STRUCTURAL NEGLECT MATTERS: OVERFITTING AND ITS MITIGATION IN GRAPH CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph contrastive learning (GCL) has emerged as an effective framework for unsupervised graph representation learning, typically relying on GNN-based encoders to capture informative representations from both node attributes and graph structure. Despite the common expectation that optimizing contrastive objectives improves the discriminability of representations for downstream tasks, we reveal that excessive optimization of the graph contrastive objective can instead degrade the learned representations. Specifically, as optimization proceeds, the encoder progressively loses structural information, ultimately reducing the discriminability of the resulting representations for downstream tasks. We refer to this phenomenon as *contrastive overfitting*. Through theoretical analysis and extensive empirical investigation, we further identify the mechanism underlying contrastive overfitting and characterize how excessive contrastive optimization leads to the progressive degradation of structural information. Guided by these findings, we propose Progress–Adaptive REgulation (PARE), a simple yet effective adaptive training strategy that can be integrated into existing GCL methods to mitigate contrastive overfitting. Extensive experiments demonstrate the effectiveness of PARE across various GCL methods.

1 INTRODUCTION

Contrastive learning has emerged as a promising framework for unsupervised representation learning, as it learns representations without relying on annotated labels (Oord et al., 2018; Chen et al., 2020; Tschannen et al., 2020). Graph contrastive learning (GCL) extends this paradigm to graph-structured data and has become a powerful approach for unsupervised graph representation learning (Liu et al., 2022; Ju et al., 2024). GCL typically trains a GNN encoder, such as a graph convolutional network (GCN) (Kipf & Welling, 2017), to learn informative and discriminative node embeddings for downstream tasks such as node classification. Among existing methods, InfoNCE-based contrastive optimization (Oord et al., 2018) has become the dominant training paradigm.

Although previous studies have reported favorable performance gains from InfoNCE-based GCL methods (Zhu et al., 2020; 2021a;b), closer inspection reveals a potential misalignment between the contrastive optimization and downstream performance. Ideally, better optimization of the GCL objective is expected to yield representations with improved downstream performance. However, empirical evidence suggests that this correspondence is inconsistent: convergence of the contrastive objective does not necessarily yield better representations and may even degrade downstream performance. We refer to this misalignment as *contrastive overfitting*.

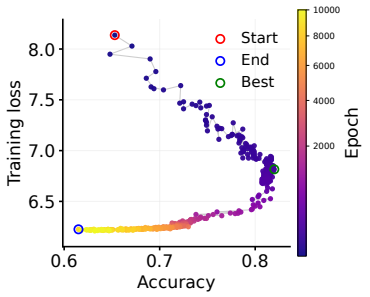


Figure 1: Contrastive overfitting of GRACE (Zhu et al., 2020) on the Cora dataset (Yang et al., 2016).

As shown in Figure 1, while the graph contrastive loss decreases monotonically toward convergence, downstream performance exhibits a clear rise-then-fall trajectory despite the continued decrease in contrastive loss. This pattern is consistently observed across different datasets and models, as detailed in Section 5. It is worth noting that this issue is problematic in a strictly unsupervised setting, where downstream labels are unavailable for model selection and therefore cannot be used for early stopping or hyperparameter tuning. Consequently, the number of training iterations is often

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.

selected through extensive trial-and-error experiments. Such biased evaluations can obscure the gap between contrastive optimization and downstream representation quality.

This anomalous behavior raises two fundamental questions: *What causes contrastive overfitting, and how can it be effectively mitigated?* In this paper, we attribute contrastive overfitting to the topology-agnostic nature of the instance-discrimination objective in GCL. Generally, although GNN encoders encode structural information into node embeddings through message passing, the contrastive objective treats nodes as independent instances and does not explicitly preserve their structural dependencies. Excessive optimization may therefore favor instance-level separability over structural preservation, causing the encoder to gradually discard encoded structural information. It is worth noting that some existing studies have addressed related issues, including false negatives (Xia et al., 2022; Li et al., 2023; Hao et al., 2024), optimization conflict (He et al., 2024; Ji et al., 2024), over-alignment (Liu et al., 2024), and downstream discrepancy (Ma et al., 2023), which align to some extent with our claim. Nevertheless, their improvements remain susceptible to over-optimization and thus fail to directly address the overfitting scenario considered in this work. Further discussion of related work is provided in Appendix A.2.

Motivated by the insight that instance-discrimination objectives in GCL do not explicitly preserve graph topology, we provide both theoretical and empirical evidence that the InfoNCE-based GCL objective can drive a GNN encoder toward a topology-insensitive solution subspace. In Section 3, we first characterize the topology-independent geometry induced by the InfoNCE objective, showing that its optimum is determined by positive-pair alignment and instance-wise separation without requiring graph structural information. We then establish that such structural degeneration can directly impair downstream performance when the target task strongly depends on graph topology. Beyond these theoretical results, we empirically investigate the optimization dynamics of GCL and reveal a consistent transition from an early high-marginal-return regime, where representation updates efficiently reduce the contrastive objective while preserving useful structure, to a late low-marginal-return regime, where continued representation drift yields only marginal contrastive improvement but progressively degrades structural information. This transition explains how excessive contrastive optimization gives rise to contrastive overfitting.

Guided by these findings, in Section 4 we propose a Progress-Adaptive REgulation (PARE) method that dynamically controls encoder updates according to the marginal return of contrastive optimization. Specifically, PARE measures the contrastive gain achieved per unit representation drift and uses this signal to regulate the update strength during training. When the marginal return remains high, the encoder largely follows standard contrastive optimization; as the return diminishes, PARE progressively suppresses further representation movement, thereby reducing unnecessary structural degeneration in the low-contrastive-return regime. The resulting mechanism is architecture-agnostic, and can be readily integrated into existing GCL methods without additional hyperparameter tuning.

We highlight *our contributions* as follows:

- We identify and empirically reveal *contrastive overfitting* in GCL, where continued reduction of the contrastive objective can progressively degrade downstream representations.
- We theoretically characterize a topology-independent optimum of InfoNCE-based GCL, and show that reduced structural information affects structure-dependent downstream tasks. We empirically validate these findings and further reveal the optimization dynamics of contrastive overfitting.
- We propose PARE, a theoretically grounded progress-adaptive regulation framework that dynamically adjusts contrastive optimization according to training progress, providing a general and architecture-agnostic mechanism for mitigating contrastive overfitting in existing GCL methods.

