving

$$\mathbb{E}\nabla^2 F_\delta(X)[D, D] \le 4B_{F,\delta}.$$

Equation (46) now bounds $F_\delta(X) - F_*$. Finally, replacing signed positive scores by their absolute values can only lower either loss, while replacing $\psi_\delta(x)$ by $|x|$ increases it by at most $A_F\delta$. Hence

$$F_{\mathrm{abs}}(H) \le F_\delta(HH^\top) + A_F\delta,$$

which proves (52). □

F.4 A SHARED-PARAMETER CERTIFICATE AND A CONVERGENT OPTIMIZER

The preceding bound concerns curvature in embedding coordinates. Full row rank of the frozen feature matrix lets us realize each such direction in the shared readout, with a cost determined by its smallest singular value.

Corollary 3 (Shared readout certificate). *Let $\Phi \in \mathbb{R}^{N\times m}$ have full row rank and let H be obtained by normalizing the nonzero rows of ΦB. Set $r_{\max} = \max_i \|(\Phi B)_i\|$ and $\sigma = \sigma_{\min}(\Phi) > 0$. If the Euclidean readout Hessian of the smooth surrogate satisfies*

$$\nabla_B^2\big[F_\delta(\mathrm{Gram}(\mathrm{normalize}(\Phi B)))\big] \succeq -\varepsilon_B I,$$

then (52) holds with $\varepsilon = \varepsilon_B r_{\max}^2/\sigma^2$. In particular, an exact local minimum satisfies that bound with $\varepsilon = 0$.

Proof. We lift a tangent direction through the feature matrix and compare the second derivatives before and after row normalization. For any tangent $U \in T_H\mathcal{M}_d$, put $R = \mathrm{Diag}(\|(\Phi B)_i\|)$ and $\dot B = \Phi^\dagger RU$. Along the straight parameter curve $B + t\dot B$, normalization gives rows $(h_i + tu_i)/\sqrt{1 + t^2\|u_i\|^2}$. Their first two derivatives at zero agree with product-sphere geodesics. Thus the Euclidean parameter second derivative equals the Riemannian Hessian evaluated on U, even without first-order stationarity. Since $\|\dot B\|_F \le r_{\max}\|U\|_F/\sigma$, the Hessian bound transfers as claimed. □

The certificate also suggests an optimizer whose metric preserves these embedding directions. Restrict the readout to $B = \Phi^\dagger H$ with $H \in \mathcal{M}_d$. Equip it with the metric $\langle \dot B_1, \dot B_2\rangle = \langle \Phi\dot B_1, \Phi\dot B_2\rangle_F$. This makes the shared-parameter manifold isometric to $\mathcal{M}_d$; a sphere update H^+ is implemented exactly by

$$B^+ = B + \Phi^\dagger(H^+ - H). \tag{53}$$

With the hidden features fixed, this lift allows us to implement each product-sphere step in shared parameters.

Theorem 10 (Convergence to prescribed original-loss accuracy). *Fix $\delta > 0$, $d \ge 2$, and $\varepsilon > 0$. From any $H_0 \in \mathcal{M}_d$, a negative-curvature method with the lift (53) reaches a shared readout satisfying (52) in finitely many steps. More explicitly, let $L_3 > 0$ bound the absolute third derivative of $f_\delta = F_\delta \circ \mathrm{Gram}$ along all unit-speed product-sphere geodesics. Until $\mathrm{Hess}\, f_\delta(H) \succeq -\varepsilon I$, choose a unit tangent U with curvature less than $-\varepsilon$, orient it so that $\langle \mathrm{grad}\, f_\delta(H), U\rangle \le 0$, and take the geodesic step $t = \varepsilon/L_3$. The number of such steps is at most*

$$\frac{3L_3^2\big(f_\delta(H_0) - F_*\big)}{\varepsilon^3}.$$

For any target excess $\zeta > 0$, it suffices to choose

$$\delta \le \frac{\zeta}{3A_F}, \qquad d - 1 \ge \frac{3(A_F + 2B_{F,\delta})}{\widehat\mu_F\zeta}, \qquad \varepsilon \le \frac{2\widehat\mu_F\zeta}{3N}. \tag{54}$$

The resulting original absolute objective is at most $F_* + \zeta$. The depth-L ReLU construction above requires $W_\ell \ge \max\{1, N-1\}$ at every hidden layer, with the hidden parameters and unused readout coefficients fixed. For fixed temperature and support degree, the sufficient output dimension is $O(\zeta^{-2})$.

Proof. Each curvature step decreases the surrogate by a uniform amount; its lower bound then limits the number of steps. For fixed $\delta > 0$, the surrogate is smooth. Compactness of the product spheres and their unit tangent bundle gives a finite uniform third-derivative bound; increase it if necessary to make it positive. Taylor's theorem along the chosen geodesic gives

$$f_\delta(H^+) \le f_\delta(H) - \frac{\varepsilon}{2}t^2 + \frac{L_3}{6}t^3 = f_\delta(H) - \frac{\varepsilon^3}{3L_3^2}.$$

Since $f_\delta \ge F_*$, only the stated number of strict decreases is possible before the Hessian stopping condition holds. Apply Thm. 9 and the exact readout lift. Each of the three error terms in (52) is at most $\zeta/3$ under (54). Choosing $\delta = \zeta/(3A_F)$ gives the stated output-dimension order. The iteration bound counts the specified ideal curvature steps; computing a direction of negative curvature has a separate arithmetic cost. □

F.5 Global gradient convergence at full graph dimension

The curvature method provides finite-accuracy guarantees at a chosen output dimension. If $d \ge \max_G n_G$, every nonglobal critical point of the surrogate instead has a descent direction, which yields an asymptotic guarantee for gradient descent. We use the interpolation coordinates to realize that descent, equivalently applying a fixed preconditioner to the original readout weights.

Theorem 11 (Global convergence of preconditioned readout descent)**.** *Use the selected invertible feature matrix Φ above and parameterize its N trainable readout coefficients by $B = \Phi^{-1}C$, keeping all hidden parameters and unused output coefficients fixed. Its raw outputs on the finite support are C. Fix $\delta > 0$, $r_0 > 0$, and $d \ge \max_G n_G$, and define*

$$\mathcal{J}(C) = F_\delta(\mathrm{Gram}(\mathrm{normalize}(C))).$$

Draw C_0 from any distribution with a Lebesgue density supported on $\min_i \|(C_0)_i\| \ge r_0$. There is a finite $L > 0$, independent of this draw, such that for every fixed $0 < \eta < 1/L$, the iterates

$$C_{t+1} = C_t - \eta\nabla\mathcal{J}(C_t), \qquad B_t = \Phi^{-1}C_t \tag{55}$$

converge almost surely to shared parameters attaining F_*. The limiting embeddings also attain the minimum of the corresponding original absolute-similarity objective. In these selected readout coordinates the update is

$$B_{t+1} = B_t - \eta(\Phi^\top\Phi)^{-1}\nabla_B[\mathcal{J}(\Phi B_t)].$$

The result gives asymptotic convergence, without an iteration bound. The random initialization of the selected readout is separate from the frozen random node features.

Proof. We show that the iterates remain bounded and converge to one critical point. The dimension hypothesis then makes every nonglobal critical point a strict saddle, whose basin has measure zero. On the open set $\mathcal{D} = \{C : \min_i \|C_i\| > r_0/2\}$ the objective is real analytic and has uniformly bounded Hessian: normalized outputs range over compact spheres, and the first two derivatives of row normalization are uniformly bounded away from zero row norm. Choose $L > 0$ bounding $\|\nabla^2\mathcal{J}(C)\|_{\mathrm{op}}$ on $\mathcal{D}$.

To keep normalization well defined, use rowwise positive-scale invariance: $\langle C_i, \nabla_i\mathcal{J}(C)\rangle = 0$. Orthogonality of the row and its update gives

$$\|(C_{t+1})_i\|^2 = \|(C_t)_i\|^2 + \eta^2\|\nabla_i\mathcal{J}(C_t)\|^2.$$

Every row norm is nondecreasing, and the entire straight update segment stays in $\mathcal{D}$. The Hessian bound along that segment implies

$$\mathcal{J}(C_t) - \mathcal{J}(C_{t+1}) \ge \frac{\eta}{2}\|\nabla\mathcal{J}(C_t)\|_F^2, \qquad \|C_t\|_F^2 \le \|C_0\|_F^2 + 2\eta(\mathcal{J}(C_0) - F_*).$$

Thus the iterates stay in a compact subset of $\mathcal{D}$, their gradients tend to zero, and they satisfy the strong-descent inequality

$$\mathcal{J}(C_t) - \mathcal{J}(C_{t+1}) \ge \tfrac{1}{2}\|\nabla\mathcal{J}(C_t)\|_F\|C_{t+1} - C_t\|_F.$$

The local analytic convergence argument of Absil et al. (2005, Theorem 3.4) applies near an accumulation point inside $\mathcal{D}$. Combined with strong descent, it shows that the whole sequence converges to one critical point.

It remains to exclude nonglobal limits. We show that each is a strict saddle by choosing a support graph with positive excess. The proof of Thm. 7 extends to F_δ without the sign condition: $(X_{vw}+t^2)/(1+t^2)$ increases every linear positive score even if $X_{vw} < 0$, and the incident smoothed negative scores decrease. Summing the color-curve second derivatives again gives the strictly negative radial derivative. If that graph's factor has rank less than d, an unused direction supplies a curve with zero first derivative and negative second derivative. If it has full row rank, the Gram direction $X^{c_G} - X_G$ lifts to a first-order descent curve, contradicting stationarity. Since $d \geq n_G$, these cases exhaust all factors. Hold the other graphs fixed and lift the curve to C by keeping each raw row norm fixed. At a critical point its negative second derivative is a negative Euclidean Hessian direction.

Finally, the gradient map $T(C) = C - \eta\nabla\mathcal{J}(C)$ is a C^1 self-map of $\mathcal{D}$ and $DT = I - \eta\nabla^2\mathcal{J}$ is nonsingular everywhere. Every strict saddle is an unstable fixed point of this map. The measure-zero basin result of Lee et al. (2019, Theorem 2 in the preprint version) applies on this open manifold. An initialization with a density therefore almost surely avoids every strict-saddle limit; the critical point reached by the whole sequence must be global. At a global surrogate minimum, every score required by a nonconstant term attains its termwise optimum. Since $\psi_\delta(x) = 0$ exactly when $x = 0$, the same embeddings minimize the original absolute objective, with its stated coverage exceptions. Multiplication by the fixed inverse Φ^{-1} gives convergence of B_t. □

Relation to the exact-objective results. Full row rank supplies the shared parameter directions used by both optimizers. The curvature method reaches prescribed original-loss accuracy at finite output dimension, while preconditioned descent reaches a global minimum almost surely when $d \geq \max_G n_G$. Both optimize the smooth oriented surrogate on the finite support with frozen features. For the exact objectives, the Gram dynamics and coherent-sign landscape theorems apply without smoothing. Convergence of the experimental AdamW updates and generalization to unseen random node features remain separate questions.

G Full-embedding margins and semidefinite certificates

The normalized analysis describes finite-loss geometry. Without row normalization, scaling the embeddings instead makes the loss depend asymptotically on their smallest positive-to-negative margin. We prove the sharp margin and ray-decay statements of Thm. 4, then derive certificates for recognizing optimal margin configurations. Throughout, vertex embeddings are independently parameterized, and (1) includes the anchor among its positives. The final paragraph treats the different behavior of (2).

Proof of Thm. 4. Self-positive pairs bound the margin by each squared row norm. We use their equality conditions to identify the optimal geometry, then sandwich the loss along a ray between two exponentials with the same rate. Every vertex v has a neighbor u. Taking $w = v$ among its positives gives

$$\mathfrak{m}(U) \leq \|u_v\|^2 - |\langle u_v, u_u\rangle| \leq \|u_v\|^2. \tag{56}$$

Summing over all vertices yields $n\mathfrak{m}(U) \leq \|U\|_{\mathrm{F}}^2 = 1$. Assigning a unit orthonormal vector to each used color and scaling all rows by $n^{-1/2}$ attains $1/n$ when $d \geq k$.

To identify the equality case, suppose the margin is $1/n$. Every squared row norm is at least $1/n$, and their sum is one, so each equals $1/n$. For every same-color positive w and neighbor u,

$$\frac{1}{n} \leq |\langle u_v, u_w\rangle| - |\langle u_v, u_u\rangle| \leq |\langle u_v, u_w\rangle| \leq \|u_v\|\|u_w\| = \frac{1}{n}. \tag{57}$$

Equality throughout forces zero edge overlaps and equality in Cauchy–Schwarz within each supplied class. Thus each class lies on one line, with the displayed row norms and arbitrary vertex signs. Edges between classes force orthogonality of the corresponding unit prototypes. Conversely, that geometry attains margin $1/n$. The supplied coloring need not be optimal: only quotient-adjacent prototypes are forced to be orthogonal.

For the ray statement, write $m = \mathfrak{m}(U) > 0$ for the margin and $D = \Delta(G) \geq 1$ for the maximum degree. The negative-to-positive exponential sum for an anchor-positive pair is

$$S_{vw}(r) = \sum_{u \in \mathcal{N}_G(v)} \exp\left[-\frac{r^2}{\tau}\left(|\langle u_v, u_w\rangle| - |\langle u_v, u_u\rangle|\right)\right]. \tag{58}$$

Then

$$1 + \ell_{\mathrm{R}}(rU) = \frac{1}{n}\sum_{v} \frac{1}{|V_{c_G(v)}|} \sum_{w \in V_{c_G(v)}} \frac{S_{vw}(r)}{1 + S_{vw}(r)}. \tag{59}$$

Every $S_{vw}(r)$ is at most $De^{-r^2 m/\tau}$. At least one pair has a summand equal to $e^{-r^2 m/\tau}$, since the minimum defining m ranges over a nonempty finite set. If $b = \max_i |V_i|$, monotonicity of $x/(1+x)$ and the positive averaging weights give

$$\frac{e^{-r^2 m/\tau}}{nb(1 + De^{-r^2 m/\tau})} \leq 1 + \ell_{\mathrm{R}}(rU) \leq De^{-r^2 m/\tau}. \tag{60}$$

Taking logarithms and multiplying by $-\tau/r^2$ gives the common exponential rate in (10). The margin thus determines decay along each prescribed ray, without a hypothesis on a gradient trajectory. □

Exact treatment of isolated vertices. Isolated vertices supply positive pairs but impose no anchor margin of their own. Their optimal norms can therefore differ from those of nonisolated vertices. Let G have at least one edge. In each supplied class i, let m_i count the nonisolated vertices and let z_i count the isolated vertices. Define the classwise norm cost ϕ and its total Λ by

$$\Lambda := \sum_i \phi(m_i, z_i), \qquad \phi(m, z) = \begin{cases} 0, & m = 0, \\ m + z, & m > 0 \text{ and } z \leq m, \\ 2\sqrt{mz}, & z > m > 0. \end{cases} \tag{61}$$

Then, provided d is at least the number of classes with $m_i > 0$,

$$\min_{\mathfrak{m}(U) \geq 1} \|U\|_{\mathrm{F}}^2 = \Lambda, \qquad \max_{\|U\|_{\mathrm{F}}^2 = 1} \mathfrak{m}(U) = \Lambda^{-1}, \tag{62}$$

where the margin includes only anchors with nonempty negative sets. To prove the lower bound, fix a class with $m = m_i > 0$ and let t be the smallest squared norm of its nonisolated vertices. Self-positive constraints give $t \geq 1$. Each isolated vertex in the class is a positive for a nonisolated anchor with that minimum norm, so Cauchy–Schwarz forces its squared norm to be at least $1/t$. The class therefore contributes at least $mt + z/t$. Minimizing this expression over $t \geq 1$ gives (61). Classes with $m_i = 0$ contribute a nonnegative amount and have no margin constraints.

To attain the bound, choose orthonormal vectors q_i on the classes with $m_i > 0$ and set

$$t_i = \max\{1, \sqrt{z_i/m_i}\}, \qquad u_v = \sqrt{t_i}\, q_i \ \ (\deg v > 0), \qquad u_v = t_i^{-1/2} q_i \ \ (\deg v = 0, m_i > 0), \tag{63}$$

and set all rows in classes with $m_i = 0$ to zero. Every edge overlap vanishes. A positive score at a nonisolated anchor is either $t_i \geq 1$ or one, so all required margins are at least one. These rows have total squared norm Λ. Homogeneity of degree two then gives the normalized identity. The ray-tail proof remains valid, because isolated anchors contribute zero to $1 + \ell_{\mathrm{R}}$. On an edgeless graph, the InfoNCE variant is constant and the margin index set is empty.

G.1 The margin problem with consistent positive signs

The sharp value identifies the optimal margin, but a certificate must recognize it from a candidate embedding. Consistent positive signs make the constraints linear in its Gram matrix, allowing a semidefinite certificate. Assume again that G has no isolated vertices. If fixed vertex sign changes make all within-color positive overlaps nonnegative, apply them first; absolute negative overlaps and row norms are unchanged. In this oriented chamber the unit-margin inequalities become

$$X_{vw} - \zeta X_{vu} \geq 1 \quad (v \in [n],\ w \in V_{c_G(v)},\ u \in \mathcal{N}_G(v),\ \zeta \in \{\pm 1\}), \qquad X = UU^{\top}. \tag{64}$$

The inequalities themselves enforce strictly positive within-color overlaps. Let $\mathcal{I}$ index these inequalities and let A_ℓ be the symmetric matrix satisfying $\langle A_\ell, X\rangle = X_{vw} - \zeta X_{vu}$, with $\langle A, X\rangle = \operatorname{tr}(AX)$. The convex lifted problem is

$$\min_{X \succeq 0} \operatorname{tr} X \quad \text{subject to} \quad \langle A_\ell, X\rangle \geq 1 \quad (\ell \in \mathcal{I}). \tag{65}$$

The optimum is n by Thm. 4, with no dimension restriction on the PSD variable. Factorizing $X = UU^\top$ recovers the oriented margin problem at width d. We next use the dual slack to certify optimality of a factor and show when its second-order conditions suffice.

Proposition 2 (Dual and factor certificates)**.** *Let U satisfy (64). Given nonnegative constraint multipliers λ_ℓ, define the dual slack*

$$Z := I - \sum_{\ell \in \mathcal{I}} \lambda_\ell A_\ell. \tag{66}$$

If

$$ZU = 0, \qquad \lambda_\ell(\langle A_\ell, UU^\top\rangle - 1) = 0 \ \ (\ell \in \mathcal{I}), \qquad Z \succeq 0, \tag{67}$$

then U is globally optimal and $\|U\|_{\mathrm{F}}^2 = n$. The PSD condition follows from the first two conditions and the second-order test

$$\operatorname{tr}(V^\top Z V) \geq 0 \quad \textit{for all } V \textit{ with } \langle A_\ell, UV^\top + VU^\top\rangle = 0 \ \ (\ell \in \mathcal{I}), \tag{68}$$

whenever $\operatorname{rank} U < d$. It also follows from $ZU = 0$ if $\operatorname{rank} U = n$. Hence, at width $d \geq n$, every feasible point satisfying the displayed first- and second-order conditions is globally optimal. Independently of multiplier existence, every local minimum of the oriented factor problem is global when $d \geq n$.

Proof. We first use primal-dual equality to certify a candidate, then derive positivity of the dual slack from its rank and second-order conditions. A separate lifting argument handles local minima without multipliers. For any feasible $Y \succeq 0$ and any multipliers with $Z \succeq 0$,

$$\operatorname{tr} Y - \sum_\ell \lambda_\ell = \langle Z, Y\rangle + \sum_\ell \lambda_\ell(\langle A_\ell, Y\rangle - 1) \geq 0. \tag{69}$$

For $Y = UU^\top$, stationarity and complementarity make both terms zero. The candidate therefore attains the dual lower bound and is globally optimal.

If $\operatorname{rank} U < d$, choose a unit vector $a \in \mathbb{R}^d$ with $Ua = 0$. For any $x \in \mathbb{R}^n$, the perturbation $V = xa^\top$ satisfies $UV^\top + VU^\top = 0$. Thus (68) gives $x^T Z x \geq 0$, proving $Z \succeq 0$. If U has full row rank, $ZU = 0$ directly implies $Z = 0$. At width $d \geq n$, these two cases exhaust all ranks. This null-column argument is standard in semidefinite factorization (Boumal et al., 2016); here the displayed assumptions give the certificate directly, without constraint qualifications or generic-cost hypotheses.