2 PRELIMINARIES

Notations. $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{Y})$ is an undirected graph, where $\mathcal{V} = \{v_i\}_{i=1}^{N}$ is the node set, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set, $\mathbf{X} \in \mathbb{R}^{N \times D}$ is the node attribute matrix, and $\mathbf{Y} \in \{1, \ldots, K\}^N$ denotes node labels. Let $\mathbf{A} \in \{0,1\}^{N \times N}$ denote the adjacency matrix and $\mathcal{N}_i = \{j | (v_i, v_j) \in \mathcal{E}\}$ denote the first-order neighborhood of node v_i. GNN encoder $f_\theta : (\{0,1\}^{N \times N}, \mathbb{R}^{N \times D}) \to \mathbb{R}^{N \times d}$ maps a given graph $(\mathbf{A}, \mathbf{X})$ to its node representation matrix $\mathbf{Z} \in \mathbb{R}^{N \times d}$. The i-th row of $\mathbf{Z}$, denoted as $z_i \in \mathbb{R}^d$, is the

normalized representation of node v_i, i.e., $\|z_i\|_2 = 1$. Let $\mathcal{G}^1$ and $\mathcal{G}^2$ be two augmented views of $\mathcal{G}$, and their corresponding node representations produced by f_θ be $\mathbf{Z}^1$ and $\mathbf{Z}^2$.

Graph contrastive learning. GCL learns node representations by contrasting different augmented views of a graph. Specifically, it pulls representations of the same node across different views, while pushing representations of different nodes apart. Consider z_i^1 as the anchor and z_i^2 as its positive counterpart from the other view. Let τ be the temperature coefficient. Since the similarity between two representations is measured by $\phi(z_i, z_j) = z_i^\top z_j$ for normalized representations, the positive and negative terms are respectively defined as $P_i := \exp(\phi(z_i^1, z_i^2)/\tau)$ and $N_i := \sum_{j\neq i} \exp(\phi(z_i^1, z_j^2)/\tau) + \sum_{j\neq i} \exp(\phi(z_i^1, z_j^1)/\tau)$, following the widely used InfoNCE-based GCL objective in GRACE (Zhu et al., 2020). The corresponding directional contrastive loss is

$$\mathcal{L}(z_i^1, z_i^2) = -\log \frac{P_i}{P_i + N_i} = \log\left(1 + \frac{N_i}{P_i}\right). \tag{1}$$

The reverse-direction loss $\mathcal{L}(z_i^2, z_i^1)$ is defined symmetrically. Averaging the two directions over all N nodes gives the GCL objective:

$$\mathcal{L}_{\text{con}} = \frac{1}{2N} \sum_{i=1}^{N} \left[\mathcal{L}(z_i^1, z_i^2) + \mathcal{L}(z_i^2, z_i^1)\right]. \tag{2}$$

3 Theoretical Analysis and Empirical Investigation

In this section, we first provide theoretical analyses that (i) characterize the geometry of GCL optima and derive conditions under which approaching this optimum limits graph-dependent structural alignment, and (ii) connect structural dependence to downstream performance. Proofs are provided in Appendix C. We then conduct exploratory experiments to demonstrate the prevalence of contrastive overfitting and empirically support the implications of these theoretical results. By jointly tracking contrastive loss reduction and representation drift, we identify two learning stages: a high-marginal-return stage of effective representation learning and a low-marginal-return stage characterized by contrastive overfitting. These insights further motivate the design of our method.

3.1 Topology-Independent Optimal Geometry of GCL

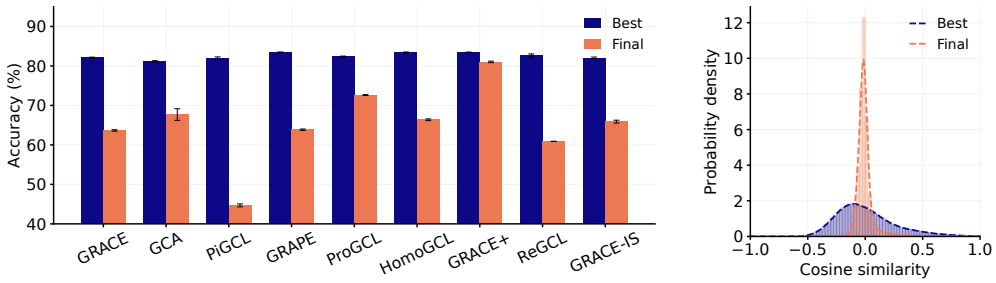
Unlike conventional instance-based data, nodes in a graph are explicitly coupled through graph topology. Informative node representations therefore typically integrate two complementary sources of information: intrinsic node attributes and structural information arising from node relations and positions in the graph. Such structural dependencies constitute a fundamental inductive bias in graph representation learning. Standard instance-level contrastive objectives, however, do not explicitly preserve this bias. InfoNCE aligns the same node across augmented views while separating different nodes irrespective of their structural relations. Although such instance-level uniformity is generally desirable in conventional contrastive learning (Wang & Isola, 2020), it may conflict with the relational nature of graph data. The following theorem characterizes the geometry favored by InfoNCE.

Theorem 1 (Topology-Independent Regular-Simplex Optimum). *If $N \geq 2$, $d \geq 1$ and $\tau > 0$, then*

$$\mathcal{L}_{\text{con}} \geq \log\left[1 + 2(N-1)\exp\Big(-\frac{N}{(N-1)\tau}\Big)\right] + C_{N,d,\tau},$$

where $C_{N,d,\tau} = e^{-2/\tau}\min\{\frac{1}{\tau}, \frac{1}{2\tau^2}\}\big[\frac{1}{d} - \frac{1}{N-1}\big]_+$ and $[x]_+ = \max\{x, 0\}$. When $d \geq N-1$, the equality holds iff $z_i^1 = z_i^2 = z_i$ and $z_i^\top z_j = -1/(N-1)$ for any $i, j \in [N]$ satisfying $i \neq j$.

Theorem 1 characterizes the topology-agnostic geometry favored by the InfoNCE objective. When $d \geq N-1$, $C_{N,d,\tau} = 0$ and the optimum is exactly an aligned regular simplex, where all distinct nodes are equally separated regardless of their structural relations. When $d < N-1$, finite-dimensional constraints induce a fixed gap from the ideal simplex, while the objective retains the same instance-separation pressure without explicitly preserving graph topology. Thus, continued contrastive optimization tends to promote increasingly uniform separation among node representations, potentially weakening topology-induced similarities. This is closely related to the alignment–uniformity perspective of contrastive learning (Wang & Isola, 2020; Yan et al., 2024), while our result



(a) Performance degradation

(b) Representation drift

Figure 2: Contrastive overfitting on Cora across GCL methods.

quantifies finite-dimensional constraints on instance separation. Building on Theorem 1, Corollary 1 in Appendix C.4 further formalizes this implication by showing that graph-dependent structural alignment is necessarily limited near the topology-independent contrastive optimum.

3.2 Structure Dependence Determines the Harm of Contrastive Overfitting

The topology-independent optimum in Theorem 1 suggests that structural information may be lost during contrastive optimization. However, such loss does not always imply degraded downstream performance: it depends on how strongly the downstream task relies on graph structure. Let Y, S, and U denote the label variable, structural variable, and attribute variable of a uniformly sampled node, respectively. We formalize this dependence as follows.

Definition 1 (Structure Dependence)**.** *The structure dependence of the downstream task is defined as*

$$\Delta_{\mathrm{str}} := I(Y;S|U) = H(Y|U) - H(Y|S,U). \tag{3}$$

Δ_{str} is essentially the label information provided by graph structure beyond node attributes. A larger Δ_{str} therefore indicates a more structure-dependent downstream task. Let Z be the embedding variable and W be an observation. The Bayes-optimal log-loss of W is defined as $\mathcal{R}^*_{\log}(W) := \inf_{p(\cdot|W)} \mathbb{E}[-\log p(Y|W)]$. Based on these, we construct the following theorem.

Theorem 2 (Structure Degeneration Hurts Structure-Dependent Tasks)**.** *If $Y \perp Z$ when given S and U, and $I(Z;S|U) \le \epsilon$, then $\mathcal{R}^*_{\log}(Z) - \mathcal{R}^*_{\log}(S,U) \ge \Delta_{\mathrm{str}} - \epsilon$.*

In Theorem 2, $\mathcal{R}^*_{\log}(Z)$ denotes the best achievable log-loss using only the learned representation Z, while $\mathcal{R}^*_{\log}(S,U)$ denotes the Bayes-optimal log-loss when both structural and attribute information are fully available. The theorem shows that their gap is lower bounded by $\Delta_{\mathrm{str}} - \epsilon$. When Δ_{str} is large, a representation with small $I(Z;S|U)$ necessarily incurs a larger lower bound on the performance gap; retaining structural information is therefore particularly important for structure-dependent tasks. In contrast, when Δ_{str} is small, the task is only weakly dependent on structure, this lower bound alone does not imply a substantial performance gap.

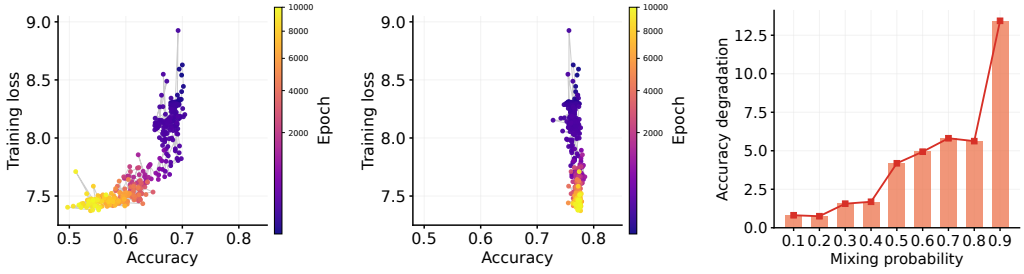
3.3 Empirical Investigations

Observation 1: Contrastive overfitting is prevalent and accompanied by topology-agnostic embedding separation. As shown in Figure 2a, we evaluate nine representative InfoNCE-based GCL methods. All methods exhibit clear degradation from their best intermediate representations to the final ones. Notably, this behavior persists across methods designed to address different limitations of GCL, suggesting that improving the attainable representation quality does not by itself eliminate the late-stage mismatch between contrastive optimization and downstream performance.

Meanwhile, Figure 2b shows the distribution of pairwise node similarities, revealing that the initially heterogeneous similarities become sharply concentrated with continued optimization. This evolution is qualitatively consistent with the instance-separation tendency characterized by Theorem 1: distinctions among pairwise relations are progressively flattened as optimization proceeds.

Table 1: Contrastive overfitting and empirical structural dependence.

Method	Cora	CiteSeer	PubMed	Am-Photo	Am-Computer	Wiki-CS
$\text{GRACE}_{\text{Best}}$	$82.48_{\pm 0.15}$	$72.48_{\pm 0.40}$	$80.78_{\pm 0.10}$	$93.24_{\pm 0.18}$	$88.64_{\pm 0.06}$	$79.64_{\pm 0.04}$
$\text{GRACE}_{\text{Final}}$	$63.66_{\pm 0.19}$	$50.82_{\pm 0.34}$	$62.33_{\pm 0.13}$	$89.84_{\pm 0.01}$	$87.43_{\pm 0.03}$	$78.84_{\pm 0.10}$
Accuracy Degradation	18.82(↓)	21.66(↓)	18.45(↓)	3.40(↓)	1.21(↓)	0.80(↓)
$\text{Supervised}_{\text{MLP}}$	$51.05_{\pm 0.85}$	$50.30_{\pm 1.17}$	$64.90_{\pm 0.59}$	$87.39_{\pm 0.47}$	$80.84_{\pm 0.22}$	$74.72_{\pm 0.11}$
$\text{Supervised}_{\text{GCN}}$	$80.69_{\pm 0.70}$	$70.41_{\pm 0.83}$	$79.41_{\pm 0.25}$	$92.96_{\pm 0.10}$	$88.33_{\pm 0.19}$	$80.38_{\pm 0.16}$
Δ_{str} (Proxy)	29.64(↑)	20.11(↑)	14.51(↑)	5.57(↑)	7.49(↑)	5.66(↑)



(a) sbm-community (b) sbm-feature (c) sbm-mix

Figure 3: Contrastive overfitting across synthetic tasks with different levels of structural dependence.

Observation 2: Structural dependence determines the downstream impact of contrastive overfitting. To isolate the role of task structure dependence, we construct a synthetic SBM graph with four communities and node features sampled from four Gaussian components. On the same graph and features, we define two tasks that differ only in their labels: **sbm-community**, whose labels depend on community structure, and **sbm-feature**, whose labels depend on node attributes (Gaussian components); see Figure B.3 for topology visualization. As shown in Figures 3a and 3b, **sbm-community** exhibits clear contrastive overfitting: downstream accuracy decreases as the contrastive loss is optimized, whereas **sbm-feature** maintains nearly constant accuracy. We further construct mixed tasks by sampling each node label from **sbm-feature** or **sbm-community**, with the mixing probability controlling the degree of structural dependence. As shown in Figure 3c, stronger structural dependence leads to larger accuracy degradation. These results are consistent with the prediction of Theorem 2 that contrastive overfitting primarily harms structure-dependent downstream tasks.

We further examine whether the same pattern holds on real-world datasets. Since Δ_{str} in Definition 1 is not directly observable, we use the supervised performance gap between a GCN and its topology-free MLP counterpart as a proxy for structural dependence. As shown in Table 1, datasets with severe contrastive overfitting (Cora, CiteSeer, and PubMed) show large GCN–MLP gaps, whereas those with mild overfitting (Am-Photo, Am-Computer, and Wiki-CS) show much smaller gaps. This further supports the link between structural dependence and contrastive overfitting.