For the local-minimum statement, write $X = UU^\top$. When $d \geq n$, a singular value decomposition extends to a matrix $R \in \mathbb{R}^{n\times d}$ with $RR^\top = I$ and $U = X^{1/2}R$. Given a feasible PSD matrix Y with smaller trace, set $X_t = (1-t)X + tY$ and $U_t = X_t^{1/2}R$. The constraints of (65) are convex, so X_t is feasible. Moreover $U_t U_t^\top = X_t$, $U_t \to U$ by continuity of the PSD square root, and $\|U_t\|_{\mathrm{F}}^2 = \operatorname{tr} X_t < \operatorname{tr} X$ for $t > 0$. This contradicts local optimality. Therefore every factor local minimum maps to a global SDP optimum. □

The displayed linearized equality directions define a sufficient second-order certificate. To apply it to a computed point, one must supply the multipliers and check these conditions; the proposition alone does not establish convergence of unnormalized gradient flow.

An approximate certificate. Approximate stationarity and dual feasibility still control the norm gap. Suppose U is feasible and $Z \succeq -\delta I$ with $\delta \geq 0$. Define the stationarity residual ϵ_{s} and complementarity residual ϵ_{c} by

$$\epsilon_{\mathrm{s}} := \|ZU\|_{\mathrm{F}}, \qquad \epsilon_{\mathrm{c}} := \sum_\ell \lambda_\ell(\langle A_\ell, UU^\top\rangle - 1) \geq 0. \tag{70}$$

Then

$$0 \leq \|U\|_{\mathrm{F}}^2 - n \leq \frac{\delta\|U\|_{\mathrm{F}}^2 + \epsilon_{\mathrm{s}}\|U\|_{\mathrm{F}} + \epsilon_{\mathrm{c}}}{1+\delta}. \tag{71}$$

Indeed, $\lambda/(1+\delta)$ is dual feasible because its slack is $(Z+\delta I)/(1+\delta) \succeq 0$. Weak duality bounds n below by $\sum_\ell \lambda_\ell/(1+\delta)$, and $\|U\|_{\mathrm{F}}^2 - \sum_\ell \lambda_\ell = \langle Z, UU^\top\rangle + \epsilon_{\mathrm{c}} \le \epsilon_{\mathrm{s}}\|U\|_{\mathrm{F}} + \epsilon_{\mathrm{c}}$. Combining these estimates gives (71), so the three explicit residuals certify proximity to a global margin direction.

G.2 Why arbitrary positive signs obstruct global optimization

The preceding certificates rely on positive signs that can be made consistent by vertex flips. Increasing the embedding width does not resolve a frustrated sign pattern: the following star graph has a suboptimal local minimum at every sufficiently large width.

Proposition 3 (A suboptimal absolute-margin local minimum)**.** *For $G = K_{3,1}$, with the three leaves assigned one color and the center a second color, the full absolute unit-margin problem has a local minimum of squared norm seven, whereas its global minimum is four. This local minimum exists for every $d \ge 3$ and satisfies second-order KKT conditions for a smooth local inequality representation. Thus increasing width alone does not remove all suboptimal local margin optima.*

Proof. We construct a feasible Gram matrix whose negative leaf overlaps force a larger trace in its sign neighborhood. Explicit multipliers then verify the second-order conditions. Choose rows with Gram matrix

$$X_\star = \begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{72}$$

Its eigenvalues are $3, 3, 1, 0$, so it has a factor with three columns, and zero columns extend that factor to any larger width. All edge overlaps vanish. Leaf self-positive scores are two, leaf distinct-positive absolute scores are one, and the center self-positive score is one. Thus every margin is at least one and the trace is seven.

In a sufficiently small neighborhood of any factor of $X_\star$, the three off-diagonal leaf Gram entries remain negative. Every feasible point in this neighborhood therefore satisfies $X_{ij} \le -1$ for distinct leaves. Positive semidefiniteness applied to $(1, 1, 1, 0)$ gives

$$X_{11} + X_{22} + X_{33} \ge -2(X_{12} + X_{13} + X_{23}) \ge 6. \tag{73}$$

The center self-positive constraint gives $X_{44} \ge 1$. Hence every nearby feasible factor has squared norm at least seven, proving local minimality. The global value is four by Thm. 4.

For an explicit second-order certificate in this local sign chamber, replace each leaf positive absolute value by $-X_{ij}$ and each negative absolute value by its two linear inequalities. For each ordered distinct leaf pair (i, j), put multiplier $1/2$ on each inequality $-X_{ij} - \zeta X_{i4} \ge 1$, $\zeta \in \{\pm 1\}$. For each leaf i, put multiplier $1/6$ on each center constraint $X_{44} - \zeta X_{4i} \ge 1$. Set all other multipliers to zero. Their sum is seven, their slack is $Z = \mathrm{diag}(\mathbf{1}_3\mathbf{1}_3^\top, 0) \succeq 0$, and $ZX_\star = 0$. Thus every factor satisfies $ZU = 0$, complementarity, and the nonnegative Lagrangian Hessian $2\,\mathrm{tr}(V^\top ZV)$ on all directions. The point is globally optimal within this local signed branch, but not for the full absolute-margin problem.

The obstruction is the leaf sign cycle. Its three signs have negative product, which vertex sign switches preserve, so no switches make all three inner products positive. The example therefore lies outside the chamber covered by the coherent-positive SDP. □

The aggregate log objective has a different ray margin. The ray argument for (1) used a separate term for every positive. In (2), the sum over the nonself-positive set P_v is instead controlled at large scale by $\max_{w \in P_v} |\langle u_v, u_w\rangle|$. Its decay therefore need not reflect the all-positive margin in (8). For example, take $K_{4,2}$ with one supplied color per side. Embed the four vertices on the first side as two identical pairs at two orthogonal unit vectors, and the second side as an identical pair at a third orthogonal unit vector. Every edge overlap is zero. Along a common scaling ray, each anchor has at least one positive score growing quadratically, so every aggregate log term tends to zero. The first class remains split across two orthogonal lines. Even without isolated vertices or singleton classes, vanishing unnormalized aggregate log loss therefore permits several lines within a class. Unit normalization prevents this scaling mechanism; the finite excess-loss theorem applies in that setting.

H PROOFS AND REFINEMENTS OF THE GEOMETRIC CERTIFICATES

The certificates convert measured edge overlaps and prototype affinities into theta bounds, conflict counts, and proper colorings after repair. We begin by canceling edge overlaps with auxiliary coordinates, which lets us apply the classical orthogonal-representation bound. Throughout, $G = ([n], E)$ is a finite simple graph with $n \geq 1$ vertices. For unit representation vectors u_i satisfying $\langle u_i, u_j \rangle = 0$ on every edge and a unit handle c with nonzero overlaps, the definition of Lovász's theta function gives (Lovász, 1979)

$$\vartheta(\overline{G}) \leq \max_i \frac{1}{\langle c, u_i \rangle^2}. \tag{74}$$

The complement in this formula follows from the original definition: the vectors form an orthonormal representation of $\overline{G}$, whose distinct nonadjacent vertices are the edges of G. Our task is to construct such a representation from embeddings that are only approximately orthogonal on edges.

H.1 POSITIVE SEMIDEFINITE CORRECTION AND THE SPECTRAL BOUND

The signed edge-overlap matrix S records the correlations to be canceled, while the diagonal handle matrix H records the available overlaps with r. A spectral bound makes their correction positive semidefinite and controls the normalization cost of the lift.

Proof of Lem. 2. We construct a positive semidefinite correction whose off-diagonal edge entries cancel S, apply the handle bound to the lifted vectors, and then compare the spectral value with the unsigned and local bounds. Let $t := \lambda_{\max}(H^{-1}(I+S)H^{-1})$ denote the signed spectral value. The diagonal entries of this symmetric matrix are h_i^{-2}, so its Rayleigh quotients give $t \geq \max_i h_i^{-2} \geq 1$. By this choice of t, the correction matrix

$$X := tH^2 - I - S = H\left[tI - H^{-1}(I+S)H^{-1}\right]H \tag{75}$$

is positive semidefinite. We can therefore factor $X_{ij} = \langle w_i, w_j \rangle$ using auxiliary vectors $w_i \in \mathbb{R}^n$. Define the lifted vectors z_i, their unit normalizations u_i, and the lifted handle c by

$$z_i := (v_i, w_i), \qquad u_i := \frac{z_i}{\sqrt{t}\, h_i}, \qquad c := (r, 0). \tag{76}$$

The diagonal identity $S_{ii} = 0$ gives $\|z_i\|^2 = 1 + X_{ii} = th_i^2$, verifying the unit norm of u_i. On each edge, $X_{ij} = -S_{ij} = -\langle v_i, v_j \rangle$ cancels the original overlap and gives $\langle z_i, z_j \rangle = 0$. The unit handle c then has absolute overlap

$$|\langle c, u_i \rangle| = \frac{|\langle r, v_i \rangle|}{\sqrt{t}\, h_i} = \frac{1}{\sqrt{t}}. \tag{77}$$

These lifted vectors meet the hypotheses of (74), yielding $\vartheta(\overline{G}) \leq t$.

To compare this value with the unsigned certificate, we take coordinatewise absolute values in a Rayleigh quotient. For a real vector x, write $|x|$ for its coordinatewise absolute value; then

$$x^T H^{-1}(I+S)H^{-1}x \leq |x|^T H^{-1}(I+W)H^{-1}|x|. \tag{78}$$

Because x and $|x|$ have equal Euclidean norm, taking Rayleigh maxima gives the second inequality of (11). Let $B := H^{-1}(I+W)H^{-1}$ denote the unsigned comparison matrix. Its symmetry and entrywise nonnegativity identify its largest eigenvalue with its spectral radius. The similar matrix $H^{-2}(I+W)$ has maximum row sum $\max_i(1+D_i)/h_i^2$, which bounds that radius and gives the final inequality in (11).

For the uniform certificate, we replace the individual handle overlaps by their minimum α. The Rayleigh bound for $I+W$ gives, for every unit vector x,

$$x^T Bx \leq (1+\rho(W))\left\|H^{-1}x\right\|^2 \leq \frac{1+\rho(W)}{\alpha^2}. \tag{79}$$

If each edge overlap is at most ε, the adjacency matrix A_G satisfies $0 \leq W \leq \varepsilon A_G$ entrywise. Taking absolute values in the Rayleigh quotient for these symmetric nonnegative matrices gives $\rho(W) \leq \varepsilon\rho(A_G)$. The row-sum estimate $\rho(A_G) \leq \Delta(G)$ completes the chain of bounds in (12). □

Sign invariance and strict improvement. The spectral bound depends on the embedding lines even though it retains signed overlaps. Replacing v_i by $\sigma_i v_i$ with $\sigma_i \in \{\pm 1\}$ conjugates S by the diagonal sign matrix $\Sigma = \operatorname{diag}(\sigma_i)$ and leaves h unchanged. Since H commutes with Σ, this conjugation preserves the certificate. To see a strict improvement over the local bound, take a star with $m \geq 1$ leaves and $v_i = r$ at every vertex. Then $H = I$, $S = W = A_G$, and $\rho(A_G) = \sqrt{m}$, giving the spectral value $1 + \sqrt{m}$ instead of $1 + m$ when $m > 1$. The graph itself has $\chi(G) = \vartheta(\overline{G}) = 2$, so the example exhibits an improvement while also showing the possible gap between a certificate and the theta value.

Weighted local corrections and their optimum. Positive vertex weights can improve the local row-sum bound by redistributing the diagonal correction across an edge. For a weight vector $x \in (0, \infty)^n$, define the weighted overlap sum $D_i(x) := \sum_j W_{ij} x_j / x_i$. To verify that $\operatorname{diag}(D_i(x)) - S$ is positive semidefinite, expand its quadratic form at $y \in \mathbb{R}^n$ as the sum of squares

$$\sum_{\{i,j\} \in E} W_{ij} \left(\sqrt{\frac{x_j}{x_i}} y_i - \operatorname{sgn}(S_{ij}) \sqrt{\frac{x_i}{x_j}} y_j \right)^2, \tag{80}$$

where terms with $W_{ij} = 0$ vanish. Repeating the lift above gives

$$\vartheta(\overline{G}) \leq \max_i \frac{1 + D_i(x)}{h_i^2}. \tag{81}$$

Optimizing these weights recovers the unsigned spectral bound:

$$\inf_{x > 0} \max_i \frac{1 + D_i(x)}{h_i^2} = \lambda_{\max}\left(H^{-1}(I + W)H^{-1} \right). \tag{82}$$

We prove this identity by a row-sum lower bound and a positive resolvent construction, which also handles disconnected graphs. Let $M := H^{-2}(I + W)$ be the matrix similar to the unsigned comparison matrix B. For each $x > 0$, similarity by $\operatorname{diag}(x)$ bounds $\rho(M)$ by $\max_i (Mx)_i / x_i$, giving one inequality. For the reverse direction, a tolerance $\eta > 0$ allows the positive approximate weight vector

$$x_\eta := \left((\rho(M) + \eta) I - M \right)^{-1} \mathbf{1} = \sum_{\ell \geq 0} \frac{M^\ell \mathbf{1}}{(\rho(M) + \eta)^{\ell + 1}} > 0. \tag{83}$$

The resolvent identity gives $M x_\eta = (\rho(M) + \eta) x_\eta - \mathbf{1}$, so the maximum ratio is at most $\rho(M) + \eta$. Letting $\eta \downarrow 0$ and using the similarity of M and B yields (82) without an irreducibility assumption. The signed spectral certificate thus improves or matches every optimized bound obtained from these weighted local corrections.

The spectral lift fixes the correction outside the edge entries. Allowing those remaining entries to vary can improve the certificate further, while keeping the original vectors and handle fixed. The following semidefinite program optimizes over that larger family of lifts.

Proposition 4 (Optimized positive semidefinite correction)**.** *With the notation of Lem. 2, define the optimal correction value $T(v, r)$ by*

$$T(v, r) := \min_{t, X} \left\{ t : \; X \succeq 0, \quad X_{ij} = -\langle v_i, v_j \rangle \;\; (\{i, j\} \in E), \quad 1 + X_{ii} \leq t h_i^2 \;\; (i \in [n]) \right\}. \tag{84}$$

The minimum exists, and $\vartheta(\overline{G}) \leq T(v, r) \leq \tau(S, h)$. Moreover, $T(v, r)$ is the best handle bound among lifts (v_i, w_i) with handle $(r, 0)$, where the number of auxiliary coordinates is unrestricted.

Proof. We identify feasible correction matrices with admissible lifts, then use the spectral construction to obtain a compact feasible sublevel set. Every feasible X is a Gram matrix of auxiliary vectors w_i that cancel the edge overlaps. Normalizing their lifts $z_i = (v_i, w_i)$ gives the handle bound $\max_i (1 + X_{ii}) / h_i^2 \leq t$. Conversely, each admissible lift yields a feasible Gram matrix with this objective value, and a factor in $\mathbb{R}^n$ suffices regardless of its original number of auxiliary coordinates. The correction in (75) is feasible at $t = \tau(S, h)$, establishing nonemptiness and the stated upper bound.

To obtain attainment, we can restrict to $1 \leq t \leq \tau(S, h)$. The diagonal constraints bound every X_{ii}, and positive semidefiniteness then gives $|X_{ij}| \leq \sqrt{X_{ii} X_{jj}}$. This nonempty restricted set is closed and bounded, so the objective attains its minimum. Applying the handle bound to every feasible lift yields the lower bound by $\vartheta(\overline{G})$. □

The optimum $T(v, r)$ is relative to the fixed embedding and handle, so it need not equal $\vartheta(\overline{G})$ and requires an additional semidefinite solve. The explicit spectral certificate instead uses a largest-eigenvalue upper bound. To turn a numerical estimate into a rigorous certificate, one can verify $tH^2 - I - S \succeq 0$ or use the explicit row-sum bound.

H.2 Prototype affinity and decoding

A theta upper bound controls the relaxation but does not generally certify a decoded coloring. For that purpose, we compare each embedding with its assigned orthonormal prototype: two vertices close to the same prototype line must have a large absolute overlap. Small overlaps on edges can therefore rule out monochromatic edges, while the same prototype affinities also provide a handle for the theta bound.

Theorem 12 (Prototype affinity certifies decoding)**.** *Let $v_i \in \mathcal{S}^{d-1}$ be the embeddings and let $q_1, \dots, q_k \in \mathcal{S}^{d-1}$ be orthonormal prototypes. For an arbitrary assignment $a : [n] \to [k]$, write $s := |\mathrm{Im}(a)|$ for the number of used prototypes, $b_i := |\langle v_i, q_{a(i)}\rangle|$ for the assigned affinity at vertex i, and $\beta := \min_i b_i$ for its minimum. Define the handle-overlap bound $\alpha_s(\beta)$ and conflict threshold $\gamma(\beta)$ by*

$$\alpha_s(\beta) := \frac{\beta - \sqrt{s-1}\sqrt{1-\beta^2}}{\sqrt{s}}, \qquad \gamma(\beta) := 2\beta^2 - 1. \tag{85}$$

Let W be the absolute edge-overlap matrix defined above, and let $E_{\mathrm{mono}}(a) := \{\{i, j\} \in E : a(i) = a(j)\}$ denote the set of monochromatic edges.

(i) *If $\beta > \sqrt{(s-1)/s}$, the unit handle $r = s^{-1/2} \sum_{b \in \mathrm{Im}(a)} q_b$ has $|\langle r, v_i\rangle| \geq \alpha_s(\beta) > 0$, and*

$$\vartheta(\overline{G}) \leq \tau(S, h) \leq \frac{1 + \rho(W)}{\alpha_s(\beta)^2}. \tag{86}$$

(ii) *Define the pairwise affinity threshold $c_{ij} := \max\{0, b_i b_j - \sqrt{1-b_i^2}\sqrt{1-b_j^2}\}$. Then*

$$\sum_{\{i,j\} \in E_{\mathrm{mono}}(a)} c_{ij} \leq \sum_{\{i,j\} \in E} |\langle v_i, v_j\rangle|. \tag{87}$$

If $\beta > 1/\sqrt{2}$, this implies

$$\gamma(\beta)|E_{\mathrm{mono}}(a)| \leq \sum_{\{i,j\} \in E} |\langle v_i, v_j\rangle|. \tag{88}$$

If also $|\langle v_i, v_j\rangle| < \gamma(\beta)$ on every edge, then a is proper and

$$\lceil \vartheta(\overline{G}) \rceil \leq \chi(G) \leq s \leq k. \tag{89}$$

The theorem applies to any measured assignment without a colorability or class-size balance assumption. Its decoding thresholds are strict: the sharpness construction below fails at $\beta = 1/\sqrt{2}$ even with orthogonal edge embeddings. Once the assignment is certified proper, $\vartheta(\overline{G}) \leq s$ improves or matches the uniform bound in (86). An independent chromatic lower bound matching s would additionally certify optimality.

Nonuniform prototype affinities. The uniform handle condition can be restrictive when only some assigned classes have weak affinities. We can accommodate this variation with the weighted handle $r_p = \sum_{b \in \mathrm{Im}(a)} p_b q_b$, using positive weights satisfying $\sum_{b \in \mathrm{Im}(a)} p_b^2 = 1$. For each used prototype $b \in \mathrm{Im}(a)$, let $\beta_b = \min_{i:a(i)=b} b_i$ denote its minimum assigned affinity. As proved in Prop. 5, a positive weighted handle has strictly positive worst-case overlap bounds for all assigned prototype neighborhoods if and only if

$$\sum_{b \in \mathrm{Im}(a)} (1 - \beta_b^2) < 1. \tag{90}$$

This condition permits heterogeneous affinities that fail the uniform threshold. Its necessity concerns the stated robust neighborhood bounds; a handle fitted to the observed embeddings may still work outside this condition.