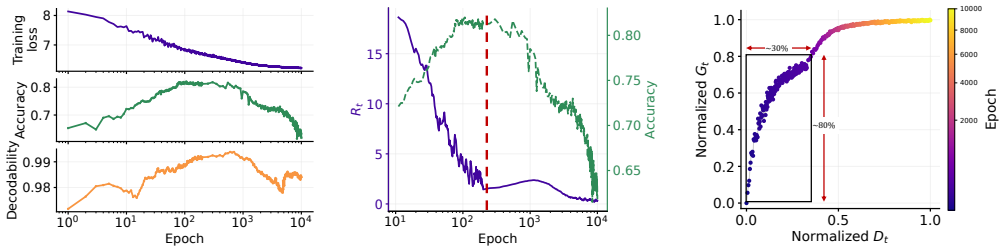
Observation 3: Contrastive optimization transitions from structural learning to structural degeneration. Since there is no direct metric for quantifying how much structural information is encoded into node representations, we use *structural decodability* as an empirical proxy. We treat graph edges as positive pairs and randomly sample an equal number of non-edges as negative pairs. We then train a low-rank bilinear structural probe that assigns each node pair (v_i, v_j) a score

$$s_{ij} = \hat{z}_i^\top \mathbf{P}\,\mathrm{diag}(\mathbf{a})\mathbf{P}^\top \hat{z}_j + b, \tag{4}$$

where $\hat{z}_i$ and $\hat{z}_j$ denote the normalized node representations, $\mathbf{P}$ is a low-rank projection matrix, and $\mathbf{a}$ contains learnable weights for the projected dimensions. The probe is evaluated on unseen node pairs using ROC-AUC. A higher AUC indicates that graph adjacency can be more accurately recovered from the learned representations, suggesting stronger preservation of structural information.

As shown in Figure 4a, structural decodability and downstream accuracy exhibit highly consistent dynamics: both increase during early training and progressively decrease under continued contrastive optimization, while the contrastive loss keeps decreasing. This suggests that GCL initially learns useful structural representations, but excessive optimization gradually degrades the encoded structural information, accompanying the observed downstream degradation.

We further investigate how the efficiency of contrastive optimization evolves during training. Let $\mathcal{L}_t$



(a) Training dynamics (b) Marginal return and accuracy (c) Cumulative trajectory

Figure 4: Optimization dynamics underlying contrastive overfitting on Cora.

and $\mathbf{Z}_t$ (obtained by encoding the original graph) denote the contrastive loss and node representations at iteration t. We define the *contrastive gain* and *representation drift* as

$$B_t := [\mathcal{L}_{t-1} - \mathcal{L}_t]_+, \quad C_t := \frac{\|\mathbf{Z}_t - \mathbf{Z}_{t-1}\|_F}{\sqrt{N}}, \tag{5}$$

B_t measures how much contrastive loss is reduced, while C_t measures how far the node representations move to achieve this reduction. We then define the *marginal contrastive return* as $R_t := B_t/C_t$, which measures the contrastive improvement obtained per unit representation drift. A large R_t therefore indicates an efficient optimization stage, whereas a small R_t indicates that further representation movement yields only marginal contrastive improvement. We further define the cumulative contrastive gain and representation drift as $G_t := \mathcal{L}_0 - \mathcal{L}_t$, and $D_t := \sum_{k=1}^{t} C_k$. As shown in Figure 4b, the optimization process exhibits two distinct regimes. In the early high-return regime, a small amount of representation movement produces substantial contrastive gain, corresponding to a large R_t. As training proceeds, R_t rapidly decreases and remains low, indicating that continued representation drift produces increasingly marginal contrastive improvement. The cumulative trajectory (D_t, G_t) in Figure 4c exhibits the same pattern. We also visualize representation drift trajectories during training in Appendix B.4. The embeddings quickly become class-separated at first, but excessive optimization later pulls them back together and weakens the learned class structure. It also explains the nonlinear relationship between representation drift and loss reduction shown in Figures 4b and 4c. Beyond InfoNCE-based objectives, we explore representative non-InfoNCE GCL methods in Appendix B.3 to assess whether our observations generalize across different contrastive objectives.

4 PARE: Progress–Adaptive Regulation

Our investigation in Section 3.3 reveals that continued GCL optimization eventually enters a low-return regime, where representation drift yields little additional contrastive gain while structural information continues to degrade. This observation motivates a simple principle: preserve encoder updates while contrastive optimization remains both effective and productive, and progressively suppress them as the optimization progress diminishes. Based on this principle, we propose a simple yet effective Progress-Adaptive REgulation (PARE) method, which dynamically regulates encoder updates according to the evolving contrastive training dynamics.

Contrastive progress estimation. We use the contrastive gain B_t, representation drift C_t, and marginal contrastive return R_t defined in Section 3.3. To reduce iteration-level fluctuations, we estimate their local trends using exponential moving averages (EMA):

$$\bar{B}_t := \mu \bar{B}_{t-1} + (1-\mu) B_t, \quad \bar{C}_t := \mu \bar{C}_{t-1} + (1-\mu) C_t, \tag{6}$$

where $\mu \in [0, 1)$ is the smoothing coefficient. The smoothed marginal contrastive return is then

$$\bar{R}_t := \bar{B}_t / (\bar{C}_t + \epsilon), \tag{7}$$

where $\epsilon > 0$ is a sufficiently small positive constant used to ensure numerical stability.

Progress-adaptive regulation. We formulate the choice of encoder update strength at each iteration as a per-step regulation problem, where the update strength is adaptively determined by the current contrastive progress. Specifically, we regulate the encoder update according to both marginal return and contrastive gain. We define the encoder update strength as

$$w_t := \min(\bar{R}_t \times \bar{B}_t, 1), \quad w_t \in [0, 1]. \tag{8}$$

Here, $\bar{R}_t$ measures the recent contrastive improvement per unit representation drift, while $\bar{B}_t$ measures the magnitude of the recent improvement. Their product assigns a large update strength when the optimization remains both efficient and productive. The clipping operation ensures that PARE never amplifies the base update.

$$\theta_{t+1} = \theta_t - \eta w_t \nabla_\theta \mathcal{L}_{\text{con}}(\theta_t), \tag{9}$$

where η is the base learning rate. When $\bar{R}_t \bar{B}_t \geq 1$, PARE recovers the standard contrastive update. As either contrastive gain or marginal return decreases, their product decreases and PARE progressively suppresses further encoder movement. Beyond regulating individual updates, PARE also provides control over the entire representation trajectory.