Proof of Thm. 12. We orient each embedding toward its assigned prototype and separate the component along that prototype from its residual. Cauchy–Schwarz then bounds both the handle overlap and the overlap of a same-prototype pair. Let $A := \mathrm{Im}(a)$ be the active prototype set. Choose signs $\sigma_i \in \{\pm 1\}$ so that the oriented embeddings $\widetilde{v}_i := \sigma_i v_i$ satisfy $\langle \widetilde{v}_i, q_{a(i)} \rangle = b_i \geq 0$, and write their orthogonal residuals as y_i:

$$\widetilde{v}_i = b_i q_{a(i)} + y_i, \qquad y_i \perp q_{a(i)}, \qquad \|y_i\| = \sqrt{1 - b_i^2}. \tag{91}$$

The sign choices preserve absolute pairwise overlaps. Because the prototypes are orthonormal, their normalized sum $r = s^{-1/2} \sum_{b \in A} q_b$ is a unit handle whose component orthogonal to $q_{a(i)}$ has norm $\sqrt{(s-1)/s}$. Pairing this component with the residual gives

$$\langle r, \widetilde{v}_i \rangle \geq \frac{b_i - \sqrt{s-1}\sqrt{1-b_i^2}}{\sqrt{s}} \geq \alpha_s(\beta). \tag{92}$$

The function of b_i in the middle expression is increasing on $[0, 1]$, which gives the final inequality. Its value at β is positive if and only if $\beta > \sqrt{(s-1)/s}$. Under that assumption, the handle overlaps satisfy $h_i = |\langle r, v_i \rangle| \geq \alpha_s(\beta) > 0$, so Lem. 2 gives part (i).

For the conflict bound, consider a pair assigned to the same prototype. If $a(i) = a(j)$, the decomposition (91) and Cauchy–Schwarz give

$$\langle \widetilde{v}_i, \widetilde{v}_j \rangle \geq b_i b_j - \sqrt{1 - b_i^2}\sqrt{1 - b_j^2}. \tag{93}$$

The absolute overlap is at least the positive part of this lower bound, so $|\langle v_i, v_j \rangle| \geq c_{ij}$. Summing over monochromatic edges gives (87). If $b_i, b_j \geq \beta > 1/\sqrt{2}$, the lower bound is at least $2\beta^2 - 1 = \gamma(\beta) > 0$, yielding

$$a(i) = a(j) \quad \Longrightarrow \quad |\langle v_i, v_j \rangle| \geq \gamma(\beta) \tag{94}$$

and

$$E_{\mathrm{mono}}(a) \subseteq \{\{i, j\} \in E : |\langle v_i, v_j \rangle| \geq \gamma(\beta)\}. \tag{95}$$

Summing the common lower bound yields (88). When every edge overlap is strictly below $\gamma(\beta)$, the containment excludes all monochromatic edges and certifies the assignment as a proper coloring using s colors. Hence $\chi(G) \leq s \leq k$; the Lovász sandwich inequality and integrality of $\chi(G)$ supply the remaining side of (89). □

Individual affinity thresholds and conflict counts. Retaining the measured affinities b_i instead of their minimum gives a pairwise properness test. The same overlap argument yields

$$|\langle v_i, v_j \rangle| < c_{ij} \quad (\{i, j\} \in E) \quad \Longrightarrow \quad a \text{ is proper.} \tag{96}$$

The test requires positive pairwise margins, since its strict inequality cannot hold when $c_{ij} = 0$, but it can succeed even if the uniform minimum-affinity criterion fails. To count only conflicts with a prescribed margin, fix a threshold $\eta > 0$. Their contribution to the weighted conflict sum gives

$$|\{\{i, j\} \in E_{\mathrm{mono}}(a) : c_{ij} \geq \eta\}| \leq \frac{1}{\eta} \sum_{\{i,j\} \in E} |\langle v_i, v_j \rangle| . \tag{97}$$

If every edge overlap is at most ε, the right-hand side of (88) is also bounded by $\varepsilon|E|$. These versions allow either measured edge overlaps or a uniform overlap bound to certify the representation, once the required affinities have been observed.

Distinct-prototype separation. Prototype affinities also bound how similar vertices assigned to different prototypes can be. Let $\theta_i := \arccos b_i \in [0, \pi/2]$ be the angle from the embedding line to its assigned prototype line. The projective angular distance $d_{\mathrm{P}}(x, y) := \arccos |\langle x, y \rangle|$ obeys the triangle inequality: we choose signs attaining the two constituent distances, apply the spherical triangle inequality, and multiply those signs to obtain an admissible sign for the third distance. For distinct assignments, the prototype lines are orthogonal, so this triangle inequality gives

$$d_{\mathrm{P}}(v_i, v_j) \geq \frac{\pi}{2} - \theta_i - \theta_j. \tag{98}$$

Whenever $\theta_i + \theta_j \le \pi/2$, the angular bound implies $|\langle v_i, v_j\rangle| \le \sin(\theta_i + \theta_j)$. In particular, $\beta > 1/\sqrt{2}$ ensures $\theta_i + \theta_j \le 2\arccos\beta < \pi/2$, yielding the uniform separation bound

$$a(i) \neq a(j) \quad \Longrightarrow \quad |\langle v_i, v_j\rangle| \le 2\beta\sqrt{1-\beta^2}. \tag{99}$$

To prove the heterogeneous-affinity criterion stated above, we let each active prototype contribute its own weight to the handle. The residual decomposition then gives a vertexwise overlap bound, and positivity of all these bounds reduces to a single condition on the class affinities.

Proposition 5 (Weighted prototype handles)**.** *Let $A := \mathrm{Im}(a)$ be the active prototype set. For positive handle weights p_b, $b \in A$, satisfying $\sum_{b\in A} p_b^2 = 1$, define the weighted handle r_p and its vertexwise overlap bound $\ell_i(p)$ by*

$$r_p := \sum_{b\in A} p_b q_b, \qquad \ell_i(p) := p_{a(i)} b_i - \sqrt{1-p_{a(i)}^2}\sqrt{1-b_i^2}. \tag{100}$$

If every $\ell_i(p) > 0$, the diagonal lower-bound matrix $L := \mathrm{diag}(\ell_i(p))$ gives

$$\vartheta(\overline{G}) \le \lambda_{\max}\big(L^{-1}(I+S)L^{-1}\big)\,. \tag{101}$$

Writing $\beta_b := \min_{i:a(i)=b} b_i$ for the minimum affinity in assigned class b, positive weights meeting these strict inequalities exist if and only if

$$\sum_{b\in A}(1-\beta_b^2) < 1. \tag{102}$$

Proof. We use the lower overlap bounds to build a feasible spectral correction, then solve the inequalities that make those bounds positive. The residual decomposition (91) gives $\langle r_p, \widetilde{v}_i\rangle \ge \ell_i(p)$, so the measured handle overlaps satisfy $h_i := |\langle r_p, v_i\rangle| \ge \ell_i(p) > 0$. Define the candidate certificate value $t := \lambda_{\max}(L^{-1}(I+S)L^{-1}) > 0$. Its defining inequality is $tL^2 - I - S \succeq 0$; adding the nonnegative diagonal correction $t\,\mathrm{diag}(h_i^2 - \ell_i(p)^2)$ gives $tH^2 - I - S \succeq 0$. The spectral lift is therefore feasible at t and yields (101).

For existence of the weights, consider $p, b \in [0,1]$. The inequality $pb > \sqrt{1-p^2}\sqrt{1-b^2}$ is equivalent to $p^2 > 1 - b^2$, since either strict inequality forces $p, b > 0$ and squaring preserves equivalence. Positivity of every $\ell_i(p)$ therefore requires and suffices for $p_b^2 > 1 - \beta_b^2$ at each active prototype. Summing these inequalities gives necessity of (102). Conversely, if their right-hand sides sum to $Q < 1$, the choice $p_b^2 := 1 - \beta_b^2 + (1-Q)/s$ distributes the remaining mass equally. These positive weights sum in square to one and satisfy every strict inequality, establishing sufficiency. □

Uniform weights recover the balanced handle, while (102) uses the weakest affinity in each active class without restricting class sizes. For example, two active prototypes with minimum affinities $\beta_1 = 3/5$ and $\beta_2 = 1$ meet the weighted condition even though the balanced threshold $\beta > 1/\sqrt{2}$ fails. The improvement supplies a theta handle; proper decoding still requires the separate edge-overlap test.

When that test allows a small number of conflicts, we can instead repair the assignment. A vertex cover of the conflict graph identifies vertices whose recoloring removes every monochromatic edge, and the conflict bound controls the size of a simple such cover.

Corollary 4 (Repair from the conflict certificate)**.** *Assume $\beta > 1/\sqrt{2}$, and write $R := \sum_{\{i,j\}\in E} |\langle v_i, v_j\rangle|$ for the total edge overlap. Let C be a vertex cover of the conflict graph $([n], E_{\mathrm{mono}}(a))$. Keeping a on $[n]\setminus C$ and properly coloring $G[C]$ using fresh colors gives*

$$\chi(G) \le s + \chi(G[C]). \tag{103}$$

In particular, there is a repair using at most $s + \lfloor R/\gamma(\beta)\rfloor$ colors.

Proof. We check the repaired coloring separately on the cover, its complement, and the edges between them. Every edge inside $[n]\setminus C$ retains distinct original labels because C meets each conflict. A proper coloring of $G[C]$ handles edges within the cover, while its fresh palette separates all edges to the complement. For the quantitative bound, choose one endpoint of each conflict and let C be their union. Then $|C| \le |E_{\mathrm{mono}}(a)| \le \lfloor R/\gamma(\beta)\rfloor$, so assigning distinct fresh colors to these vertices suffices; an empty cover needs no new colors. □

The extra colors are part of the guarantee: a conflicting assignment need not admit a proper coloring with its original s labels. In fact, no additive term sublinear in the conflict count can work uniformly over s. In K_{s+m} with $m \le s$, assign m classes of size two and $s-m$ singleton classes. This produces m conflicting edges, whereas the chromatic number is $s+m$. Taking s orthonormal prototypes and $v_i = q_{a(i)}$ meets all prototype hypotheses with $\beta = 1$, $\gamma(\beta) = 1$, and $R = m$, attaining the repair bound.

The measured overlap R links these decoding certificates to supervised training. The excess-loss estimate in Thm. 5 bounds it directly, so the same quantity can control theta, conflicts, and the additional palette required for repair.

Corollary 5 (Achieved loss, certificates, and repair)**.** *Assume the hypotheses of Thm. 5, and identify its unit embeddings with v_i. For the InfoNCE variant, define the loss-derived overlap bound $Q := nC_{\mathrm{R}}\epsilon_{\mathrm{R}}/2$. For the aggregate log objective, additionally assume that every supplied color class has size at least two and set $Q := nC_{\log}(J_G - b_G^{\log})/2$. In either case,*

$$R := \sum_{\{i,j\}\in E} |\langle v_i, v_j\rangle| \le Q. \tag{104}$$

For a unit handle r with positive minimum overlap $\alpha := \min_i |\langle r, v_i\rangle| > 0$,

$$\vartheta(\overline{G}) \le \frac{1+\min\{Q, \sqrt{2Q}\}}{\alpha^2}. \tag{105}$$

For any decoded assignment with orthonormal prototypes and minimum assigned affinity $\beta > 1/\sqrt{2}$,

$$|E_{\mathrm{mono}}(a)| \le \left\lfloor \frac{Q}{\gamma(\beta)} \right\rfloor. \tag{106}$$

Thus $Q < \gamma(\beta)$ certifies properness; in general, a repair uses at most $s + \lfloor Q/\gamma(\beta)\rfloor$ colors.

Proof. We extract the edge-overlap bound from the geometric defect and pass it through the spectral and decoding estimates. The edge part of $\sum_i T_i$ counts every overlap twice; for the log objective, the class-size hypothesis makes all vertices active and ensures the same counting. Dropping the nonnegative dispersion term in Thm. 5 yields (104). To control theta, the row-sum estimate gives $\rho(W) \le R$, while the Frobenius norm gives the complementary bound

$$\rho(W) \le \|W\|_{\mathrm{F}} = \sqrt{2\sum_{\{i,j\}\in E} |\langle v_i, v_j\rangle|^2} \le \sqrt{2R}, \tag{107}$$

because unit-vector overlaps have absolute value at most one. Substituting the better of these two estimates into Lem. 2 gives (105). For the decoded assignment, (88) and integrality bound the number of conflicts; Cor. 4 converts that count into the stated palette bound. □

The supplied proper labels define the loss reference, and the decoded assignment a is the object being certified. Applying the corollary therefore requires an achieved loss gap and measured prototype affinities. Whether optimization reaches those values, or whether training controls them on a test distribution, remains separate from this deterministic implication.

H.3 Signed Euclidean distance and sharpness

For comparison with Euclidean clustering errors, we can express the affinity thresholds as distances between lines. Unit vectors satisfy $\min_{\sigma\in\{\pm1\}} \|v - \sigma q\|^2 = 2 - 2|\langle v, q\rangle|$, so the maximum signed Euclidean line distance

$$\delta := \max_i \min_{\sigma\in\{\pm1\}} \left\|v_i - \sigma q_{a(i)}\right\| \tag{108}$$

satisfies $\delta^2 = 2 - 2\beta$ and $0 \le \delta \le \sqrt{2}$. Substituting this relation into the handle and conflict thresholds gives

$$\alpha_s(\beta) = \frac{1-\delta^2/2}{\sqrt{s}} - \sqrt{1-\frac{1}{s}}\,\delta\sqrt{1-\frac{\delta^2}{4}}, \tag{109}$$

$$\gamma(\beta) = 1 - 2\delta^2 + \frac{\delta^4}{2}. \tag{110}$$

Thus the uniform sufficient proper-decoding conditions are $\delta < \sqrt{2-\sqrt{2}}$ and $\varepsilon < 1 - 2\delta^2 + \delta^4/2$, where ε bounds every edge overlap.

Sharpness, including nearest-prototype decoding. A single edge shows why equality must be excluded from both decoding thresholds. Let e_1, e_2, e_3 be the standard basis of $\mathbb{R}^3$, choose orthonormal prototypes $q_1 = e_1$ and $q_2 = e_2$, and take a graph with one edge $\{u, v\}$. For an angle $\rho \in [0, \pi/4]$, define embeddings symmetrically about the first prototype and assign both endpoints to it:

$$v_u = \cos\rho\, e_1 + \sin\rho\, e_3, \qquad v_v = \cos\rho\, e_1 - \sin\rho\, e_3, \qquad a(u) = a(v) = 1. \tag{111}$$

Both assigned affinities are $\beta = \cos\rho$, while the edge overlap is $\cos(2\rho) = 2\beta^2 - 1 = \gamma(\beta)$. The monochromatic edge therefore violates properness as soon as equality is allowed in the overlap threshold. Its line distance is $\delta = 2\sin(\rho/2)$; at $\rho = \pi/4$, this reaches $\sqrt{2-\sqrt{2}}$ even though the edge overlap vanishes. Each endpoint still has q_1 as its unique nearest prototype line, because its affinity with q_2 is zero. Equality at the affinity threshold can thus fail under nearest-prototype decoding as well. Adding an isolated vertex embedded at and assigned to q_2 makes both prototypes active without changing either failure.

I CORA ABLATIONS

We study the main design choices on Cora before transferring the resulting pipeline to the later citation-graph, COLOR, and cycle experiments. Holding the sampled-graph protocol fixed lets us compare feature and loss choices first, then examine how the encoder and decoder affect the palette–conflict tradeoff.

Reporting conventions. The neural results are aggregate statistics over the stated seeds. Individual predictions and checkpoints are unavailable, so the aggregates cannot be independently reconstructed from those outputs. Some configurations have differing summaries across separate experiment campaigns; we identify these discrepancies in the relevant comparisons.

Our reported mean $\pm$ standard deviation describes variation across the stated seeds, rather than uncertainty over new graph families. Because model choices used the reported evaluation scores and an untouched selection/test split is not documented, we interpret the selected rows as exploratory comparisons. A threshold hit means $\mathrm{Mono} \leq .05$ and can still contain monochromatic edges, while a missing output does not establish that the coloring problem is infeasible. These conventions apply throughout the citation, COLOR, and cycle studies below.

CORA graph construction and statistics. Each training example is one induced 2-hop Cora subgraph, with pseudo-color labels produced by NetworkX greedy coloring with DSATUR. Tables 5 and 6 give the sampling rule and the resulting graph sizes. On average, a training subgraph has 75.42 nodes and 127.86 undirected edges; a test subgraph has 71.80 nodes and 117.08 undirected edges. The accepted test subgraphs range from 50 to 107 nodes, so this ablation evaluates local Cora neighborhoods rather than the full citation graph.

Table 5: CORA full-graph and sampling specification.

Item	Value
Full graph	2708 nodes, 5278 undirected edges, density 0.00144
Node features	1433-dimensional Cora bag-of-words features
Center split	Train/test ratio $0.5/0.5$, split seed 0
Center pools	1354 train-center nodes, 1354 test-center nodes
Subgraph sampling	Full-graph induced 2-hop neighborhoods, relabeled locally
Accepted size range	50–120 nodes
Sampling seeds	Train seed 123, test seed 999
Label proxy	DSATUR greedy coloring on each sampled subgraph
Evaluation	Held-out test graphs

Table 6: Single-subgraph statistics for the sampled CORA train/test sets. Values are means with min–max ranges.

Split	Graphs	Nodes	Undirected edges	Density	Greedy χ proxy
Train	200	75.42 (50–119)	127.86 (65–278)	0.0460	4.27 (3–5)
Test	50	71.80 (50–107)	117.08 (61–255)	0.0461	4.32 (3–5)

Experimental protocol. Starting from the full Cora citation graph, candidate center nodes are split into disjoint train and test pools with ratio $0.5/0.5$. From each center pool, we sample induced 2-hop subgraphs with size range $(50, 120)$, yielding 200 training graphs and 50 test graphs. Because these neighborhoods are induced from the full graph, non-center nodes can appear in both training and test subgraphs even though the center-node pools are disjoint. The recorded protocol specifies this sampling construction for the Cora variants, but the available summaries do not identify the exact evaluated subset.

I.1 Node features and loss function ablation with vanilla GCN

Objective definitions. We write L_{InfoNCE} for the implemented term (2), using signed similarity in v1 and absolute similarity in v1abs. This term pools all non-self same-label positives before the logarithm and omits singleton-color anchors, so it differs from the per-pair objective (1). The code adds a 10^{-12} stabilizer to the positive and total sums and excludes same-label edges from the negative set; the latter agrees with $\mathcal{N}_G(v)$ under proper supervision. Within a minibatch, it averages over eligible anchors across graphs, weighting each graph by its eligible-anchor count. The theoretical per-graph statements apply before this averaging and use ideal arithmetic on unit-norm outputs.

The following auxiliary terms define the loss variants. Reproducing a run also requires its regularization and warmup settings, which these formulas alone do not specify.

We first compare a small family of feature and loss variants under the same vanilla GCN encoder. For a graph $G = (V, E)$, let x_u denote the input features, z_u the learned embedding of node u, and $\ell_u \in \mathbb{R}^C$ its logits for C auxiliary color slots. At softmax temperature $\tau_{\text{soft}} > 0$, the soft color probabilities are

$$p_u = \text{softmax}(\ell_u/\tau_{\text{soft}}).$$

v1. The v1 objective is contrastive only:

$$L_{\text{v1}} = L_{\text{InfoNCE}}.$$

v1abs. This replaces signed similarity by absolute similarity in the implemented aggregated logarithmic objective.

v2. The v2 objective augments the same contrastive term with a soft conflict loss:

$$L_{\text{v2}} = L_{\text{InfoNCE}} + \lambda_{\text{soft}} L_{\text{soft}}.$$

To weight conflicts by their endpoint degrees, we use a degree exponent p and edge weights w_{uv}, normalized within each graph to weights w'_{uv} of mean one. The soft conflict loss is

$$L_{\text{soft}} = \frac{1}{|E|} \sum_{(u,v)\in E} w'_{uv} \sum_{c=1}^{C} p_{u,c} p_{v,c},$$

$$w'_{uv} = \frac{w_{uv}}{|E|^{-1} \sum_{\{a,b\}\in E} w_{ab}}, \qquad w_{uv} = \frac{(\deg u + 1)^p + (\deg v + 1)^p}{2}.$$

For $E = \varnothing$, set $L_{\text{soft}} = 0$. Since $p_{u,c}$ is the probability assigned to auxiliary color slot c, the sum $\sum_{c=1}^{C} p_{u,c} p_{v,c}$ measures the soft overlap between adjacent nodes' assignments. The implementation also exposes an optional margin penalty and warmup; the displayed v2 expression assumes that the margin coefficient is zero. A learned readout head, with weight matrix W_{col} and bias b_{col}, produces the logits from the node embedding:

$$z_u = f_\theta(x_u, G), \qquad \ell_u = W_{\text{col}} z_u + b_{\text{col}}, \qquad p_u = \text{softmax}(\ell_u/\tau_{\text{soft}}).$$

v2abs. This replaces the signed logarithmic contrastive term of v2 by its absolute-similarity counterpart.

v3. The v3 objective is again contrastive only, but with structural node features instead of Cora BOW:

$$L_{\text{v3}} = L_{\text{InfoNCE}}.$$

v4. The v4 objective is contrastive only, but with i.i.d. random Gaussian node features:

$$L_{\text{v4}} = L_{\text{InfoNCE}}.$$

v5. The v5 objective uses a weighted soft conflict loss in place of (2). Let A be the adjacency matrix, D the diagonal degree matrix, and P the matrix of row-wise softmax probabilities. Then

$$L_{\text{weighted-soft}} = \tfrac{1}{2}\langle D^p A, PP^\top \rangle_F,$$

and we set

$$L_{\text{v5}} = L_{\text{weighted-soft}}.$$

Table 7: Comparison of CORA variants on the common CORA split-subgraph test set. All methods use the same sampled train/test subgraphs; they differ in node features, training objective, and inference head. Metrics are averaged over the evaluated test graphs.

Variant	Train Features	Training Objective	Inference	k mean	Mono mean	k/χ mean	Hit-rate
v1	Raw CORA BOW	L_{InfoNCE}	Embedding → k-medoids	6.20	0.0327	1.42	1.00
v1abs	Raw CORA BOW	L_{InfoNCE}	Embedding → k-medoids	6.40	0.0251	1.44	1.00
v2	Raw CORA BOW	$L_{\text{InfoNCE}} + \lambda_{\text{soft}} L_{\text{soft}}$	Embedding → k-medoids	6.33	0.0299	1.46	1.00
v3	Structural only	L_{InfoNCE}	Embedding → k-medoids	9.87	0.0545	2.28	0.63
v4	Random Gaussian	L_{InfoNCE}	Embedding → k-medoids	12.77	0.0376	2.94	0.97
v5	Raw CORA BOW	$L_{\text{weighted-soft}}$	Embedding → k-medoids	8.97	0.0365	2.03	1.00
PI-GNN	Learnable node embeddings	Potts-model soft conflict objective	Direct color logits → arg max	4.33	0.0012	1.00	0.83

Reading the comparison. In Table 7, *BOW* denotes the raw Cora bag-of-words node features, k is the tested palette size selected by the decoder, χ is the greedy-coloring proxy, and *Mono* is the monochromatic-edge fraction. Tied embeddings can leave fewer occupied clusters than the requested palette, so k describes the decoder's request rather than an independently counted number of occupied colors.

Among the contrastive CORA variants in this untuned comparison, v1 uses the smallest palette relative to the greedy proxy, while v2 slightly reduces conflicts at the cost of a larger k/χ. Structural-only features (v3), Gaussian random features (v4), and the soft-loss-only variant (v5) all underperform v1/v2 on the CORA split. PI-GNN (Schuetz et al., 2022), evaluated on the same sampled test graphs, has lower reported mean color count and conflict rate, but its recorded hit rate is also lower (0.83). The available per-instance records do not resolve whether completion or threshold failures account for that hit value. We next tune the auxiliary term to examine the tradeoff observed between v1 and v2.

I.2 v2 Hyperparameter tuning experiments

Tuning rationale. We tune v2 in stages to examine the contribution of its auxiliary soft conflict term. Trial 1 first tunes the loss weight λ_{soft} to determine how strongly the soft conflict term should

remain active. Trial 2 then tunes the two key parameters inside L_{soft}: the degree exponent p, which controls how strongly dense-region conflicts are upweighted, and the softmax temperature τ_{soft}, which controls how sharp the soft color probabilities are. This staging keeps the search focused on the auxiliary term while retaining the rest of the pipeline.

Table 8: Summary of v2 hyperparameter tuning. Metrics are averaged over the evaluated test graphs.

Stage	Search space	Best setting	k mean	Mono	k/χ	Hit-rate
Baseline	Default v2	$\lambda_{\text{soft}} = 0.20$ $p = 3$, $\tau_{\text{soft}} = 1.0$	6.33	0.0299	1.46	1.00
Trial 1	Tune λ_{soft}	$\lambda_{\text{soft}} = 0.30$	6.17	0.0300	1.41	1.00
Trial 2	Tune p and τ_{soft}	$p = 4.0$, $\tau_{\text{soft}} = 1.25$ $\lambda_{\text{soft}} = 0.30$	5.80	0.0291	1.32	1.00

Tuning results. Table 8 follows the roles of the three parameters defined above: λ_{soft} sets the auxiliary weight, p controls degree weighting, and τ_{soft} controls the softness of the color probabilities. In Trial 1, moving from the default setting to $\lambda_{\text{soft}} = 0.30$ improves k/χ from 1.46 to 1.41 with no loss in hit-rate. Trial 2 then searches within L_{soft}: changing $p = 3$ to $p = 4$ together with $\tau_{\text{soft}} = 1.25$ reduces mean k from 6.17 to 5.80 and k/χ from 1.41 to 1.32, with a slight improvement in Mono. These are observations from the tuning sweep, without an independent confirmatory evaluation; the joint change in Trial 2 does not isolate the effect of either hyperparameter.

I.3 Finding the right encoders

I.3.1 Encoder implementations

To compare architectures with common input features, width, and depth, all encoders in the initial sweep use raw CORA BOW features, two encoder layers, hidden dimension 128, and dropout (Srivastava et al., 2014) 0.1 where applicable.

GCN variants. The `gcn` baseline uses PyG (Fey & Lenssen, 2019) `GCNConv` without self-loops, with ReLU between layers and no residual connections or normalization. The custom `res_gcn` adds an input projection to the hidden dimension and combines repeated `GCNConv`, ReLU, and dropout with residual addition (He et al., 2016) and LayerNorm. The `gin` encoder (Xu et al., 2019) instead uses PyG `GINConv` with a two-layer MLP and trainable ϵ, followed by residual addition, LayerNorm, ReLU, and dropout.

Other local encoders. The custom `gated_gcn` computes vector gates from source and target node states, degree-normalizes the gated messages, and applies a residual LayerNorm update without edge features. For comparison, `sage` uses PyG `SAGEConv` with ReLU and dropout between layers, while `gat` (Veličković et al., 2018) uses PyG `GATConv` with 4 attention heads, ELU (Clevert et al., 2016), and dropout between layers. Neither of these latter two encoders uses a residual connection.

GPS variants. The `gps` encoder combines a local `GCNConv` module with global multi-head self-attention (Vaswani et al., 2017) in PyG `GPSConv`, using 4 attention heads. The `gps_gcn`, `gps_sage`, `gps_gin`, and `gps_gat` variants retain that global block and use `GCNConv`, `SAGEConv`, `GINConv`, and `GATConv`, respectively, for local message passing.

Other attention-based encoders. The custom `graph_vit` projects node features to the hidden dimension and processes them with a two-layer `TransformerEncoder`, without explicit graph positional encoding. The custom `exphormer`, inspired by Shirzad et al. (2023), instead uses sparse attention over original graph edges, self-edges, and deterministic cyclic-offset edges at offsets $\pm 1, \ldots, \pm 4$.

Orthogonal message transformations. The real-valued `unitary_mp`, inspired by Kiani et al. (2024), uses Cayley-transform orthogonal maps for self and neighbor messages, degree-normalized aggregation, a learned self/message mixing coefficient, residual addition, and LayerNorm. We describe this construction, GatedGCN, and residual GCN in more detail because these choices motivate the targeted comparisons below.

GatedGCN-style encoder. The gate controls which coordinates of a source node enter each neighbor message. For node state h_u, each directed edge (u, v) computes a vector gate g_{uv} and gated message m_{uv} using trainable matrices W_s, W_t, W_m and the sigmoid σ:

$$g_{uv} = \sigma(W_s h_u + W_t h_v), \qquad m_{uv} = W_m h_u \odot g_{uv}.$$

Messages into node v are degree-normalized and combined with a self transformation whose trainable matrix is W_0:

$$\tilde{h}_v = W_0 h_v + \frac{1}{\max(1, \deg v)} \sum_{u:(u,v)\in E} m_{uv}.$$

The layer output is

$$h_v^{\text{out}} = \text{LayerNorm}\left(h_v + \text{Dropout}(\text{ReLU}(\tilde{h}_v))\right).$$

This design uses vector-valued gates, residual connections, and LayerNorm, but does not use edge features.

Residual GCN encoder. The residual GCN uses the same local aggregation family as the plain GCN baseline, but wraps each layer with residual normalization:

$$\tilde{h}^{(\ell+1)} = \text{ReLU}(\text{GCNConv}(h^{(\ell)}, E)),$$

$$h^{(\ell+1)} = \text{LayerNorm}\left(h^{(\ell)} + \text{Dropout}(\tilde{h}^{(\ell+1)})\right).$$

Comparing this encoder with the plain GCN tests the role of residual connections and normalization without introducing attention or global mixing.

Unitary message-passing encoder. The unitary_mp encoder is a real-valued orthogonal/unitary-style message-passing model. It is not a complex-valued unitary neural network; instead, each layer constrains its self and message transformations to be orthogonal through a Cayley transform. For a trainable square matrix A, the skew-symmetric part is

$$S = A - A^\top,$$

and the corresponding orthogonal map is

$$Q(A) = (I - S)(I + S)^{-1}.$$

Each layer uses separate orthogonal maps for the self state and neighbor messages:

$$q_v = Q_{\text{self}} h_v, \qquad r_u = Q_{\text{msg}} h_u.$$

Neighbor messages are mean-aggregated,

$$\bar{r}_v = \frac{1}{\max(1, \deg v)} \sum_{u:(u,v)\in E} r_u,$$

and then mixed with a learned gate $\beta \in [0, 1]$:

$$\tilde{h}_v = \beta q_v + (1 - \beta)\bar{r}_v.$$

The layer output uses the same residual-normalized form as the other stable MP encoders:

$$h_v^{\text{out}} = \text{LayerNorm}\left(h_v + \text{Dropout}(\text{ReLU}(\tilde{h}_v))\right).$$

This encoder tests whether norm-preserving message transformations stabilize the embeddings before k-medoids decoding.

I.3.2 All Encoder Results

Table 9 compares these encoders under the common CORA protocol. The v1 and v2 rows show whether an architectural change remains useful when the auxiliary conflict term is added.

Table 9: All CORA encoder results, grouped by message-passing (MP) and non-MP/global-attention encoders. Entries are mean ± standard deviation over ten random seeds, except parameter counts, which are deterministic.

Encoder	Pipeline	k mean	Mono mean	k/χ mean	Hit-rate	Params
Message-passing encoders						
gcn	v1	5.88±0.54	0.0290±0.0032	1.40±0.14	0.97±0.04	200064
	v2	5.99±0.48	0.0281±0.0032	1.43±0.13	0.97±0.03	200709
res_gcn	v1	4.66±0.37	0.0209±0.0032	1.11±0.10	1.00±0.00	217088
	v2	4.73±0.24	0.0221±0.0032	1.13±0.07	1.00±0.00	217733
gat	v1	7.15±0.81	0.0364±0.0112	1.70±0.20	0.91±0.08	200576
	v2	7.61±1.48	0.0436±0.0224	1.80±0.34	0.87±0.11	201221
gin	v1	4.99±0.39	0.0220±0.0030	1.19±0.10	1.00±0.00	250114
	v2	5.01±0.42	0.0229±0.0035	1.19±0.11	1.00±0.00	250759
gated_gcn	v1	4.26±0.19	0.0207±0.0028	1.01±0.05	1.00±0.00	316160
	v2	4.38±0.19	0.0197±0.0039	1.04±0.05	1.00±0.00	316805
sage	v1	4.29±0.25	0.0223±0.0027	1.02±0.06	1.00±0.00	399872
	v2	4.48±0.28	0.0239±0.0031	1.06±0.08	1.00±0.00	400517
unitary_mp	v1	4.48±0.28	0.0233±0.0034	1.06±0.08	1.00±0.00	249602
	v2	4.43±0.29	0.0233±0.0039	1.05±0.08	1.00±0.00	250247
Non-MP / global-attention encoders						
gps	v1	5.03±0.35	0.0228±0.0039	1.20±0.09	1.00±0.00	482048
	v2	5.06±0.35	0.0248±0.0031	1.21±0.09	1.00±0.00	482693
graph_vit	v1	12.73±2.28	0.0444±0.0062	3.04±0.54	0.82±0.16	580352
	v2	12.63±1.36	0.0412±0.0030	3.00±0.33	0.88±0.05	580997
exphormer	v1	15.40±1.14	0.0480±0.0043	3.65±0.27	0.70±0.08	316160
	v2	15.94±0.89	0.0515±0.0040	3.78±0.19	0.59±0.10	316805

Across ten seeds, GatedGCN is the strongest encoder in the initial sweep by both k and k/χ (v1: $k/\chi = 1.01 \pm 0.05$), while GatedGCN v2 gives the lowest average Mono (0.0197 ± 0.0039). Residual GCN, GraphSAGE, and UnitaryMP remain stable with hit-rate 1.00. Among attention-based encoders, untuned GPS remains the strongest, while GraphViT and Exphormer underperform in this setup.

I.3.3 Tuned GPS and Other Attention-Based Results

We tune the attention-based variants in a separate sweep. In particular, we tried different local message-passing blocks inside GPS and tuned the other attention-based encoders. The sweep used the same CORA split, losses, and k-medoids decoder. Table 10 reports the best setting found for each of these encoder families.

Table 10: Best configurations from the non-MP/global-attention tuning sweep. Entries are mean ± standard deviation over ten random seeds for the fixed reported configurations.

Encoder	Pipeline	Layers	Heads	Drop Out	LR	Exp.	k mean	Mono	k/χ	Hit-rate
gps_gcn	v1	3	8	0.2	0.003	–	4.53±0.19	0.0196±0.0043	1.08±0.06	1.00±0.00
	v2	2	4	0.2	0.003	–	4.42±0.25	0.0199±0.0029	1.05±0.07	1.00±0.00
gps_sage	v1	2	4	0.2	0.003	–	4.45±0.17	0.0201±0.0037	1.06±0.06	1.00±0.00
	v2	3	4	0.2	0.01	–	4.44±0.19	0.0219±0.0037	1.06±0.05	1.00±0.00
gps_gin	v1	2	8	0.2	0.003	–	4.67±0.31	0.0187±0.0041	1.11±0.08	1.00±0.00
	v2	2	4	0.2	0.01	–	4.59±0.26	0.0207±0.0036	1.09±0.07	1.00±0.00
gps_gat	v1	2	8	0.2	0.003	–	4.44±0.22	0.0189±0.0039	1.05±0.05	1.00±0.00
	v2	2	8	0.1	0.003	–	4.50±0.23	0.0185±0.0035	1.06±0.06	1.00±0.00
graph_vit	v1	2	8	0.1	0.003	–	6.28±0.30	0.0229±0.0029	1.51±0.08	1.00±0.00
	v2	2	8	0.2	0.003	–	6.17±0.51	0.0224±0.0037	1.48±0.14	1.00±0.00
exphormer	v1	2	4	0.2	0.003	2	8.10±3.75	0.0323±0.0090	1.92±0.87	0.96±0.08
	v2	2	4	0.2	0.003	2	13.08±3.36	0.0435±0.0065	3.09±0.80	0.81±0.16

For the fixed tuned configurations, GPS variants are competitive with the strongest message-passing encoders: v2 GPS-GCN reports $k/\chi = 1.05 \pm 0.07$, and the GPS variants exhibit different palette/conflict tradeoffs. GraphViT improves relative to the untuned sweep but remains behind the GPS variants, and Exphormer is still weaker in this CORA setting.

I.3.4 Targeted GatedGCN/ResGCN Ablation

The initial sweep makes GatedGCN and residual GCN natural candidates for a closer comparison. We therefore change their architectural components while keeping the v1/v2 losses, train/test split, and k-medoids decoder fixed.

Reference and gate controls. The `gated_gcn_full` reference uses vector gates, degree-normalized aggregation, residual addition, and LayerNorm. The `gated_gcn_no_gate` variant removes the learned source-target gate while keeping the self-linear term and the remaining update; `gated_gcn_scalar_gate` replaces the vector gate by one scalar per edge.

Normalization and aggregation controls. The `gated_gcn_no_layernorm` and `gated_gcn_no_residual` variants remove LayerNorm and the residual skip connection, respectively, while retaining the vector gate, degree-normalized aggregation, and other components. The `gated_gcn_sum_agg` variant changes only the aggregation, replacing the degree-normalized mean by an unnormalized sum.

Capacity control. The `res_gcn` baseline combines `GCNConv`, dropout, residual addition, and LayerNorm. Adding an FFN block gives `res_gcn_param_matched`, which tests whether additional capacity explains the difference from GatedGCN.

Table 11: Targeted ablation results for the strongest message-passing encoder components. Entries are mean ± standard deviation over ten random seeds.

Encoder	Pipeline	k mean	Mono	k/χ	Hit-rate
gated_gcn_full	v1	4.26±0.20	0.0206±0.0037	1.01±0.06	1.00±0.00
	v2	4.38±0.28	0.0206±0.0034	1.04±0.07	1.00±0.00
gated_gcn_no_gate	v1	4.31±0.25	0.0211±0.0037	1.02±0.05	1.00±0.00
	v2	4.39±0.16	0.0193±0.0017	1.04±0.04	1.00±0.00
gated_gcn_no_layernorm	v1	4.24±0.19	0.0205±0.0029	1.01±0.05	1.00±0.00
	v2	4.37±0.29	0.0189±0.0049	1.04±0.07	1.00±0.00
gated_gcn_no_residual	v1	4.27±0.41	0.0219±0.0034	1.01±0.09	1.00±0.00
	v2	4.29±0.27	0.0216±0.0029	1.02±0.07	1.00±0.00
gated_gcn_scalar_gate	v1	4.31±0.28	0.0214±0.0042	1.02±0.07	1.00±0.00
	v2	4.45±0.24	0.0200±0.0022	1.06±0.06	1.00±0.00
gated_gcn_sum_agg	v1	4.45±0.25	0.0196±0.0031	1.06±0.07	1.00±0.00
	v2	4.56±0.43	0.0192±0.0036	1.08±0.11	1.00±0.00
res_gcn	v1	4.69±0.29	0.0226±0.0035	1.11±0.08	1.00±0.00
	v2	4.80±0.42	0.0222±0.0032	1.15±0.11	1.00±0.00
res_gcn_param_matched	v1	4.75±0.35	0.0241±0.0023	1.13±0.09	1.00±0.00
	v2	5.95±1.59	0.0277±0.0063	1.41±0.37	0.99±0.01

The v1 no-LayerNorm variant has the lowest mean k/χ (1.01 ± 0.05), very close to the full model (1.01 ± 0.06), and the v2 no-LayerNorm variant has the lowest Mono (0.0189 ± 0.0049). The full GatedGCN design is therefore stable, but the evidence does not isolate any single gate, residual, or normalization choice as uniformly dominant. The parameter-matched residual GCN is worse than the simpler residual GCN, which does not establish a capacity-only explanation of the difference.

I.3.5 Absolute-Value InfoNCE with SignNet Post-Processing

To test an alternative to hard line canonicalization, we ran one CORA split-subgraph ablation with a SignNet-style sign-invariant post-head after training with the absolute-similarity version of (2). The experiment uses the same CORA split-subgraph construction as the preceding ablations and evaluates on held-out test graphs with the same k-medoids decoding rule with mono threshold 0.05. For the canonical baseline, we sweep the canonicalization pass count over $\{1, 2, 4\}$ and report the best row per encoder. For the SignNet variant, we use the same top three encoder configurations selected from the CORA encoder study: `gated_gcn`, `gps_gcn`, and `gps_sage`.

With MLP blocks ϕ and ρ, the SignNet post-head combines the two signs before its final transformation:

$$\mathrm{SignNet}(z) = \rho\big(\phi(z) + \phi(-z)\big).$$

This replaces the hard line-canonicalization step by a continuous, piecewise differentiable transformation that is sign-invariant at evaluation with dropout disabled, applied after the encoder and before the final embedding-based decoding stage.

Table 12: CORA absolute-value InfoNCE: best canonicalization pass count versus the SignNet post-head. Entries are mean ± standard deviation over ten random seeds.