Theorem 3 (Bounded Cumulative Representation Drift by PARE)**.** *Assume that: (i) $\mathcal{L}_{\text{con}}(\theta)$ is a fixed, β-smooth objective bounded below by F_{low}; (ii) $\theta \mapsto Z(\theta)/\sqrt{N}$ is K-Lipschitz for a fixed input. Let $F_t := \mathcal{L}_{\text{con}}(\theta_t)$. Consider the update in Equation (9) with $0 < \eta \leq 1/\beta$ and $B_t = [F_{t-1} - F_t]_+$. For any $s \geq 1$, if $\bar{R}_t \leq R_{\max} < \infty$ for all $t \geq s$, then*

$$\sum_{t=s+1}^{\infty} C_t \leq K\sqrt{2\eta R_{\max}}\left(F_s - F_{\text{low}} + \frac{\bar{B}_s}{2(1-\mu)}\right).$$

Theorem 3 provides a trajectory-level interpretation of PARE. Under standard smoothness conditions, the total future representation movement remains finite even when the number of optimization steps is unbounded. The bound becomes smaller when the remaining contrastive improvement, the residual EMA gain, or the marginal-return bound decreases. Thus, after optimization enters a low-gain and low-return regime, PARE limits the additional representation movement induced by continued training. This result directly supports the design goal of suppressing late-stage representation drift. Building on Theorem 3, Corollary 2 in Appendix C.4 further shows that PARE limits the subsequent loss of structural alignment and the increase in distance from structurally optimal representations.

5 Evaluation of PARE

In this section, we evaluate PARE along four dimensions: (i) its effectiveness in mitigating overfitting, (ii) its robustness to training configurations, (iii) whether its training behavior supports the proposed structural-preservation mechanism, and (iv) its component contributions.

Table 2: Node classification accuracy on Cora, CiteSeer and Am-Photo. **Best** and second best results are highlighted. Oracle results served as an optimistic reference and are excluded from ranking.

Dataset	Method	Oracle	Final	+ α-ReQ	+ RankMe	+ LiDAR	+ Contrast-Reg	+ PARE
Cora	GRACE	$82.48_{\pm 0.15}$	$63.66_{\pm 0.19}$	$63.24_{\pm 0.38}$	$63.66_{\pm 0.19}$	$63.24_{\pm 0.38}$	$70.48_{\pm 0.21}$	$\mathbf{82.43}_{\pm 0.09}$
	GCA	$82.30_{\pm 0.23}$	$67.70_{\pm 1.47}$	$67.70_{\pm 1.47}$	$68.85_{\pm 0.62}$	$67.41_{\pm 1.19}$	$72.32_{\pm 0.81}$	$\mathbf{82.21}_{\pm 0.14}$
	PiGCL	$81.95_{\pm 0.34}$	$44.68_{\pm 0.36}$	$49.83_{\pm 0.16}$	$44.68_{\pm 0.36}$	$56.85_{\pm 0.38}$	$69.76_{\pm 0.32}$	$\mathbf{74.19}_{\pm 0.34}$
	GRAPE	$83.53_{\pm 0.11}$	$63.84_{\pm 0.19}$	$64.56_{\pm 0.09}$	$81.59_{\pm 0.03}$	$81.59_{\pm 0.03}$	$70.32_{\pm 0.42}$	$\mathbf{83.46}_{\pm 0.31}$
	ProGCL	$82.36_{\pm 0.14}$	$72.63_{\pm 0.11}$	$77.60_{\pm 0.04}$	$72.86_{\pm 0.16}$	$73.02_{\pm 0.19}$	$78.31_{\pm 0.46}$	$\mathbf{81.63}_{\pm 0.72}$
	HomoGCL	$83.40_{\pm 0.14}$	$66.37_{\pm 0.22}$	$70.49_{\pm 0.33}$	$67.26_{\pm 0.16}$	$66.12_{\pm 0.15}$	$72.25_{\pm 0.08}$	$\mathbf{82.02}_{\pm 0.63}$
	GRACE+	$83.43_{\pm 0.09}$	$81.02_{\pm 0.19}$	$80.97_{\pm 0.06}$	$81.02_{\pm 0.19}$	$80.43_{\pm 0.11}$	$81.36_{\pm 0.27}$	$\mathbf{82.41}_{\pm 0.08}$