Encoder	Canon passes	Canon k	Canon Mono	Canon k/χ	SignNet k	SignNet Mono	SignNet k/χ
GatedGCN	4	4.64±0.15	0.0198±0.0032	1.10±0.05	4.98±0.29	0.0155±0.0043	1.19±0.07
GPS-GCN	1	5.97±0.38	0.0201±0.0032	1.43±0.11	5.17±0.41	0.0156±0.0065	1.23±0.11
GPS-SAGE	2	6.14±0.46	0.0222±0.0048	1.47±0.11	5.34±0.52	0.0161±0.0052	1.27±0.12

Across ten seeds, the SignNet post-head has lower reported mean Mono than hard canonicalization for each encoder, but it does not lower k/χ for GatedGCN. It does improve both GPS encoders on k/χ and Mono, with equal reported absolute ratio reductions of 0.20 for the two GPS variants.

I.4 SINGLE-RUN ABSOLUTE-VALUE SWEEP WITH LINE CANONICALIZATION

Implementation. The following sweep reports one run per configuration. It is a separate experiment from the ten-seed SignNet comparison, so their numerical summaries do not support a matched comparison. This experiment replaces signed inner products with absolute inner products in the v1 training term. Let $\hat{z}_u = z_u/\|z_u\|_2$ be the normalized embedding, $P(u)$ the non-self same-greedy-color positives in the same sampled graph, and $N(u)$ the adjacent different-color negatives. Writing $\mathcal{V}_{\text{valid}}$ for anchors with at least one positive, the objective is the implemented form of (2):

$$L_{\text{abs-InfoNCE}} = -\frac{1}{|\mathcal{V}_{\text{valid}}|} \sum_{u \in \mathcal{V}_{\text{valid}}} \log \frac{\sum_{v \in P(u)} \exp(|\langle \hat{z}_u, \hat{z}_v \rangle|/\tau)}{\sum_{v \in P(u) \cup N(u)} \exp(|\langle \hat{z}_u, \hat{z}_v \rangle|/\tau)},$$

Before k-medoids evaluation, each normalized embedding is canonicalized as an unoriented line: choose the coordinate with largest magnitude and flip the whole vector if that coordinate is negative. We ablate the number of canonicalization evaluation passes over $\{1, 2, 4\}$ and report the best pass for each encoder. This is a v1-style experiment for all encoders: it uses only (2), without the v2 soft-conflict auxiliary objective. The GPS and other attention-based rows use the v1-tuned hyperparameter settings from Table 10; the message-passing rows use their established default settings.

Table 13: Absolute-value InfoNCE with line canonicalization on the CORA split-subgraph test set, using the best available encoder hyperparameters. Metrics are averaged over the evaluated test graphs.

Encoder setting	Passes	k mean	Mono	k/χ	Hit-rate	Params
gcn	4	6.87	0.0354	1.58	1.00	200064
Normalized GCN	1	7.80	0.0287	1.78	1.00	200064
res_gcn	4	5.10	0.0174	1.19	1.00	217088
gat	4	9.03	0.0496	2.08	0.80	200576
gin	1	5.73	0.0222	1.31	1.00	250114
gated_gcn	4	4.27	0.0241	1.00	1.00	316160
sage	1	6.47	0.0308	1.49	1.00	399872
unitary_mp	2	4.27	0.0237	0.99	1.00	249602
gps_gcn	4	5.63	0.0218	1.31	1.00	631296
gps_sage	2	6.90	0.0261	1.59	1.00	482048
gps_gin	4	6.17	0.0214	1.43	1.00	515074
gps_gat	1	5.77	0.0230	1.34	1.00	482560
graph_vit	2	6.17	0.0293	1.43	1.00	580352
exphormer	1	6.40	0.0235	1.49	1.00	316160

The single-run sweep reports its smallest ratio for UnitaryMP, but the unmatched records do not support ranking it against the separate ten-seed tuned table. The pass-count comparison also needs care: a deterministic canonicalization map is idempotent, so a reported gain cannot be attributed to repeated canonicalization without determining whether other evaluation settings or random seeds also changed.

J EXPERIMENTS ON CITATION GRAPHS

The Cora ablations identify configurations to examine on other citation graphs. We first repeat the sampled-subgraph comparison on CiteSeer and PubMed, then test full-graph fitting and transfer from a frozen Cora-trained encoder. The reporting conventions and limits of the available records in Sec. I apply throughout.

Citation graph scale. The full-graph sizes in Table 14, taken from the PI-GNN paper of Schuetz et al. (2022), provide context for the sampled-subgraph experiments. We retain the Cora edge count measured in our own PyG preprocessing separately in Table 5, so that the reference graph specifications are not confused with the graph used by our sampling code.

Table 14: Reference full-graph specifications for the three citation graphs. These are full-graph scales; our experiments use sampled local subgraphs unless stated otherwise.

Dataset	Nodes	Edges	Density	Reported colors q
Cora	2708	5429	0.15%	5
CiteSeer	3327	4732	0.09%	6
PubMed	19717	44338	0.02%	8

CiteSeer and PubMed graph construction. The CiteSeer and PubMed sampled-subgraph runs reuse the same split-subgraph protocol as Cora so that the later cross-dataset comparisons stay controlled: 200 training subgraphs, 50 test subgraphs, induced 2-hop neighborhoods, accepted size range 50–120, DSATUR greedy-color pseudo-labels. Table 15 reports the resulting sampled-subgraph scale for these two datasets.

Table 15: Single-subgraph statistics for sampled CiteSeer and PubMed train/test sets. Values are means with min–max ranges.

Dataset	Split	Nodes	Undirected edges	Density	Greedy χ proxy
CiteSeer	Train	76.81 (50–119)	152.78 (63–333)	0.0550	4.40 (3–6)
CiteSeer	Test	74.38 (51–119)	138.88 (63–304)	0.0503	4.28 (3–6)
PubMed	Train	76.06 (50–119)	133.59 (51–393)	0.0462	4.16 (3–8)
PubMed	Test	78.94 (50–120)	141.38 (64–414)	0.0462	4.28 (3–7)

J.1 CiteSeer and PubMed Experiments

We repeat the encoder sweep on CiteSeer and PubMed using the same v1/v2 objectives, split-subgraph construction, and k-medoids decoder. The GPS and other attention-based rows use pipeline-specific Cora-tuned hyperparameters from Table 10, while the message-passing rows use their established default settings. Each encoder name in the tables therefore denotes its corresponding v1 or v2 tuned configuration.

Table 16: CiteSeer encoder sweep with selected available encoder hyperparameters. Entries are mean ± standard deviation over ten random seeds.

Encoder setting	Pipeline	k mean	Mono	k/χ	Hit-rate
gcn	v1	5.73±0.50	0.0245±0.0043	1.33±0.12	0.99±0.02
	v2	5.79±0.58	0.0265±0.0025	1.35±0.15	0.99±0.02
res_gcn	v1	4.68±0.23	0.0206±0.0033	1.07±0.06	1.00±0.00
	v2	4.64±0.28	0.0222±0.0024	1.07±0.07	1.00±0.00
gat	v1	6.62±1.00	0.0288±0.0047	1.51±0.23	0.97±0.05
	v2	7.01±1.30	0.0357±0.0159	1.61±0.27	0.93±0.13
gin	v1	4.67±0.41	0.0232±0.0036	1.07±0.11	1.00±0.00
	v2	4.77±0.34	0.0227±0.0038	1.10±0.07	1.00±0.00
gated_gcn	v1	4.43±0.27	0.0190±0.0029	1.01±0.07	1.00±0.00
	v2	4.49±0.30	0.0188±0.0038	1.03±0.07	1.00±0.00
sage	v1	4.33±0.30	0.0208±0.0040	0.99±0.08	1.00±0.00
	v2	4.42±0.16	0.0213±0.0048	1.01±0.06	1.00±0.00
unitary_mp	v1	4.58±0.31	0.0236±0.0030	1.05±0.09	1.00±0.01
	v2	4.47±0.33	0.0226±0.0029	1.03±0.09	1.00±0.00
gps_gcn	v1	4.62±0.21	0.0188±0.0038	1.06±0.05	1.00±0.00
	v2	4.58±0.12	0.0206±0.0029	1.05±0.04	1.00±0.00
gps_sage	v1	4.61±0.25	0.0217±0.0031	1.06±0.06	1.00±0.00
	v2	4.75±0.33	0.0200±0.0044	1.09±0.08	1.00±0.00
gps_gin	v1	4.66±0.33	0.0198±0.0030	1.07±0.08	1.00±0.00
	v2	4.63±0.29	0.0185±0.0034	1.06±0.08	1.00±0.00
gps_gat	v1	4.58±0.16	0.0184±0.0042	1.05±0.05	1.00±0.00
	v2	4.59±0.31	0.0191±0.0056	1.05±0.08	1.00±0.00
graph_vit	v1	5.49±0.40	0.0229±0.0037	1.28±0.10	0.99±0.02
	v2	5.43±0.39	0.0213±0.0047	1.26±0.09	0.99±0.02
exphormer	v1	6.56±3.11	0.0283±0.0080	1.48±0.66	1.00±0.00
	v2	8.32±4.58	0.0313±0.0102	1.88±0.99	0.95±0.10

Table 17: PubMed encoder sweep with selected available encoder hyperparameters. Entries are mean ± standard deviation over ten random seeds.

Encoder setting	Pipeline	k mean	Mono	k/χ	Hit-rate
gcn	v1	7.75±0.72	0.0338±0.0033	1.82±0.13	0.98±0.03
	v2	7.64±0.74	0.0333±0.0026	1.80±0.15	0.98±0.02
res_gcn	v1	5.10±0.37	0.0292±0.0022	1.24±0.07	1.00±0.00
	v2	5.13±0.34	0.0297±0.0026	1.26±0.06	1.00±0.00
gat	v1	9.30±0.64	0.0352±0.0031	2.19±0.13	0.93±0.05
	v2	9.31±0.72	0.0371±0.0036	2.18±0.13	0.91±0.05
gin	v1	6.79±0.89	0.0292±0.0035	1.65±0.20	0.99±0.03
	v2	7.06±0.75	0.0307±0.0038	1.71±0.16	0.98±0.04
gated_gcn	v1	4.56±0.27	0.0297±0.0022	1.12±0.05	1.00±0.00
	v2	4.43±0.26	0.0295±0.0034	1.09±0.04	1.00±0.00
sage	v1	5.16±0.26	0.0317±0.0035	1.27±0.06	1.00±0.00
	v2	5.15±0.37	0.0307±0.0020	1.26±0.05	1.00±0.00
unitary_mp	v1	5.07±0.34	0.0308±0.0038	1.24±0.06	1.00±0.00
	v2	5.04±0.31	0.0299±0.0027	1.23±0.04	1.00±0.00
gps_gcn	v1	4.67±0.36	0.0288±0.0031	1.15±0.08	1.00±0.00
	v2	4.76±0.28	0.0291±0.0020	1.17±0.05	1.00±0.00
gps_sage	v1	4.99±0.27	0.0311±0.0028	1.23±0.05	1.00±0.00
	v2	4.61±0.27	0.0303±0.0026	1.13±0.06	1.00±0.00
gps_gin	v1	5.72±1.62	0.0296±0.0023	1.39±0.37	1.00±0.00
	v2	9.49±1.65	0.0365±0.0031	2.21±0.35	0.90±0.07
gps_gat	v1	4.77±0.43	0.0308±0.0022	1.18±0.09	1.00±0.00
	v2	4.97±0.51	0.0287±0.0027	1.22±0.10	1.00±0.00
graph_vit	v1	15.62±1.22	0.1276±0.0562	3.95±0.40	0.19±0.17
	v2	15.72±0.75	0.1268±0.0427	3.97±0.24	0.20±0.12
exphormer	v1	6.51±0.60	0.0325±0.0024	1.62±0.14	1.00±0.00
	v2	6.55±0.48	0.0329±0.0025	1.63±0.10	1.00±0.00

The ranking changes with the dataset: v1 SAGE has the best mean color efficiency on CiteSeer ($k/\chi = 0.99 \pm 0.08$), with GatedGCN and GIN close behind, whereas v2 GatedGCN leads the reported PubMed encoders ($k/\chi = 1.09 \pm 0.04$). Several GPS and residual/message-passing rows remain competitive on PubMed, but GraphViT is unstable there and Exphormer is substantially weaker on CiteSeer. These local-neighborhood results motivate a separate test of whether the same configurations work on full graphs.

J.2 Full-Graph to Full-Graph Extension

Implementation. The previous subsection trains on sampled local citation subgraphs and tests on disjoint-center subgraphs from the same base graph. We now extend that methodology to transductive fitting on the full Cora, CiteSeer, and PubMed graphs. The v1/v2 pipelines retain the signed variant of (2), raw BOW features, DSATUR pseudo-color labels, and k-medoids decoding over the learned embeddings. The selected five encoders are res_gcn, gated_gcn, unitary_mp, gps_gcn, and gps_sage. For v2, we keep the previously tuned auxiliary setting $\lambda_{\text{soft}} = 0.30$, $p = 4.0$, and $\tau_{\text{soft}} = 1.25$.

Full-graph protocol. For a full citation graph $G = (V, E)$, we train on that same graph and evaluate on that same graph. The greedy full-graph coloring is used only as a pseudo-label source for the contrastive loss and as a reference χ proxy. To evaluate (2) on PubMed, we process anchor rows in chunks while retaining each anchor's comparisons against the full node set. At evaluation, we sweep $k = 1, \ldots, 16$. Write k_{zero} for the smallest successful zero-conflict palette and k_τ for the smallest palette meeting conflict tolerance τ; in this local notation, τ is a tolerance rather than the loss temperature. When the corresponding tested set is nonempty, these quantities are

$$k_{\text{zero}} = \min\{k : \text{conflicts}(k) = 0\},$$
$$k_\tau = \min\{k : \text{Mono}(k) \le \tau\}, \qquad \tau \in \{0.005, 0.01, 0.02, 0.05\}.$$

We also report the conflict count and normalized conflict rate at $k = \chi_{\text{greedy}}$, with unavailable threshold summaries marked in the tables. This separates the stricter tolerances ($k_{0.005}$ and $k_{0.01}$) from the looser low-conflict settings ($k_{0.02}$ and $k_{0.05}$).

Table 18: Full-graph specifications used in the transductive contrastive benchmark. The edge counts are the undirected counts observed in the local PyG preprocessing used by the experiment.

Dataset	Nodes	Features	Undirected edges	Density	Greedy χ proxy
Cora	2708	1433	5278	0.001440	5
CiteSeer	3327	3703	4552	0.000823	6
PubMed	19717	500	44324	0.000228	8

Table 19: Full-graph-to-full-graph contrastive results. k_{zero} is the smallest zero-conflict k found within the search cap; $k_{0.05}$ is the smallest k with monochromatic-edge fraction at most 0.05. Entries are mean $\pm$ standard deviation over ten random seeds; NA denotes an unavailable threshold summary; per-seed failure counts are unavailable.

Dataset / Encoder	Pipeline	k_{zero}	$k_{0.05}$	$k_{0.05}/\chi$	Conflicts@χ_{greedy}	ϵ@χ_{greedy}
Cora full graph						
gps_sage	v1	13.5±2.4	4.4±0.5	0.88±0.10	63.5±24.7	0.012031±0.004688
gps_gcn	v1	NA	4.1±0.3	0.82±0.06	48.1±14.7	0.009113±0.002785
gated_gcn	v1	NA	4.0±0.0	0.80±0.00	71.8±9.8	0.013604±0.001847
res_gcn	v1	NA	4.3±0.5	0.86±0.10	144.2±23.0	0.027321±0.004350
unitary_mp	v1	NA	5.7±0.5	1.14±0.10	276.8±26.4	0.052444±0.005010
CiteSeer full graph						
gps_gcn	v1	NA	4.0±0.0	0.67±0.00	12.1±8.4	0.002658±0.001836
gps_sage	v1	NA	4.0±0.0	0.67±0.00	24.7±16.0	0.005426±0.003521
gated_gcn	v1	NA	4.0±0.0	0.67±0.00	31.0±14.3	0.006810±0.003133
res_gcn	v1	NA	4.0±0.0	0.67±0.00	74.8±10.7	0.016432±0.002341
unitary_mp	v1	NA	5.0±0.0	0.83±0.00	118.5±15.7	0.026033±0.003451
PubMed full graph						
gated_gcn	v1	NA	6.0±0.5	0.75±0.06	1282.3±191.1	0.028930±0.004312
gps_sage	v1	NA	5.3±0.7	0.66±0.08	1152.0±147.7	0.025990±0.003331
gps_gcn	v1	NA	5.4±0.7	0.68±0.09	1207.5±277.2	0.027243±0.006253
res_gcn	v1	NA	6.4±0.5	0.80±0.06	1731.1±275.1	0.039056±0.006207
unitary_mp	v1	NA	7.7±0.5	0.96±0.06	1926.3±39.5	0.043460±0.000892

Table 20: Full-graph comparison at multiple conflict thresholds. Entries are mean ± standard deviation over ten random seeds; NA denotes an unavailable threshold summary; per-seed failure counts are unavailable.

Dataset / Encoder	Pipeline	k_{zero}	$k_{0.005}$	$k_{0.01}$	$k_{0.02}$	$k_{0.05}$
Cora full graph						
gps_sage	v1	13.5±2.4	6.5±1.0	5.9±0.6	5.0±0.0	4.4±0.5
gps_gcn	v1	NA	6.2±0.6	5.4±0.5	5.0±0.0	4.1±0.3
gated_gcn	v1	NA	7.5±0.7	6.3±0.5	5.0±0.0	4.0±0.0
res_gcn	v1	NA	10.0±1.1	8.2±0.9	6.3±0.5	4.3±0.5
unitary_mp	v1	NA	14.0±1.2	11.3±1.2	8.4±0.5	5.7±0.5
CiteSeer full graph						
gps_gcn	v1	NA	6.1±0.3	5.5±0.5	5.1±0.3	4.0±0.0
gps_sage	v1	NA	6.3±0.7	5.5±0.7	5.0±0.0	4.0±0.0
gated_gcn	v1	NA	7.4±1.0	5.9±0.9	5.0±0.0	4.0±0.0
res_gcn	v1	NA	10.5±1.1	7.8±0.8	5.8±0.6	4.0±0.0
unitary_mp	v1	NA	12.6±1.4	9.8±0.9	7.3±0.5	5.0±0.0
PubMed full graph						
gated_gcn	v1	NA	NA	NA	11.4±2.0	6.0±0.5
gps_sage	v1	NA	NA	NA	9.9±1.0	5.3±0.7
gps_gcn	v1	NA	NA	NA	11.7±1.9	5.4±0.7
res_gcn	v1	NA	NA	NA	NA	6.4±0.5
unitary_mp	v1	NA	NA	NA	13.9±0.3	7.7±0.5

The tables show that the full-graph setting is substantially harder than the sampled-subgraph setting, but that the learned embeddings still produce stable low-conflict operating points. Exact zero-conflict recovery is rare under the tested cap: among the rows in Tab. 20, only Cora with GPS-SAGE consistently reaches zero conflicts across all ten seeds, with $k_{\text{zero}} = 13.5 \pm 2.4$. The stricter thresholds separate the encoders more clearly. On Cora, the GPS rows reach $k_{0.005} = 6.5 \pm 1.0$ for GPS-SAGE and 6.2 ± 0.6 for GPS-GCN, while GatedGCN needs 7.5 ± 0.7; all three then collapse to the greedy proxy scale by $k_{0.02}$, with $k_{0.02} = 5.0 \pm 0.0$. CiteSeer is similar at moderate thresholds: GPS-GCN, GPS-SAGE, and GatedGCN all reach $k_{0.05} = 4.0 \pm 0.0$, and their $k_{0.02}$ values are 5.1 ± 0.3, 5.0 ± 0.0, and 5.0 ± 0.0, respectively.

PubMed is the main stress case. None of the PubMed rows consistently reaches the 0.005 or 0.01 thresholds within the tested range, and the 0.02 threshold requires noticeably more colors: $k_{0.02} = 9.9 \pm 1.0$ for GPS-SAGE, 11.4 ± 2.0 for GatedGCN, 11.7 ± 1.9 for GPS-GCN, and 13.9 ± 0.3 for UnitaryMP, while ResGCN does not consistently reach the threshold. At the looser 0.05 operating point, however, the decoder still finds relatively small palettes with nonzero tolerated conflicts: Cora reaches $k_{0.05} = 4.0 \pm 0.0$ with GatedGCN, CiteSeer reaches 4.0 ± 0.0 for all three strongest GPS/GatedGCN rows, and PubMed reaches 5.3 ± 0.7 with GPS-SAGE. Thus the full-graph results describe decoder operating points, without measuring the proposed embedding geometry, but they also show that the best encoder depends on the graph and on how strict the conflict tolerance is: GPS variants are strongest at the stricter thresholds, GatedGCN is competitive at the loose Cora operating point, and PubMed requires larger palettes despite the same training pipeline.