	ReGCL	$82.64_{\pm 0.40}$	$60.90_{\pm 0.00}$	$76.22_{\pm 0.27}$	$60.74_{\pm 0.14}$	$60.79_{\pm 0.07}$	$72.71_{\pm 0.50}$	$\mathbf{82.16}_{\pm 0.09}$
	GRACE-IS	$82.05_{\pm 0.20}$	$65.90_{\pm 0.36}$	$65.90_{\pm 0.36}$	$64.32_{\pm 0.91}$	$64.63_{\pm 0.77}$	$67.95_{\pm 0.26}$	$\mathbf{81.99}_{\pm 0.08}$
CiteSeer	GRACE	$72.48_{\pm 0.40}$	$50.82_{\pm 0.34}$	$50.82_{\pm 0.34}$	$50.82_{\pm 0.34}$	$50.87_{\pm 0.36}$	$61.28_{\pm 0.92}$	$\mathbf{72.38}_{\pm 0.07}$
	GCA	$72.80_{\pm 0.30}$	$45.20_{\pm 0.00}$	$45.63_{\pm 0.05}$	$45.20_{\pm 0.00}$	$44.36_{\pm 0.52}$	$57.34_{\pm 0.09}$	$\mathbf{72.65}_{\pm 0.44}$
	PiGCL	$71.57_{\pm 0.56}$	$51.68_{\pm 0.14}$	$50.22_{\pm 0.19}$	$70.08_{\pm 0.17}$	$49.55_{\pm 0.26}$	$59.92_{\pm 0.07}$	$\mathbf{71.45}_{\pm 0.37}$
	GRAPE	$73.03_{\pm 0.49}$	$51.95_{\pm 0.52}$	$53.47_{\pm 1.13}$	$71.57_{\pm 0.42}$	$51.80_{\pm 0.57}$	$63.22_{\pm 0.72}$	$\mathbf{73.03}_{\pm 0.51}$
	ProGCL	$71.22_{\pm 1.11}$	$67.48_{\pm 0.33}$	$69.09_{\pm 0.78}$	$69.09_{\pm 0.78}$	$69.45_{\pm 0.28}$	$68.75_{\pm 0.39}$	$\mathbf{70.26}_{\pm 0.73}$
	HomoGCL	$73.41_{\pm 0.57}$	$55.14_{\pm 1.04}$	$58.05_{\pm 1.06}$	$67.92_{\pm 0.53}$	$57.34_{\pm 0.86}$	$59.86_{\pm 0.75}$	$\mathbf{73.01}_{\pm 0.78}$
	GRACE+	$72.73_{\pm 0.25}$	$71.79_{\pm 0.52}$	$70.84_{\pm 0.23}$	$71.87_{\pm 0.42}$	$71.15_{\pm 1.20}$	$70.82_{\pm 0.67}$	$\mathbf{72.39}_{\pm 0.34}$
	ReGCL	$70.85_{\pm 0.13}$	$54.78_{\pm 0.04}$	$68.52_{\pm 0.47}$	$54.52_{\pm 0.06}$	$68.65_{\pm 0.85}$	$69.72_{\pm 0.80}$	$\mathbf{70.57}_{\pm 1.37}$
	GRACE-IS	$72.18_{\pm 0.72}$	$47.16_{\pm 0.09}$	$46.70_{\pm 0.00}$	$47.16_{\pm 0.09}$	$46.54_{\pm 0.12}$	$60.42_{\pm 0.15}$	$\mathbf{71.66}_{\pm 0.77}$
Am-Photo	GRACE	$93.24_{\pm 0.18}$	$89.84_{\pm 0.01}$	$91.54_{\pm 0.01}$	$89.80_{\pm 0.03}$	$89.63_{\pm 0.02}$	$92.55_{\pm 0.00}$	$\mathbf{93.20}_{\pm 0.00}$
	GCA	$91.78_{\pm 0.01}$	$86.99_{\pm 0.05}$	$88.57_{\pm 0.01}$	$90.21_{\pm 0.30}$	$88.81_{\pm 0.01}$	$91.10_{\pm 0.02}$	$\mathbf{91.32}_{\pm 0.00}$
	PiGCL	$92.86_{\pm 0.00}$	$90.02_{\pm 0.25}$	$91.80_{\pm 0.00}$	$90.02_{\pm 0.25}$	$90.02_{\pm 0.25}$	$91.15_{\pm 0.15}$	$\mathbf{92.75}_{\pm 0.15}$
	GRAPE	$92.60_{\pm 0.02}$	$91.36_{\pm 0.01}$	$92.23_{\pm 0.09}$	$92.21_{\pm 0.01}$	$91.50_{\pm 0.02}$	$92.32_{\pm 0.04}$	$\mathbf{92.39}_{\pm 0.01}$
	ProGCL	$93.04_{\pm 0.01}$	$92.05_{\pm 0.02}$	$90.52_{\pm 0.01}$	$91.97_{\pm 0.01}$	$91.97_{\pm 0.01}$	$92.62_{\pm 0.00}$	$\mathbf{92.96}_{\pm 0.00}$
	HomoGCL	$93.35_{\pm 0.01}$	$93.00_{\pm 0.02}$	$92.37_{\pm 0.01}$	$92.60_{\pm 0.05}$	$92.53_{\pm 0.16}$	$92.52_{\pm 0.03}$	$\mathbf{93.10}_{\pm 0.31}$
	GRACE+	$92.55_{\pm 0.01}$	$91.39_{\pm 0.01}$	$91.52_{\pm 0.01}$	$91.64_{\pm 0.01}$	$91.49_{\pm 0.02}$	$91.01_{\pm 0.08}$	$\mathbf{91.73}_{\pm 0.02}$
	ReGCL	$93.27_{\pm 0.03}$	$92.30_{\pm 0.01}$	$92.33_{\pm 0.01}$	$92.33_{\pm 0.01}$	$92.39_{\pm 0.05}$	$92.54_{\pm 0.02}$	$\mathbf{93.03}_{\pm 0.07}$
	GRACE-IS	$93.14_{\pm 0.00}$	$89.45_{\pm 0.03}$	$92.46_{\pm 0.12}$	$91.42_{\pm 0.02}$	$90.18_{\pm 0.03}$	$92.09_{\pm 0.02}$	$\mathbf{92.86}_{\pm 0.01}$