J.3 CORA TRANSFER WITH SHARED STRUCTURAL FEATURES

Implementation. Raw BOW features cannot be transferred directly across Cora, CiteSeer, and PubMed because the feature dimensions and vocabularies are dataset-specific. To test scale and dataset transfer while preserving the contrastive pipeline, we therefore replace raw BOW by a shared structural feature map with fixed dimension 16. The model is trained on Cora and evaluated without retraining on CiteSeer and PubMed. We test the same five high-performing encoders used in the full-graph tuning study: res_gcn, gated_gcn, unitary_mp, gps_gcn, and gps_sage. The subgraph and full-graph transfer settings are run separately because their inference distributions are different: subgraph transfer tests local-neighborhood generalization, while full-graph transfer tests whether the learned representation scales to the whole target graph.

Table 21: Cora-trained transfer on held-out target subgraphs. Entries are mean ± standard deviation over ten random seeds.

Train → Test	Encoder / Method	Pipeline	k mean	Mono mean	k/χ mean	Hit-rate
Cora → CiteSeer	gated_gcn	v1	6.17±0.54	0.0490±0.0035	1.37±0.09	0.74±0.10
Cora → CiteSeer	gated_gcn	v2	6.21±0.74	0.0487±0.0039	1.40±0.15	0.74±0.11
Cora → CiteSeer	gps_gcn	v1	6.91±0.67	0.0489±0.0034	1.56±0.15	0.74±0.11
Cora → PubMed	gated_gcn	v2	5.08±0.42	0.0309±0.0037	1.20±0.06	0.99±0.02
Cora → PubMed	gps_sage	v2	6.01±0.73	0.0341±0.0028	1.40±0.10	0.99±0.02
Cora → PubMed	gated_gcn	v1	5.02±0.53	0.0301±0.0027	1.19±0.08	0.99±0.02

Table 22: Cora-trained transfer to full target graphs. Entries are mean ± standard deviation over ten random seeds; NA denotes an unavailable threshold summary; per-seed failure counts are unavailable.

Train → Test	Encoder	Pipeline	$k_{0.05}$	Mono@0.05	Conflicts@χ_{greedy}	Mono@χ_{greedy}
Cora → CiteSeer	gps_gcn	v1	NA	–	646.2±36.7	0.1420±0.0081
Cora → CiteSeer	res_gcn	v1	NA	–	776.8±30.3	0.1707±0.0067
Cora → CiteSeer	gps_gcn	v2	NA	–	790.7±58.4	0.1737±0.0128
Cora → PubMed	gps_gcn	v1	9.8±2.4	0.0448±0.0033	2834.3±797.6	0.0639±0.0180
Cora → PubMed	gps_gcn	v2	11.7±2.2	0.0455±0.0039	3143.1±442.6	0.0709±0.0100
Cora → PubMed	res_gcn	v1	13.4±2.5	0.0453±0.0040	3118.6±440.9	0.0704±0.0099

On held-out target subgraphs, Cora-trained structural features retain useful performance on PubMed and partially transfer to CiteSeer, but remain weaker than same-dataset raw-BOW training. The full-graph results provide less support for direct transfer: an all-seed threshold value is unavailable for the displayed CiteSeer rows, while the PubMed GPS-GCN v1 row has $k_{0.05} = 9.8 \pm 2.4$. This gap from same-dataset fitting motivates feature adaptations that carry BOW information into a common input space.

J.3.1 BOW Adaptation Experiments

Implementation. For a node u in dataset d, let F_d be the vocabulary-dependent feature dimension and $x_u^{(d)} \in \mathbb{R}^{F_d}$ its raw BOW vector. Since F_{Cora}, F_{CiteSeer}, and F_{PubMed} are different, we cannot reuse the raw input projection across datasets directly. Each adaptation therefore constructs a shared feature vector $\tilde{x}_u \in \mathbb{R}^r$ of fixed dimension r before applying the same v1/v2 contrastive encoder.

First, all four adaptations use a deterministic signed hash projection. Let $H_d \in \mathbb{R}^{F_d \times m}$ be a deterministic sparse projection with one signed nonzero per original BOW coordinate, divided by the square root of the corresponding bucket count as in the implementation. The shared hashed BOW vector is

$$h_u = x_u^{(d)} H_d \in \mathbb{R}^m.$$

Write $\widehat{h}_u$ for the row-normalized sketch, with the implementation's numerical epsilon floor. This signed hash map places all citation graphs into the same input dimension without learning a dataset-specific projection. Hash collisions can change sparsity and inner products, and no distortion bound is established for this deterministic map. In the hybrid variants, let $s_u \in \mathbb{R}^{32}$ denote the structural feature vector built from degree, clustering coefficient, core number, PageRank (Brin & Page, 1998), and deterministic transforms. Then

$$\tilde{x}_u^{\text{hyb}} = [\widehat{h}_u; s_u] \in \mathbb{R}^{288}, \qquad \tilde{x}_u^{\text{hyb+LN}} = \text{LN}([\widehat{h}_u; s_u]),$$

where LN denotes layer normalization applied after concatenation. Because the projection hashes coordinate indices rather than aligned vocabulary tokens, its common output dimension does not establish semantic alignment across datasets.

The BOW-only variants compress the source hashed matrix $H_{\text{src}} \in \mathbb{R}^{n \times 1024}$, with one row for each of the n source nodes, to dimension 256. For PCA, let μ denote the source mean and let the columns

of V_{256} be its top principal directions. The projected feature is

$$\tilde{x}_u^{\mathrm{PCA}} = (h_u - \mu)V_{256}.$$

For SVD, if the source hashed matrix admits the rank-256 approximation $H_{\mathrm{src}} \approx U_{256}\Sigma_{256}V_{256}^{\top}$, then the transferred target feature is

$$\tilde{x}_u^{\mathrm{SVD}} = h_u V_{256}.$$

The PCA/SVD vectors in these formulas are row-normalized before entering the encoder. For Cora $\rightarrow$ CiteSeer/PubMed transfer, the PCA/SVD basis is fit on Cora and reused unchanged on the target graph. All four adaptations keep the downstream v1/v2 contrastive objectives and k-medoids decoder unchanged. The experiment uses the same top-3 encoder family as the later transfer study: gated_gcn, gps_gcn, and gps_sage. The tables below report only the actual transfer settings, not the same-dataset control runs.

Table 23: Cora subgraph transfer with the v1 pipeline. Entries are mean $\pm$ standard deviation over ten random seeds.

Adaptation	Encoder	Cora $\rightarrow$ CiteSeer		Cora $\rightarrow$ PubMed	
		k/χ	Mono	k/χ	Mono
Hybrid	gated_gcn	1.30±0.08	0.0342±0.0040	1.29±0.08	0.0316±0.0038
	gps_gcn	1.29±0.18	0.0330±0.0047	1.24±0.12	0.0302±0.0023
	gps_sage	1.47±0.16	0.0369±0.0030	1.49±0.12	0.0330±0.0035
Hybrid+LN	gated_gcn	1.27±0.08	0.0340±0.0032	1.30±0.07	0.0315±0.0022
	gps_gcn	1.44±0.18	0.0350±0.0032	1.35±0.14	0.0310±0.0021
	gps_sage	1.41±0.11	0.0346±0.0025	1.41±0.05	0.0320±0.0028
PCA	gated_gcn	1.54±0.18	0.0373±0.0018	1.99±0.20	0.0352±0.0025
	gps_gcn	1.30±0.11	0.0344±0.0024	1.58±0.20	0.0354±0.0017
	gps_sage	1.70±0.23	0.0369±0.0042	1.94±0.26	0.0338±0.0021
SVD	gated_gcn	1.45±0.10	0.0376±0.0025	1.41±0.11	0.0339±0.0023
	gps_gcn	1.32±0.09	0.0356±0.0032	1.32±0.12	0.0314±0.0021
	gps_sage	1.74±0.17	0.0376±0.0030	1.63±0.16	0.0350±0.0035

Table 24: Cora subgraph transfer with the v2 pipeline. Entries are mean $\pm$ standard deviation over ten random seeds.

Adaptation	Encoder	Cora $\rightarrow$ CiteSeer		Cora $\rightarrow$ PubMed	
		k/χ	Mono	k/χ	Mono
Hybrid	gated_gcn	1.27±0.10	0.0340±0.0030	1.26±0.12	0.0310±0.0037
	gps_gcn	1.37±0.13	0.0343±0.0022	1.36±0.11	0.0318±0.0026
	gps_sage	1.31±0.09	0.0340±0.0034	1.23±0.11	0.0321±0.0026
Hybrid+LN	gated_gcn	1.29±0.10	0.0357±0.0016	1.29±0.14	0.0302±0.0032
	gps_gcn	1.34±0.10	0.0344±0.0028	1.37±0.15	0.0310±0.0026
	gps_sage	1.35±0.17	0.0339±0.0030	1.23±0.11	0.0307±0.0019
PCA	gated_gcn	1.64±0.18	0.0363±0.0030	2.25±0.47	0.0362±0.0025
	gps_gcn	1.57±0.15	0.0367±0.0028	1.76±0.25	0.0356±0.0032
	gps_sage	1.35±0.10	0.0358±0.0039	1.58±0.17	0.0342±0.0019
SVD	gated_gcn	1.48±0.10	0.0370±0.0042	1.42±0.12	0.0336±0.0023
	gps_gcn	1.53±0.18	0.0370±0.0032	1.45±0.21	0.0317±0.0012
	gps_sage	1.27±0.11	0.0352±0.0048	1.32±0.17	0.0334±0.0028

Table 25: Cora full-graph transfer with the v1 pipeline. Entries are mean ± standard deviation over ten random seeds; NA denotes an unavailable threshold summary; per-seed failure counts are unavailable.

Threshold aggregation caveat: The aggregation routine included a -1 failure sentinel in palette means. NR replaces summaries for which sentinel contamination cannot be excluded from the available moments; NA denotes unavailable summaries. Monochromatic-edge averages describe returned operating points and do not supply missing success counts.

Adaptation	Encoder	Cora → CiteSeer			Cora → PubMed		
		$k_{0.05}$	Mono	$k_{0.02}$	$k_{0.05}$	Mono	$k_{0.02}$
Hybrid	gated_gcn	6.7±0.7	0.0453±0.0034	NR	10.8±1.3	0.0458±0.0023	NA
	gps_gcn	6.2±0.4	0.0412±0.0029	12.1±1.3	7.5±1.1	0.0458±0.0032	NR
	gps_sage	6.7±0.5	0.0453±0.0025	13.9±1.5	8.4±0.5	0.0462±0.0029	NR
Hybrid+LN	gated_gcn	6.9±1.0	0.0448±0.0054	NR	11.3±2.1	0.0476±0.0030	NA
	gps_gcn	6.2±0.4	0.0448±0.0045	13.1±1.1	7.7±1.1	0.0457±0.0039	NR
	gps_sage	7.1±0.3	0.0449±0.0030	14.9±1.5	9.0±0.7	0.0463±0.0033	NR
PCA	gated_gcn	5.9±0.3	0.0436±0.0040	11.9±0.7	10.0±2.2	0.0467±0.0024	NA
	gps_gcn	6.0±0.7	0.0444±0.0036	12.3±1.1	9.3±2.5	0.0454±0.0023	NA
	gps_sage	6.3±0.5	0.0438±0.0039	13.4±0.5	10.6±2.4	0.0448±0.0028	NR
SVD	gated_gcn	5.8±0.4	0.0436±0.0034	11.2±0.6	8.2±0.4	0.0466±0.0023	NR
	gps_gcn	6.2±0.4	0.0437±0.0038	12.4±1.7	7.2±0.8	0.0457±0.0040	NR
	gps_sage	6.4±0.7	0.0453±0.0029	12.9±1.0	8.8±0.4	0.0471±0.0015	NR

Table 26: Cora full-graph transfer with the v2 pipeline. Entries are mean ± standard deviation over ten random seeds; NA denotes an unavailable threshold summary; per-seed failure counts are unavailable.

Threshold aggregation caveat: The aggregation routine included a -1 failure sentinel in palette means. NR replaces summaries for which sentinel contamination cannot be excluded from the available moments; NA denotes unavailable summaries. Monochromatic-edge averages describe returned operating points and do not supply missing success counts.

Adaptation	Encoder	Cora → CiteSeer			Cora → PubMed		
		$k_{0.05}$	Mono	$k_{0.02}$	$k_{0.05}$	Mono	$k_{0.02}$
Hybrid	gated_gcn	9.2±2.3	0.0464±0.0026	NR	NR	0.0449±0.0034	NA
	gps_gcn	8.6±1.3	0.0457±0.0028	NR	9.5±0.8	0.0458±0.0031	NA
	gps_sage	NR	0.0432±0.0024	NR	9.3±1.9	0.0457±0.0032	NR
Hybrid+LN	gated_gcn	NR	0.0438±0.0025	NA	NR	0.0464±0.0038	NA
	gps_gcn	8.5±1.6	0.0475±0.0023	NR	9.8±0.8	0.0457±0.0041	NA
	gps_sage	6.9±2.2	0.0470±0.0023	NR	NR	0.0458±0.0020	NR
PCA	gated_gcn	6.6±0.8	0.0440±0.0053	14.0±1.8	NR	0.0470±0.0030	NA
	gps_gcn	8.0±1.1	0.0452±0.0033	NR	12.2±2.8	0.0441±0.0038	NA
	gps_sage	NR	0.0449±0.0054	NR	NR	0.0436±0.0038	NR
SVD	gated_gcn	7.1±1.0	0.0408±0.0035	NR	10.1±1.2	0.0476±0.0017	NA
	gps_gcn	8.3±0.5	0.0457±0.0036	NR	8.8±1.0	0.0441±0.0037	NR
	gps_sage	NR	0.0448±0.0040	NR	NR	0.0456±0.0031	NR

Among the reported subgraph-transfer rows, v2 Hybrid/GPS-SAGE has the best mean k/χ on CiteSeer, while v2 Hybrid/GatedGCN and Hybrid/GPS-SAGE give the best PubMed results by mean k/χ and Mono, respectively. Full-graph transfer is more variable: NA and NR mark missing or uninterpretable threshold summaries, and the available aggregates do not reliably determine partial-success counts. We therefore exclude unavailable or sentinel-contaminated summaries from comparisons and treat the remaining selected rows as exploratory, since no independent confirmatory selection split is available.

J.4 Absolute-value transfer experiments and baseline methods

Implementation. To compare with a feature choice that does not use the citation vocabularies, we evaluate the absolute-similarity form of (2) with 64-dimensional random node features. As in the preceding transfer experiments, the model is trained on Cora and evaluated without retraining on CiteSeer and PubMed. We compare these learned runs against PI-GNN and full-GCN, imposing the same 15-minute per-graph cap for full-graph transfer.

Table 27: Cora-trained absolute-value transfer with random features on held-out target subgraphs, together with the two baselines. Rows are mean ± standard deviation over ten random seeds.

Train → Test	Encoder / Method	Pipeline	Completed	k mean	Mono mean	k/χ mean	Hit-rate
Cora → CiteSeer	gated_gcn	v2abs	30/30	5.88±0.38	0.0332±0.0033	1.33±0.09	1.00±0.00
Cora → CiteSeer	gated_gcn	v1abs	30/30	5.95±0.29	0.0341±0.0031	1.34±0.06	1.00±0.00
Cora → CiteSeer	PI-GNN	–	30/30	3.10	0.0283	0.72	–
Cora → CiteSeer	full-GCN	–	19/30	3.16	0.0192	0.75	–
Cora → PubMed	gated_gcn	v2abs	30/30	5.08±0.35	0.0306±0.0017	1.24±0.07	1.00±0.00
Cora → PubMed	gated_gcn	v1abs	30/30	5.22±0.35	0.0306±0.0026	1.26±0.09	1.00±0.00
Cora → PubMed	PI-GNN	–	30/30	2.93	0.0209	0.77	–
Cora → PubMed	full-GCN	–	17/30	2.71	0.0228	0.74	–

Results. Table 27 reports the held-out target-subgraph results. The absolute-value variants remain low-conflict on both targets across seeds, but the learned rows are less color-efficient than the external baselines on the subgraphs those baselines complete. This comparison remains runtime-sensitive: full-GCN does not complete all target subgraphs, whereas the learned absolute-value runs complete the full evaluation set.

Table 28 reports the corresponding full-graph transfer results. Here the contrast with the earlier shared-structural transfer setting is sharper: the absolute-value random-feature rows recover assignments meeting the nonzero-conflict threshold on both targets, while both baselines fail to finish within the common 15-minute cap. Among the displayed rows, the selected `gated_gcn` settings have the smaller reported palettes, reaching $k_{0.05} = 5.2 \pm 0.4$ on CiteSeer and $k_{0.05} = 7.9 \pm 0.3$ on PubMed.

Table 28: Cora-trained absolute-value transfer with random features on full target graphs, together with the two baselines. Rows are mean ± standard deviation over ten random seeds.

Train → Test	Encoder / Method	Pipeline	$k_{0.05}$	Mono@0.05	Conflicts@χ_{greedy}	Mono@χ_{greedy}
Cora → CiteSeer	gated_gcn	v1abs	5.2±0.4	0.0418±0.0059	160.9±19.0	0.0353±0.0042
Cora → CiteSeer	gated_gcn	v2abs	5.7±0.5	0.0438±0.0041	184.8±18.8	0.0406±0.0041
Cora → CiteSeer	PI-GNN	–	timeout	–	–	–
Cora → CiteSeer	full-GCN	–	timeout	–	–	–
Cora → PubMed	gated_gcn	v2abs	7.9±0.3	0.0460±0.0027	2026.5±114.7	0.0457±0.0026
Cora → PubMed	gps_gcn	v1abs	10.3±1.8	0.0460±0.0036	2718.6±493.5	0.0613±0.0111
Cora → PubMed	PI-GNN	–	timeout	–	–	–
Cora → PubMed	full-GCN	–	timeout	–	–	–

K Extra citation network experiments

K.1 Full-Graph Hyperparameter Tuning

Implementation. We ran a dedicated full-graph hyperparameter study on the three strongest encoder families: gated_gcn, gps_gcn, and gps_sage. Unlike the earlier subgraph tuning, all evaluation is now done directly on the full BOW graphs, and all runs are ranked by the primary operating-point objective of this benchmark: smallest $k_{0.05}$, then smaller $\mathrm{Mono}(k_{0.05})$, then smaller $k_{0.02}$, then smaller $\mathrm{Mono}(k_{0.02})$, and finally smaller conflicts at $k = \chi_{\text{greedy}}$. The tuning is split into three stages:

- **Stage 1 (v1 training sweep):** temperature $\{0.10, 0.20, 0.30\}$, learning rate $\{0.003, 0.01\}$, dropout $\{0.0, 0.1, 0.2\}$, and epochs $\{80, 120\}$ for each of the three encoders on each dataset.
- **Stage 2 (v1 architecture sweep):** encoder-capacity refinement with hidden dimension / depth / heads; for GatedGCN this is $(\text{hidden}, \text{layers}) \in \{128, 192, 256\} \times \{2, 3, 4\}$, while for the two GPS variants it is $(\text{hidden}, \text{layers}, \text{heads}) \in \{128, 192\} \times \{2, 3\} \times \{4, 8\}$.
- **Stage 3 (v2 soft-loss sweep):** soft degree power $\{1, 2, 4, 6\}$, soft temperature $\{0.75, 1.0, 1.25, 1.5\}$, and warmup epochs $\{0, 10, 20\}$, keeping $\lambda_{\text{soft}} = 0.30$.

The listed Cartesian grids contain 324, 75, and 432 configurations respectively, totaling 831 planned configurations. Completion is reported for the grid, but individual run records are unavailable; the following comparisons use aggregate summaries.

The selected settings differ across datasets. Writing T for the training temperature, Cora uses gated_gcn v1 with $T = 0.30$, learning rate 0.003, dropout 0.2, and 120 epochs; CiteSeer uses gated_gcn v2 with soft degree power $p = 1$, $\tau_{\text{soft}} = 1.25$, and warmup 10; PubMed uses gps_gcn v1 with $T = 0.30$, learning rate 0.003, dropout 0.1, and 120 epochs.

The selected dataset-specific rows all reach the 0.05 threshold in every seed, but their means differ from the separate single-run operating points. Cora and CiteSeer both have $k_{0.05} = 4.0 \pm 0.0$, with CiteSeer cleaner at Mono@0.05 0.0258 ± 0.0056 and Cora at 0.0380 ± 0.0036. PubMed remains harder, with $k_{0.05} = 5.6 \pm 1.0$ and $k_{0.02} = 11.5 \pm 1.4$.

If a single pipeline and a single hyperparameter setting must be reused across all three full graphs, the reported shared recipe is gps_gcn v1 with $T = 0.30$, learning rate 0.003, dropout 0.1, and 120 epochs. Its aggregate results give $k_{0.05} = 4.0 \pm 0.0$ on Cora and CiteSeer and reaches $k_{0.05} = 4.2 \pm 0.4$ on PubMed, with lower conflicts at χ_{greedy} than the dataset-specific PubMed row.

Table 29: Dataset-specific selected settings; these are not verified global best configurations. Entries are mean ± standard deviation over ten random seeds; NA denotes an unavailable threshold summary; per-seed failure counts are unavailable.

Dataset	$k_{0.05}$	Mono@0.05	$k_{0.02}$	Conflicts@χ_{greedy}	Mono@χ_{greedy}
Cora	4.0±0.0	0.0380±0.0036	5.0±0.0	71.3±9.4	0.0135±0.0018
CiteSeer	4.0±0.0	0.0258±0.0056	4.8±0.4	24.4±10.4	0.0054±0.0023
PubMed	5.6±1.0	0.0391±0.0038	11.5±1.4	1214.4±281.1	0.0274±0.0063

Table 30: Shared full-graph configuration under the constraint that one pipeline and one hyperparameter setting are reused on all three datasets. Entries are mean ± standard deviation over ten random seeds; NA denotes an unavailable threshold summary; per-seed failure counts are unavailable.

Dataset	$k_{0.05}$	Mono@0.05	$k_{0.02}$	Conflicts@χ_{greedy}	Mono@χ_{greedy}
Cora	4.0±0.0	0.0197±0.0102	4.3±0.5	18.9±11.3	0.0036±0.0021
CiteSeer	4.0±0.0	0.0196±0.0056	4.5±0.5	11.6±4.6	0.0025±0.0010
PubMed	4.2±0.4	0.0406±0.0050	7.5±1.1	718.3±117.3	0.0162±0.0026

K.2 GPS With Unitary First Layer

Implementation. The preceding tuning changes hyperparameters within each architecture. Here we test an architectural alternative: replacing the first local GPS message-passing layer by the real-valued orthogonal update of unitary_mp, while keeping the remaining GPS transformer structure unchanged. We test the GCN and GraphSAGE backbones shown in Table 31, using the same v1/v2 objectives, greedy pseudo-labels, embedding inference, and k-medoids decoder as in the earlier sweep.

Table 31: Cora subgraph-to-subgraph GPS unitary-first results. Baseline GPS rows are included for direct comparison. Entries are mean ± standard deviation over ten random seeds.

Encoder	Pipeline	k mean	Mono mean	k/χ mean	Hit-rate
gps_unitary_first_sage	v1	4.35±0.30	0.0221±0.0044	1.03±0.08	1.00±0.00
gps_sage	v1	4.48±0.19	0.0186±0.0025	1.06±0.04	1.00±0.00
gps_unitary_first_gcn	v1	4.47±0.36	0.0196±0.0034	1.06±0.09	1.00±0.00
gps_sage	v2	4.50±0.19	0.0197±0.0037	1.07±0.06	1.00±0.00
gps_unitary_first_sage	v2	4.50±0.27	0.0198±0.0024	1.07±0.06	1.00±0.00
gps_gcn	v2	4.53±0.29	0.0191±0.0033	1.08±0.07	1.00±0.00
gps_unitary_first_gcn	v2	4.57±0.20	0.0218±0.0039	1.09±0.07	1.00±0.00
gps_gcn	v1	4.61±0.28	0.0189±0.0031	1.09±0.06	1.00±0.00

Table 32: Full-graph GPS unitary-first comparison. Entries are mean ± standard deviation over ten random seeds; NA denotes an unavailable threshold summary; per-seed failure counts are unavailable.

Dataset	Encoder group	Pipeline / Encoder	$k_{0.05}$	Mono@0.05	$k_{0.02}$	Conflicts@χ_{greedy}
Cora	GPS baseline	v2 gps_sage	NA	0.0261±0.0114	NA	96.6±236.1
Cora	GPS+unitary first	v2 gps_unitary_first_sage	4.8±2.5	0.0314±0.0076	NA	94.1±209.2
CiteSeer	GPS baseline	v1 gps_sage	4.0±0.0	0.0321±0.0056	5.1±0.3	22.6±15.3
CiteSeer	GPS+unitary first	v2 gps_unitary_first_sage	4.0±0.0	0.0250±0.0071	4.9±0.3	11.3±9.0
PubMed	GPS baseline	v1 gps_gcn	5.7±0.8	0.0394±0.0056	11.4±1.6	1176.8±156.6
PubMed	GPS+unitary first	v1 gps_unitary_first_gcn	5.1±0.3	0.0434±0.0030	10.6±1.1	1120.5±96.9

The unitary-first GPS modification is useful but not uniformly dominant. On the Cora subgraph benchmark, gps_unitary_first_sage v1 gives the best color efficiency, while the lowest Mono remains the plain gps_sage v1 baseline. On full graphs, the result is mixed: the unitary-first variant improves the selected CiteSeer GPS row and lowers the PubMed $k_{0.05}$ and conflicts at χ_{greedy}, but it does not uniformly improve Mono and the Cora comparison remains noisy. This suggests that the unitary first layer is a useful candidate architecture, but it should not replace the plain GPS backbones as the default without dataset-specific validation.

K.3 Cross-study operating points

To compare the scales and tradeoffs across these studies, the following table gathers selected operating points discussed in the main text. The corresponding experiment appendices provide the protocols, failures, and selection limitations needed to interpret each row.

Table 33: Representative operating points. Cora ablation rows report subgraph means without seed uncertainty; remaining rows report ten-seed mean ± standard deviation. ρ uses DSATUR for citation graphs and the known chromatic number for the synthetic suite. The experiment descriptions specify the aggregation procedures and exclusions.

Setting	Recipe	ρ	Mono	Hit
Cora / GCN	v1	1.42	0.0327	1.00
Cora / GCN	v1abs	1.44	0.0251	1.00
Cora→CiteSeer	GatedGCN, v2abs	1.33 ± .09	.0332 ± .0033	1.00 ± .00
Cora→PubMed	GatedGCN, v2abs	1.24 ± .07	.0306 ± .0017	1.00 ± .00
40-graph suite	GatedGCN, v1abs	1.48 ± .10	.0216 ± .0026	.89 ± .01
Large cycles	GatedGCN, v1abs	1.42 ± .22	.0133 ± .0100	1.00 ± .00

L COLOR experiments

We next test whether a model trained on several graphs from one family transfers to other members of that family. COLOR graphs have no native node attributes, so this study also compares random

inputs with structural descriptors. The reporting conventions and limits of the available records in Appendix I apply throughout.

Implementation. The benchmark is constructed as an in-family generalization study on COLOR graphs. For each family, training graphs are drawn from the same family as the evaluation graphs, and performance is measured on a held-out subset denoted ID and a larger OOD subset from that family. We use three encoders selected from the citation experiments: gated_gcn, gps_gcn, and gps_sage, each evaluated under both the v1 and v2 pipelines.

All runs use the same contrastive training and k-medoids decoding pipeline as in the Cora experiments. The v1 objective is the signed variant of the training loss (2), while v2 adds the tuned soft conflict term inherited from the later Cora study. Across all COLOR runs, we use 80 training epochs, common input dimension $d = 64$, and mono-threshold target 0.05.

Feature modes. Because COLOR graphs have no native node attributes, we compare two synthetic feature constructions.

- **Random:** random unit vectors $x_i \in \mathbb{R}^{64}$.
- **Struct:** deterministic structural descriptors. The map ϕ_{struct} expands four normalized node statistics—degree, local clustering coefficient, core number, and PageRank—to give the input vector

  $$x_i = \phi_{\text{struct}}(i).$$

  For a base scalar feature $x \in [0, 1]$, the expansion cycles through

  $$x, \qquad x^2, \qquad \sin(\pi x), \qquad \cos(\pi x),$$

  repeating these transforms across the four base statistics until 64 coordinates are filled. This matches the chosen input dimension but does not create 64 independent structural statistics; the encoders could in principle accept four-dimensional inputs.

Notation and comparisons. Let k be the smallest tested palette meeting the conflict threshold when one is found, or the recorded fallback palette otherwise. We normalize it by the benchmark's greedy-coloring proxy χ_{greedy} to obtain

$$\rho = \frac{k}{\chi_{\text{greedy}}}.$$

Table 34 summarizes the family splits, while Table 35 reports a separate summary of selected configurations. The three per-family tables give the full breakdown for book, Myciel, and queen graphs. For absolute-value variants, we report only the single best instance across v1abs, v2abs, and the different architectures; these selected rows therefore do not constitute a complete matched signed/absolute ablation. We run the unsupervised baselines on the ID or OOD test graphs themselves.

Table 34: Family splits used in the completed COLOR benchmark.

Family	Train graphs	ID graphs	OOD graphs	Train node range	ID node range	OOD node range
Book	3	1	1	74–138	87	561
Myciel	2	1	1	47–95	191	383
Queen	6	2	5	64–144	169–196	225–484

Table 35: Selected configuration per family from a separate experiment summary. These aggregates conflict with the detailed per-family tables and are excluded from best-configuration conclusions. Entries are mean ± standard deviation over ten random seeds.

Family	Feature mode	Pipeline	Encoder	ρ_{ID}	Mono_{ID}	Hit_{ID}	ρ_{OOD}	Mono_{OOD}	Hit_{OOD}
Book	Random	v1	gps_sage	$1.41_{\pm 0.18}$	$0.0451_{\pm 0.0035}$	$1.00_{\pm 0.00}$	$1.08_{\pm 0.08}$	$0.0452_{\pm 0.0034}$	$1.00_{\pm 0.00}$
Myciel	Struct	v1	gps_gcn	$1.45_{\pm 0.28}$	$0.0395_{\pm 0.0041}$	$1.00_{\pm 0.00}$	$1.54_{\pm 0.40}$	$0.0438_{\pm 0.0045}$	$1.00_{\pm 0.00}$
Queen	Random	v2	gps_sage	$0.90_{\pm 0.03}$	$0.0474_{\pm 0.0018}$	$1.00_{\pm 0.00}$	$0.76_{\pm 0.02}$	$0.0471_{\pm 0.0008}$	$1.00_{\pm 0.00}$

Table 36: Book-family results. All book evaluation graphs have known exact chromatic number, and in this family the greedy proxy equals the exact value for the evaluation split, so the reported denominator is the known chromatic number; nonzero-conflict outputs are still not proper colorings. For baseline rows, Hit records the reported completion fraction. Rows are mean ± standard deviation over ten random seeds.

Feature mode	Pipeline/Encoder	ρ_{ID}	$\mathrm{Mono}_{\mathrm{ID}}$	$\mathrm{Hit}_{\mathrm{ID}}$	ρ_{OOD}	$\mathrm{Mono}_{\mathrm{OOD}}$	$\mathrm{Hit}_{\mathrm{OOD}}$
Random	v1 / gps_sage	1.39±0.20	0.0463±0.0028	1.00±0.00	1.05±0.06	0.0454±0.0031	1.00±0.00
	v1 / gps_gcn	1.38±0.18	0.0451±0.0045	1.00±0.00	1.13±0.14	0.0459±0.0028	1.00±0.00
	v1 / gated_gcn	1.31±0.24	0.0446±0.0051	1.00±0.00	1.05±0.14	0.0469±0.0020	1.00±0.00
	v2 / gps_sage	1.40±0.24	0.0426±0.0065	1.00±0.00	1.04±0.14	0.0456±0.0029	1.00±0.00
	v2 / gps_gcn	1.38±0.21	0.0461±0.0031	1.00±0.00	1.21±0.10	0.0468±0.0028	1.00±0.00
	v2 / gated_gcn	1.36±0.18	0.0416±0.0073	1.00±0.00	1.00±0.13	0.0469±0.0027	1.00±0.00
	v2abs / gated_gcn	1.48±0.24	0.0414±0.0054	1.00±0.00	1.28±0.17	0.0466±0.0018	1.00±0.00
Struct	v1 / gated_gcn	2.79±0.15	0.0744±0.0034	0.00±0.00	2.05±0.31	0.1300±0.0099	0.00±0.00
	v1 / gps_gcn	2.77±0.15	0.0808±0.0055	0.00±0.00	2.19±0.23	0.1079±0.0116	0.00±0.00
	v1 / gps_sage	2.73±0.20	0.0791±0.0067	0.00±0.00	2.23±0.27	0.1244±0.0095	0.00±0.00
	v2 / gated_gcn	2.62±0.38	0.0749±0.0031	0.00±0.00	2.01±0.27	0.1270±0.0082	0.00±0.00
	v2 / gps_gcn	2.58±0.33	0.0805±0.0060	0.00±0.00	2.10±0.25	0.1203±0.0070	0.00±0.00
	v2 / gps_sage	2.69±0.17	0.0798±0.0073	0.00±0.00	2.20±0.16	0.1102±0.0157	0.00±0.00
Baseline	PI-GNN	–	–	0.00	–	–	0.00
	full-GCN	0.55	0.0394	1.00	–	–	0.00

Table 37: Myciel-family results. All evaluation graphs have known exact chromatic number, so the reported denominator is the known chromatic number; nonzero-conflict outputs are still not proper colorings. For baseline rows, Hit records the reported completion fraction. Rows are mean ± standard deviation over ten random seeds; baseline rows are unchanged.

Feature mode	Pipeline/Encoder	ρ_{ID}	$\mathrm{Mono}_{\mathrm{ID}}$	$\mathrm{Hit}_{\mathrm{ID}}$	ρ_{OOD}	$\mathrm{Mono}_{\mathrm{OOD}}$	$\mathrm{Hit}_{\mathrm{OOD}}$
Random	v1 / gated_gcn	2.09±0.13	0.0457±0.0027	1.00±0.00	2.21±0.18	0.0464±0.0036	1.00±0.00
	v1 / gps_gcn	1.62±0.16	0.0427±0.0030	1.00±0.00	1.60±0.17	0.0458±0.0034	1.00±0.00
	v1 / gps_sage	2.26±0.30	0.0451±0.0041	1.00±0.00	2.39±0.31	0.0455±0.0049	1.00±0.00
	v1abs / gps_gcn	2.66±0.17	0.0456±0.0030	1.00±0.00	2.46±0.15	0.0476±0.0013	1.00±0.00
	v2 / gated_gcn	2.11±0.34	0.0446±0.0028	1.00±0.00	2.16±0.14	0.0448±0.0037	1.00±0.00
	v2 / gps_gcn	1.65±0.20	0.0434±0.0036	1.00±0.00	1.66±0.17	0.0466±0.0026	1.00±0.00
	v2 / gps_sage	1.52±0.24	0.0439±0.0060	1.00±0.00	1.80±0.22	0.0439±0.0049	1.00±0.00
Struct	v1 / gated_gcn	1.65±0.16	0.0416±0.0074	1.00±0.00	1.51±0.17	0.0438±0.0055	1.00±0.00
	v1 / gps_gcn	1.24±0.25	0.0456±0.0039	1.00±0.00	1.49±0.27	0.0431±0.0060	1.00±0.00
	v1 / gps_sage	1.61±0.28	0.0389±0.0071	1.00±0.00	1.81±0.29	0.0453±0.0039	1.00±0.00
	v2 / gated_gcn	1.59±0.29	0.0397±0.0063	1.00±0.00	1.59±0.34	0.0411±0.0067	1.00±0.00
	v2 / gps_gcn	1.51±0.39	0.0444±0.0040	1.00±0.00	1.76±0.30	0.0425±0.0058	1.00±0.00
	v2 / gps_sage	1.93±0.38	0.0425±0.0061	1.00±0.00	1.76±0.36	0.0440±0.0050	1.00±0.00
Baseline	PI-GNN	–	–	0.00	–	–	0.00
	full-GCN	0.50	0.0373	1.00	–	–	0.00

Table 38: Queen-family results. Here $\rho = k/\chi_{\text{greedy}}$ is a proxy metric, because exact chromatic numbers are unavailable for most larger queen instances. For baseline rows, Hit records the reported completion fraction. Rows are mean ± standard deviation over ten random seeds.

Feature mode	Pipeline/Encoder	ρ_{ID}	Mono_{ID}	Hit_{ID}	ρ_{OOD}	Mono_{OOD}	Hit_{OOD}
Random	v1 / gated_gcn	1.05±0.04	0.0469±0.0015	1.00±0.00	0.83±0.02	0.0472±0.0013	1.00±0.00
	v1 / gps_gcn	0.98±0.05	0.0460±0.0023	1.00±0.00	0.79±0.02	0.0478±0.0008	1.00±0.00
	v1 / gps_sage	1.04±0.05	0.0477±0.0011	1.00±0.00	0.83±0.02	0.0471±0.0008	1.00±0.00
	v1abs / gated_gcn	1.09±0.05	0.0475±0.0008	1.00±0.00	0.87±0.03	0.0475±0.0007	1.00±0.00
	v2 / gated_gcn	1.03±0.04	0.0475±0.0011	1.00±0.00	0.83±0.02	0.0472±0.0013	1.00±0.00
	v2 / gps_gcn	0.97±0.05	0.0471±0.0015	1.00±0.00	0.81±0.01	0.0470±0.0011	1.00±0.00
	v2 / gps_sage	0.97±0.05	0.0459±0.0021	1.00±0.00	0.77±0.02	0.0473±0.0011	1.00±0.00
Struct	v1 / gated_gcn	1.29±0.18	0.2809±0.0471	0.00±0.00	0.88±0.10	0.2831±0.0425	0.00±0.00
	v1 / gps_gcn	1.73±0.05	0.0885±0.0010	0.00±0.00	1.31±0.01	0.0957±0.0010	0.00±0.00
	v1 / gps_sage	1.73±0.04	0.0903±0.0051	0.00±0.00	1.32±0.02	0.0961±0.0029	0.00±0.00
	v2 / gated_gcn	1.12±0.28	0.2548±0.0577	0.00±0.00	0.93±0.17	0.2620±0.0553	0.00±0.00
	v2 / gps_gcn	1.75±0.03	0.0886±0.0011	0.00±0.00	1.32±0.01	0.0961±0.0010	0.00±0.00
	v2 / gps_sage	1.43±0.19	0.1436±0.0527	0.00±0.00	1.10±0.13	0.1572±0.0565	0.00±0.00
Baseline	PI-GNN	–	–	0.00	–	–	0.00
	full-GCN	–	–	0.00	–	–	0.00

Interpretation. Random-feature rows provide low-conflict operating points, whereas structural features miss the threshold in the listed Book and Queen rows. The evaluation set is small (2 Book, 2 Myciel, 7 Queen graphs), however, and configurations were selected using OOD scores. These results consequently describe an exploratory comparison within the tested families. The ID designation also does not mean within the training size range: Myciel ID size 191 exceeds training sizes 47–95, and Queen ID sizes 169–196 exceed 64–144.

A further limitation concerns the configuration-selection summary, which disagrees with the detailed tables for identically named configurations. For Book v1/GPS-SAGE the OOD ratios are $1.08 \pm .08$ and $1.05 \pm .06$; for structural Myciel v1/GPS-GCN they are $1.54 \pm .40$ and $1.49 \pm .27$; for Queen v2/GPS-SAGE they are $.76 \pm .02$ and $.77 \pm .02$. We report both aggregates and do not infer a verified winner from them. Likewise, baseline completion failures do not establish graph infeasibility or learned solver superiority.

M CYCLE EXPERIMENTS

The cycle benchmark separates transfer to larger graphs from transfer to other graph structures: we train only on cycles, then evaluate on both much larger cycles and a small suite containing non-cycle graphs. The reporting conventions and limits of the available records in Appendix I apply throughout.

Implementation. This section studies a cycle generalization benchmark in which training is performed only on cycles C_n with $50 \leq n \leq 200$, using random node initialization. The learned methods are GatedGCN, UnitaryMP, GPS-GCN, and GPS-SAGE, each run in a v1abs-style and v2abs-style pipeline. The test set contains 40 small graphs and 20 large cycles. The small subset contains 20 cycles $C_{20}, \ldots, C_{39}$ and 20 non-cycle graphs, while the large subset contains $C_{7000}, \ldots, C_{7019}$. We additionally evaluate the PI-GNN baseline of Schuetz et al. (2022) and the warm-start Coloring-GNN baseline of Vanderbush & Weber (2026). Both baselines are run on the same test graphs with a hard per-graph runtime budget of 900 seconds. Experiments not completed within this time budget are denoted as "timeout" in the report.