Table 3: Node classification accuracy on PubMed, Am-Computer and Wiki-CS. **Best** and second best results are highlighted. Oracle results served as an optimistic reference excluded from ranking.

Dataset	Method	Oracle	Final	+ α-ReQ	+ RankMe	+ LiDAR	+ Contrast-Reg	+ PARE
PubMed	GRACE	80.78$_{\pm 0.10}$	62.33$_{\pm 0.13}$	62.44$_{\pm 0.09}$	61.72$_{\pm 0.10}$	62.34$_{\pm 0.08}$	72.46$_{\pm 0.11}$	**78.19**$_{\pm 0.19}$
	GCA	80.42$_{\pm 0.34}$	57.29$_{\pm 0.16}$	56.14$_{\pm 0.41}$	57.29$_{\pm 0.16}$	57.29$_{\pm 0.16}$	65.26$_{\pm 0.22}$	**77.67**$_{\pm 0.11}$
	PiGCL	80.11$_{\pm 0.13}$	70.45$_{\pm 0.32}$	70.45$_{\pm 0.32}$	69.36$_{\pm 0.11}$	68.09$_{\pm 0.34}$	69.94$_{\pm 0.20}$	**78.10**$_{\pm 0.00}$
	GRAPE	80.48$_{\pm 0.04}$	62.89$_{\pm 0.27}$	62.89$_{\pm 0.27}$	62.89$_{\pm 0.27}$	61.90$_{\pm 0.92}$	69.16$_{\pm 0.14}$	**79.71**$_{\pm 0.11}$
	ProGCL	82.85$_{\pm 0.11}$	71.09$_{\pm 0.05}$	81.47$_{\pm 0.15}$	71.41$_{\pm 0.25}$	70.96$_{\pm 0.20}$	80.34$_{\pm 0.28}$	**82.13**$_{\pm 0.08}$
	HomoGCL	80.73$_{\pm 0.16}$	64.68$_{\pm 0.22}$	64.68$_{\pm 0.22}$	64.68$_{\pm 0.22}$	64.84$_{\pm 0.21}$	74.18$_{\pm 0.43}$	**79.63**$_{\pm 0.23}$
	ReGCL	81.59$_{\pm 0.05}$	62.95$_{\pm 0.05}$	62.95$_{\pm 0.05}$	63.47$_{\pm 0.06}$	63.13$_{\pm 0.06}$	66.41$_{\pm 0.12}$	**79.89**$_{\pm 0.02}$
	GRACE-IS	82.15$_{\pm 0.17}$	62.95$_{\pm 0.17}$	62.95$_{\pm 0.17}$	62.95$_{\pm 0.17}$	63.72$_{\pm 0.40}$	67.48$_{\pm 0.13}$	**81.17**$_{\pm 0.05}$
Am-Computer	GRACE	88.64$_{\pm 0.06}$	87.43$_{\pm 0.03}$	84.03$_{\pm 0.04}$	86.92$_{\pm 0.03}$	87.37$_{\pm 0.13}$	81.66$_{\pm 0.58}$	**88.07**$_{\pm 0.01}$
	GCA	88.59$_{\pm 0.01}$	85.96$_{\pm 0.03}$	84.90$_{\pm 0.01}$	86.94$_{\pm 0.03}$	87.27$_{\pm 0.01}$	88.40$_{\pm 0.01}$	**88.44**$_{\pm 0.00}$
	PiGCL	88.55$_{\pm 0.03}$	88.12$_{\pm 0.02}$	83.93$_{\pm 0.04}$	**88.45**$_{\pm 0.02}$	88.27$_{\pm 0.01}$	86.98$_{\pm 0.01}$	88.40$_{\pm 0.02}$
	GRAPE	88.18$_{\pm 0.01}$	87.26$_{\pm 0.01}$	83.63$_{\pm 0.06}$	87.69$_{\pm 0.01}$	**87.76**$_{\pm 0.00}$	87.64$_{\pm 0.01}$	87.53$_{\pm 0.06}$
	ProGCL	88.91$_{\pm 0.04}$	88.18$_{\pm 0.01}$	83.33$_{\pm 0.15}$	88.47$_{\pm 0.01}$	88.43$_{\pm 0.04}$	87.57$_{\pm 0.01}$	**88.75**$_{\pm 0.01}$
	HomoGCL	88.54$_{\pm 0.02}$	87.14$_{\pm 0.21}$	84.00$_{\pm 0.02}$	86.60$_{\pm 0.01}$	87.11$_{\pm 0.12}$	87.74$_{\pm 0.11}$	**87.93**$_{\pm 0.00}$
	ReGCL	88.02$_{\pm 0.01}$	83.50$_{\pm 0.04}$	86.70$_{\pm 0.01}$	87.39$_{\pm 0.04}$	86.85$_{\pm 0.00}$	86.89$_{\pm 0.10}$	**87.90**$_{\pm 0.01}$
	GRACE-IS	88.47$_{\pm 0.02}$	88.12$_{\pm 0.06}$	84.17$_{\pm 0.01}$	**88.31**$_{\pm 0.02}$	87.94$_{\pm 0.02}$	87.69$_{\pm 0.01}$	88.21$_{\pm 0.01}$
Wiki-CS	GRACE	79.64$_{\pm 0.04}$	78.84$_{\pm 0.10}$	78.87$_{\pm 0.13}$	78.93$_{\pm 0.23}$	78.79$_{\pm 0.08}$	79.47$_{\pm 0.22}$	**79.57**$_{\pm 0.30}$
	GCA	79.77$_{\pm 0.04}$	77.84$_{\pm 0.39}$	78.45$_{\pm 0.09}$	78.17$_{\pm 0.17}$	78.14$_{\pm 0.07}$	**79.35**$_{\pm 0.05}$	78.98$_{\pm 0.01}$
	PiGCL	80.12$_{\pm 0.19}$	78.93$_{\pm 0.28}$	79.09$_{\pm 0.15}$	79.15$_{\pm 0.11}$	78.74$_{\pm 0.31}$	77.79$_{\pm 0.10}$	**79.65**$_{\pm 0.05}$
	GRAPE	80.40$_{\pm 0.03}$	78.99$_{\pm 0.13}$	78.93$_{\pm 0.06}$	78.79$_{\pm 0.04}$	78.96$_{\pm 0.07}$	**79.81**$_{\pm 0.06}$	78.91$_{\pm 0.02}$
	ProGCL	80.90$_{\pm 0.18}$	80.02$_{\pm 0.11}$	80.56$_{\pm 0.18}$	80.02$_{\pm 0.04}$	79.92$_{\pm 0.06}$	80.41$_{\pm 0.11}$	**80.62**$_{\pm 0.03}$
	HomoGCL	80.66$_{\pm 0.04}$	75.13$_{\pm 0.89}$	**80.66**$_{\pm 0.04}$	79.26$_{\pm 0.34}$	75.14$_{\pm 0.71}$	79.16$_{\pm 0.09}$	79.58$_{\pm 0.07}$
	ReGCL	79.87$_{\pm 0.06}$	79.49$_{\pm 0.04}$	79.42$_{\pm 0.05}$	79.43$_{\pm 0.04}$	79.35$_{\pm 0.03}$	78.62$_{\pm 0.07}$	**79.60**$_{\pm 0.03}$
	GRACE-IS	80.33$_{\pm 0.24}$	79.76$_{\pm 0.16}$	79.72$_{\pm 0.15}$	79.42$_{\pm 0.22}$	79.85$_{\pm 0.09}$	79.28$_{\pm 0.07}$	**79.87**$_{\pm 0.02}$

* GRACE+ runs out of memory on these datasets and is therefore omitted.

Datasets and methods. We evaluate PARE on six benchmark datasets: Cora, CiteSeer, PubMed (Yang et al., 2016), Am-Photo, Am-Computer (Shchur et al., 2018), and Wiki-CS (Mernyei & Cangea, 2020). We compare PARE with three unsupervised representation selection methods: α-ReQ (Agrawal et al., 2022), RankMe (Garrido et al., 2023), and LiDAR (Thilak et al., 2024), and one GCL regularizer Contrast-Reg (Ma et al., 2023). Additional comparisons with fixed learning-rate decay schedules and projection-head variants are provided in Appendices B.1 and B.2, respectively. We integrate these methods and PARE into nine representative InfoNCE-based GCL variants: GRACE (Zhu et al., 2020), GCA (Zhu et al., 2021b), PiGCL (He et al., 2024), ReGCL (Ji et al., 2024), ProGCL (Xia et al., 2022), GRACE+ (Chi & Ma, 2024), HomoGCL (Li et al., 2023), GRAPE (Hao et al., 2024), and GRACE-IS (Liu et al., 2024).

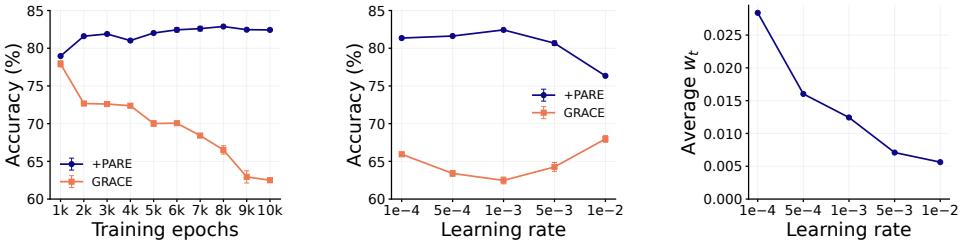
Evaluation protocol. For each backbone–dataset pair, the standard backbone and its variants share the same hyperparameter configuration and are trained for 10,000 epochs, with their final checkpoints (or selected checkpoints) used for evaluation. We evaluate frozen embeddings using logistic regression and report the mean test accuracy and standard deviation over 10 random seeds. We also report the **Oracle** accuracies, obtained from the best downstream test performance, as a reference for peak performance, and the **Final** accuracies, obtained at the end of standard training, as the overfitted performance. Detailed experimental settings and evaluation protocols can be found in Appendix D.