Combined averages. Tables 39 and 40 summarize the 40-graph small suite and the 20 large-cycle stress suite under ten random seeds, allowing us to compare the two objectives across the four encoders.

Comparison with baselines. We use v1abs + GatedGCN as the learned reference in Table 41: it has the best large-cycle ratio among the reported learned configurations, competitive small-suite behavior, and assignments meeting the conflict threshold on all large cycles. The comparison asks how this selected operating point compares with the two per-graph solvers.

Table 39: Combined averages for the v1abs-style random-feature benchmark. Entries are mean ± standard deviation over ten random seeds.

Encoder	Small k/χ	Small Mono	Small Hit	Large k/χ	Large Mono	Large Hit
GatedGCN	1.48±0.10	0.0216±0.0026	0.89±0.01	1.42±0.22	0.0133±0.0100	1.00±0.00
UnitaryMP	1.48±0.06	0.0214±0.0024	0.91±0.01	1.67±0.00	0.0096±0.0012	1.00±0.00
GPS-GCN	1.64±0.11	0.0213±0.0019	0.92±0.02	1.93±0.19	0.0270±0.0082	1.00±0.00
GPS-SAGE	1.74±0.11	0.0238±0.0021	0.90±0.02	2.01±0.19	0.0261±0.0110	1.00±0.00

Table 40: Combined averages for the v2abs-style random-feature benchmark. Entries are mean ± standard deviation over ten random seeds.

Encoder	Small k/χ	Small Mono	Small Hit	Large k/χ	Large Mono	Large Hit
GatedGCN	1.44±0.06	0.0224±0.0020	0.90±0.01	1.62±0.13	0.0058±0.0076	1.00±0.00
UnitaryMP	1.44±0.03	0.0226±0.0025	0.92±0.02	1.62±0.13	0.0131±0.0073	1.00±0.00
GPS-GCN	1.87±0.11	0.0235±0.0021	0.90±0.02	2.20±0.28	0.0269±0.0107	1.00±0.00
GPS-SAGE	1.57±0.10	0.0196±0.0028	0.94±0.03	1.66±0.15	0.0199±0.0101	1.00±0.00

Table 41: Aggregate comparison against PI-GNN and the warm-start ColoringGNN baseline. Learned rows are mean ± standard deviation over ten random seeds; external baseline rows are unchanged. Total learned runtime is the sum of reported mean training and evaluation time. Its standard deviation is omitted because the covariance needed to combine these terms is unavailable.

Method	Split	Completed	Timeouts	k/χ	Mono	Hit	Total runtime (s)
v1abs + GatedGCN	Small 40 graphs	40	0	1.48±0.10	0.0216±0.0026	0.89±0.01	6.82
PI-GNN	Small 40 graphs	39	1	1.009	0.0126	1.000	4701.1
Warm-start ColoringGNN	Small 40 graphs	40	0	0.987	0.0141	1.000	2282.7
v1abs + GatedGCN	Large 20 cycles	20	0	1.42±0.22	0.0133±0.0100	1.00±0.00	6.38
PI-GNN	Large 20 cycles	0	20	–	–	–	18000.0
Warm-start ColoringGNN	Large 20 cycles	0	20	–	–	–	18000.0

Runtime. To distinguish training cost from the cost of a new instance, we report learned-model training time once per configuration and test-time runtime as the average decode time per graph. For PI-GNN and the warm-start ColoringGNN baseline, the runtime table reports per-graph solve time because these baselines are evaluated graph by graph under the same k sweep.

Table 42: Runtime summary for the learned random-feature cycle models. Entries are mean ± standard deviation over ten random seeds.

Pipeline	Encoder	Train (s)	Avg small test (s)	Avg large test (s)
v1abs	GatedGCN	5.88±0.36	0.0236±0.0199	0.0253±0.0075
v1abs	UnitaryMP	7.57±0.05	0.0162±0.0008	0.0345±0.0006
v1abs	GPS-GCN	12.12±0.07	0.0187±0.0014	0.0640±0.0076
v1abs	GPS-SAGE	9.17±0.05	0.0190±0.0014	0.0537±0.0079
v2abs	GatedGCN	10.38±0.06	0.0147±0.0007	0.0333±0.0049
v2abs	UnitaryMP	12.36±0.07	0.0155±0.0004	0.0329±0.0044
v2abs	GPS-GCN	13.96±0.09	0.0200±0.0016	0.0603±0.0107
v2abs	GPS-SAGE	17.00±0.09	0.0179±0.0009	0.0446±0.0056

Table 43: Runtime summary for PI-GNN and the warm-start ColoringGNN baseline under a 15-minute per-graph budget.

Method	Split	Avg all (s)	Avg completed (s)	Max (s)	Timeouts
PI-GNN	Small 40 graphs	117.53	97.47	900.0	1
PI-GNN	Large 20 cycles	900.00	–	900.0	20
Warm-start ColoringGNN	Small 40 graphs	57.07	57.07	415.5	0
Warm-start ColoringGNN	Large 20 cycles	900.00	–	900.0	20

Results. The next per-graph comparison focuses on two representative learned configurations together with the two external baselines. We compare the balanced v1abs + GatedGCN reference recipe, the v2abs + GPS-SAGE recipe used in the main-text per-graph table, PI-GNN, and the warm-start ColoringGNN baseline on each graph. Learned cells report seed means and standard deviations, while external baseline cells are single-run values; timed-out baseline graphs are shown as "timeout".

For v1abs + GatedGCN, the ten-seed aggregate reaches small-suite $k/\chi = 1.48 \pm 0.10$, Mono $= 0.0216 \pm 0.0026$, and hit-rate 0.89 ± 0.01; on the 20 large cycles it reaches $k/\chi = 1.42 \pm 0.22$, Mono $= 0.0133 \pm 0.0100$, and hit-rate 1.00 ± 0.00. The v2abs GatedGCN and UnitaryMP rows have slightly lower small-suite ratios (1.44), but they use larger tested palettes on the large-cycle stress test. Thus the preferred learned configuration depends on which suite and palette–conflict tradeoff we consider.

On the small 40-graph subset, PI-GNN and the warm-start ColoringGNN baseline are more color-efficient, but both time out on all large-cycle instances under the 15-minute cap. The learned models pay a one-time training cost and then decode the large cycles quickly, while these two baseline implementations reach the cap on this stress suite. This supports the value of amortized inference for the tested implementations. Cycles themselves admit elementary linear-time proper colorings, so the timeouts establish neither intrinsic coloring difficulty nor an advantage over specialized algorithms.

N Comparisons with per-instance coloring solvers

Evidence and protocol. To broaden the comparison with per-instance coloring solvers, we evaluate three methods under fixed-palette and legal-search protocols. The per-attempt records contain 1452 distinct attempts: 121 instances, two seeds, three methods, and two protocols. The instances comprise 2 Book, 2 Myciel, 7 Queen, 50 Cora subgraphs, and 60 graphs in the cycle-training test suite (including its non-cycle test graphs). We compute counts, target rates, proper-output medians, and completed-run runtime medians from these 1452 records. The summaries reflect recorded solver outcomes; the individual assignments have not been independently revalidated.

Table 44: Per-graph results on the 20 small cycle graphs. Cells are reported as k / Mono. Learned-method cells are mean ± standard deviation over ten random seeds; baseline cells are unchanged.

Graph	n	χ	v1abs + GatedGCN	v2abs + GPS-SAGE	PI-GNN	Warm-start ColoringGNN
C_20	20	2	3.4±0.5 / 0.0000±0.0000	4.0±0.8 / 0.0000±0.0000	3 / 0.0000	3 / 0.0000
C_21	21	3	3.3±0.5 / 0.0190±0.0246	4.3±2.1 / 0.0143±0.0230	3 / 0.0000	2 / 0.0476
C_22	22	2	3.4±0.5 / 0.0227±0.0240	4.3±0.7 / 0.0182±0.0235	2 / 0.0000	2 / 0.0000
C_23	23	3	3.2±0.4 / 0.0217±0.0229	3.8±0.8 / 0.0087±0.0183	2 / 0.0435	2 / 0.0435
C_24	24	2	3.3±0.5 / 0.0083±0.0176	4.1±1.1 / 0.0292±0.0201	3 / 0.0000	3 / 0.0000
C_25	25	3	3.2±0.4 / 0.0160±0.0207	4.1±0.6 / 0.0120±0.0193	2 / 0.0400	3 / 0.0000
C_26	26	2	3.5±0.7 / 0.0115±0.0186	3.7±0.9 / 0.0192±0.0203	3 / 0.0000	3 / 0.0000
C_27	27	3	3.6±1.0 / 0.0185±0.0195	4.5±1.3 / 0.0074±0.0156	3 / 0.0000	2 / 0.0370
C_28	28	2	3.4±0.5 / 0.0107±0.0173	3.9±0.6 / 0.0143±0.0184	3 / 0.0000	3 / 0.0000
C_29	29	3	3.3±0.5 / 0.0034±0.0109	4.1±1.2 / 0.0241±0.0167	2 / 0.0345	3 / 0.0000
C_30	30	2	3.3±0.5 / 0.0067±0.0141	3.8±0.6 / 0.0133±0.0172	3 / 0.0000	3 / 0.0000
C_31	31	3	3.3±0.5 / 0.0032±0.0102	4.0±1.1 / 0.0032±0.0102	3 / 0.0000	3 / 0.0000
C_32	32	2	3.6±0.8 / 0.0125±0.0161	4.2±0.8 / 0.0063±0.0132	3 / 0.0000	2 / 0.0000
C_33	33	3	3.4±0.7 / 0.0121±0.0156	4.1±0.9 / 0.0121±0.0156	2 / 0.0303	2 / 0.0303
C_34	34	2	3.7±0.8 / 0.0059±0.0124	3.7±0.5 / 0.0118±0.0152	2 / 0.0000	3 / 0.0000
C_35	35	3	3.6±0.7 / 0.0086±0.0138	4.0±0.8 / 0.0114±0.0148	3 / 0.0000	2 / 0.0286
C_36	36	2	3.6±0.5 / 0.0111±0.0143	4.0±0.5 / 0.0167±0.0143	3 / 0.0000	3 / 0.0000
C_37	37	3	3.4±0.5 / 0.0027±0.0085	4.3±1.1 / 0.0081±0.0131	3 / 0.0000	2 / 0.0270
C_38	38	2	3.5±0.5 / 0.0079±0.0127	3.9±1.0 / 0.0132±0.0139	3 / 0.0000	3 / 0.0000
C_39	39	3	3.4±0.5 / 0.0103±0.0132	4.5±1.1 / 0.0077±0.0124	3 / 0.0000	3 / 0.0000

Table 45: Per-graph results on the 20 non-cycle graphs. Cells are reported as k / Mono. A timed-out baseline graph is shown as "timeout". Learned-method cells are mean ± standard deviation over ten random seeds; baseline cells are unchanged.

Graph	n	χ	v1abs + GatedGCN	v2abs + GPS-SAGE	PI-GNN	Warm-start ColoringGNN
K_5	5	5	5.0±0.0 / 0.0000±0.0000	5.0±0.0 / 0.0000±0.0000	5 / 0.0000	5 / 0.0000
K_8	8	8	7.0±0.0 / 0.0357±0.0000	7.0±0.0 / 0.0357±0.0000	8 / 0.0357	7 / 0.0357
K_10	10	10	8.1±0.3 / 0.0422±0.0070	8.3±0.5 / 0.0378±0.0107	8 / 0.0444	8 / 0.0444
K_{5,5}	10	2	3.0±1.9 / 0.0040±0.0126	3.3±1.6 / 0.0040±0.0126	2 / 0.0000	2 / 0.0000
K_{4,8}	12	2	4.7±2.9 / 0.0063±0.0132	2.8±0.6 / 0.0000±0.0000	2 / 0.0000	2 / 0.0000
K_{3,10}	13	2	4.2±2.4 / 0.0067±0.0141	3.4±0.8 / 0.0033±0.0105	2 / 0.0000	2 / 0.0000
W_11	11	3	7.1±2.0 / 0.0000±0.0000	6.6±1.6 / 0.0000±0.0000	3 / 0.0000	3 / 0.0000
W_12	12	4	5.0±1.3 / 0.0273±0.0235	4.6±1.3 / 0.0318±0.0220	3 / 0.0455	3 / 0.0455
W_13	13	3	5.6±2.2 / 0.0208±0.0220	5.5±1.4 / 0.0292±0.0201	3 / 0.0000	3 / 0.0000
W_14	14	4	5.1±0.9 / 0.0231±0.0199	5.9±2.5 / 0.0269±0.0186	3 / 0.0385	3 / 0.0385
Petersen	10	3	4.6±1.3 / 0.0000±0.0000	5.2±1.7 / 0.0000±0.0000	3 / 0.0000	3 / 0.0000
Icosahedral	12	4	7.0±1.4 / 0.0300±0.0105	6.4±1.4 / 0.0267±0.0141	4 / 0.0333	5 / 0.0000
KG(7,2)	21	5	8.9±1.0 / 0.0610±0.0225	8.6±1.2 / 0.0457±0.0040	4 / 0.0286	4 / 0.0286
KG(9,3)	84	5	8.7±1.1 / 0.0969±0.0147	9.6±0.5 / 0.0661±0.0133	4 / 0.0119	4 / 0.0119
KG(10,3)	120	6	10.0±0.0 / 0.1071±0.0220	9.2±1.0 / 0.0924±0.0160	timeout	4 / 0.0333
Mycielski(C5)^0	5	3	3.0±0.0 / 0.0000±0.0000	3.4±0.5 / 0.0000±0.0000	3 / 0.0000	3 / 0.0000
Mycielski(C5)^1	11	4	5.9±2.2 / 0.0050±0.0158	4.5±1.0 / 0.0000±0.0000	3 / 0.0500	3 / 0.0500
Mycielski(C5)^2	23	5	7.1±1.6 / 0.0310±0.0129	6.6±2.1 / 0.0352±0.0137	4 / 0.0141	4 / 0.0141
Mycielski(C5)^3	47	6	9.2±1.0 / 0.0665±0.0182	9.5±0.5 / 0.0449±0.0100	4 / 0.0169	4 / 0.0254
Mycielski(C5)^4	95	7	9.1±0.9 / 0.0873±0.0167	9.4±0.5 / 0.0519±0.0155	4 / 0.0225	4 / 0.0212

We compare the gCol HybridEA implementation of HEA (Galinier & Hao, 1999), the official ANYCSP implementation with a retrained COL configuration (Tönshoff et al., 2023), and an independent GIN-based baseline inspired by Zhang et al. (2024), with TabuCol refinement (Hertz & de Werra, 1987). The latter is not author-released code. Each method is evaluated under two protocols. Fixed-k supplies the recorded target palette, while legal search starts from a shared DSATUR assignment and searches for smaller proper palettes. Targets are exact where recorded as known and otherwise use the benchmark proxy, so reaching a target need not certify chromatic optimality. The harness checks properness and palette constraints separately: for example, HEA returns proper Queen assignments in every fixed-k attempt, but two exceed their prescribed target palette.

The accounting allows 900 seconds per attempt. For ANYCSP, it charges an approximately 800-second pretrained snapshot to every instance, leaving approximately 100 seconds for search; this conservative convention differs from amortizing training once over the benchmark and cannot rank the usual pretrained inference protocol. Zhang-GIN's per-instance optimization is included in runtime. We report median times conditional on completed attempts and success percentages over all attempts, including timeouts.

Table 46: Aggregate solver results. Each paired cell is fixed-k/legal search. Target success requires a proper output using at most the prescribed number of colors. Proper colors and seconds are medians among proper and completed outputs respectively. A dash means no eligible output, and all such attempts timed out.

Family	Method	Attempts	Target success (%)	Proper colors	Seconds
Book	HEA	4/4	100.0/100.0	12.0/12.0	0.245/0.686
Book	ANYCSP	4/4	0.0/0.0	–/–	–/–
Book	Zhang-GIN (reimpl.)	4/4	100.0/100.0	12.0/12.0	89.950/283.469
Myciel	HEA	4/4	100.0/100.0	8.5/8.5	0.264/0.582
Myciel	ANYCSP	4/4	0.0/0.0	–/–	–/–
Myciel	Zhang-GIN (reimpl.)	4/4	100.0/100.0	8.5/8.5	91.130/296.534
Queen	HEA	14/14	85.7/85.7	22.0/18.0	0.296/1.667
Queen	ANYCSP	14/14	0.0/0.0	–/–	–/–
Queen	Zhang-GIN (reimpl.)	14/14	85.7/0.0	23.5/–	89.801/–
Cora	HEA	100/100	100.0/100.0	4.0/4.0	0.239/0.751
Cora	ANYCSP	100/100	0.0/0.0	–/–	–/–
Cora	Zhang-GIN (reimpl.)	100/100	100.0/100.0	4.0/4.0	77.084/157.560
Cycle suite	HEA	120/120	100.0/100.0	3.0/3.0	0.240/0.673
Cycle suite	ANYCSP	120/120	26.7/0.0	3.0/–	822.962/–
Cycle suite	Zhang-GIN (reimpl.)	120/120	99.2/83.3	3.0/3.0	78.393/109.930

Failures and attribution. HEA completes all 484 attempts with proper outputs and reaches the target in 480, making it the strongest proper-coloring comparator in this campaign. ANYCSP completes 32 of 484 attempts, all in fixed-k cycle-suite trials; the remaining 452 time out under the stated accounting. Zhang-GIN completes 449 of 484 attempts, returning proper outputs in 447; 35 attempts time out, and two completed fixed-k Queen outputs contain conflicts. No recorded attempt has an error status.

Completion alone does not identify a method's contribution to legal search. All 208 successful Zhang-GIN legal-search attempts on Book, Myciel, Cora, and the cycle suite have no recorded method-produced feasible event: their proper assignments are inherited from the shared DSATUR initialization. We therefore cannot attribute these successes to a learned improvement over DSATUR.

O Details on Experimental Setups

For reproducibility, we record the software and datasets used in the learned-model and baseline experiments. The Python neural-network models were built using PyTorch (v2.5.1+cu121) and PyTorch Geometric (v2.7.0), with graph construction and synthetic benchmark generation handled through NetworkX (v3.4.2). Numerical computation and data processing used NumPy (v2.2.6),

SciPy (v1.15.2), and pandas (v2.3.3). For the DGL-based baseline runs we used DGL (v1.1.3). Models were trained using the optimization and initialization routines provided by the default PyTorch stack, with AdamW Loshchilov & Hutter (2019) used as the optimizer in our learned pipelines.

The reported hardware configuration for these runs is summarized in Table 47.

Table 47: Hardware specifications.

Component	Specification
Architecture	x86_64
OS	Rocky Linux 8.10 (Green Obsidian)
CPU	AMD EPYC 9374F 32-Core Processor ($\times 2$)
GPU	NVIDIA H200
GPU memory	143771 MiB (approximately 141 GB)
RAM	1.5 TiB

The experiments use three groups of datasets. First, the citation-graph experiments use the Cora, CiteSeer, and PubMed citation graphs, loaded through the standard Planetoid benchmark interface in PyTorch Geometric (Yang et al., 2016; Fey & Lenssen, 2019). Second, the COLOR-family experiments use book, Myciel, and queen graphs; the book instances are taken from the CMU COLOR benchmark collection[1] associated with the DIMACS coloring benchmarks (Johnson & Trick, 1996), while the Myciel and queen families are generated programmatically from their standard graph constructions. Third, the cycle experiments use synthetic cycle graphs for training and held-out evaluation, together with additional manually specified benchmark graphs such as complete graphs, complete bipartite graphs, wheel graphs, Kneser graphs, Mycielski graphs, Petersen, Icosahedral, and very large cycles; these graphs are generated through NetworkX or through explicit graph-construction utilities in our codebase.

Table 48 lists the licenses of the main software libraries used in the experiments.

Table 48: Main software licenses.

Software	License
PyTorch	BSD-3-Clause
PyTorch Geometric	MIT
NetworkX	BSD-3-Clause
NumPy	BSD-3-Clause
SciPy	BSD-3-Clause
pandas	BSD-3-Clause
DGL	Apache-2.0

[1] https://mat.tepper.cmu.edu/COLOR02/