5.1 Main Results

Overfitting mitigation. In Tables 2 and 3, PARE ranks first among the compared methods in 45 of 51 settings. On Cora, CiteSeer, and PubMed, it ranks first in all 26 settings and outperforms the strongest competing method for each backbone by 6.92, 5.31, and 8.70 percentage points on average, respectively. Its average gains over the unregulated final checkpoints are 16.20, 16.82, and 15.23 points. In contrast, the improvements are smaller on Am-Photo, Am-Computer, and Wiki-CS, where the final–oracle gaps are also smaller. This pattern is consistent with our conclusion in Theorem 2. In Table B.1, PARE consistently outperforms fixed learning-rate decay schedules, while avoiding the need for a predefined decay rule. As shown in Table B.2, the projection head is insufficient to prevent overfitting, and PARE is complementary to the projection-head design and can be seamlessly combined with it to further mitigate contrastive overfitting.

Robustness to hyperparameters. We examine the sensitivity of PARE to the training horizon and base learning rate. As shown in Figure 5a, PARE reaches a stable accuracy level after the initial training stage and remains effective throughout prolonged training, whereas standard GRACE progressively degrades. Figure 5b shows that PARE consistently outperforms standard GRACE

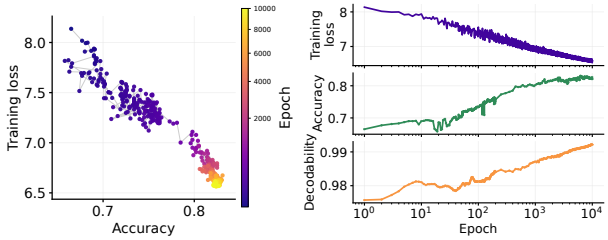
across base learning rates from 10^{-4} to 10^{-2}. Figure 5c further shows that larger base learning rates lead to smaller average regulation weights. This adaptive adjustment of w_t helps explain the robustness observed in the first two panels: it compensates for stronger base updates and suppresses low-return late-stage updates, thereby preventing excessive representation drift.



(a) Training horizon (b) Base learning rate (c) PARE update weight

Figure 5: Robustness to training and adaptive regulation on the Cora dataset. (a) Final accuracy across training horizons with a fixed base learning rate of 10^{-3}. (b) Final accuracy across base learning rates after 10,000 training epochs. (c) Average regulation weight w_t across base learning rates.

Behavior analysis. Figure 6 shows the optimization process of GRACE equipped with PARE on Cora. Compared with standard GRACE in Figure 1, the pronounced late-stage decline in downstream accuracy is largely suppressed, even as the contrastive loss continues to decrease. Structural decodability and downstream accuracy also exhibit aligned temporal trajectories: both improve during early training and remain high in the later stage.



(a) Loss-accuracy trajectory (b) Training dynamics

Figure 6: Training process of GRACE with PARE on Cora.

Component ablation. As shown in Figure 7, the complete PARE formulation achieves the highest accuracy on both Cora and CiteSeer. Removing EMA causes a relatively small performance decrease, supporting the usefulness of temporal smoothing for estimating local optimization progress. Using only the marginal-return term $\bar{R}_t$ (Equation (7)) results in a larger decrease, whereas using only the contrastive-gain term $\bar{B}_t$ (Equation (6)) produces the largest degradation. These results suggest that the two components capture complementary signals and that their combination is essential for achieving better performance.

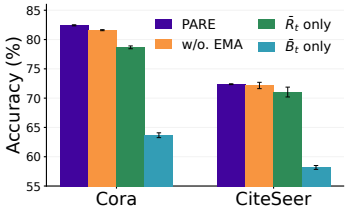


Figure 7: Component ablation.

6 Conclusion

In this paper, we investigate contrastive overfitting in graph contrastive learning, where continued optimization of the contrastive objective degrades downstream performance despite a decreasing training loss. We show that this phenomenon is closely related to the topology-agnostic nature of instance-level contrastive objectives. While GNN encoders initially exploit graph structure, excessive contrastive optimization can gradually drive representations toward a topology-independent geometry and weaken useful structural information. Our theoretical analysis characterizes this topology-independent optimum and explains why such degradation is particularly harmful to structure-dependent downstream tasks. We further identify a low-marginal-return regime, where continued representation drift yields little additional contrastive improvement. Based on this observation, we propose PARE, a progress-adaptive regulation framework that adjusts encoder updates according to the marginal return and gain of contrastive optimization. Experiments across multiple datasets and GCL methods show that PARE effectively mitigates contrastive overfitting.

AI USE STATEMENT

We used ChatGPT to assist with language editing, including improving grammar, clarity, and readability of the manuscript. We also used ChatGPT to help identify potential logical inconsistencies and reasoning errors in drafts written by the authors. In addition, ChatGPT was used to assist in writing and refining code for result visualization. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The anonymous code is available at: https://anonymous.4open.science/r/pare.

REFERENCES

Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-ReQ: Assessing Representation Quality in Self-Supervised Learning by Measuring Eigenspectrum Decay. In *Advances in Neural Information Processing Systems*, 2022.

Han Bao, Yoshihiro Nagano, and Kento Nozawa. On the surrogate gap between contrastive and supervised losses. In *International conference on machine learning*, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 2020.

Hongliang Chi and Yao Ma. Enhancing contrastive learning on graphs with node similarity. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in neural information processing systems*, 2020.

Zhezheng Hao, Haonan Xin, Long Wei, Liaoyuan Tang, Rong Wang, and Feiping Nie. Towards expansive and adaptive hard negative mining: Graph contrastive learning via subspace preserving. In *Proceedings of the ACM Web Conference 2024*, 2024.

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International conference on machine learning*, pp. 4116–4126, 2020.

Dongxiao He, Jitao Zhao, Cuiying Huo, Yongqi Huang, Yuxiao Huang, and Zhiyong Feng. A new mechanism for eliminating implicit conflict in graph contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Cheng Ji, Zixuan Huang, Qingyun Sun, Hao Peng, Xingcheng Fu, Qian Li, and Jianxin Li. Regcl: rethinking message passing in graph contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Wei Ju, Yifan Wang, Yifang Qin, Zhengyang Mao, Zhiping Xiao, Junyu Luo, Junwei Yang, Yiyang Gu, Dongjie Wang, Qingqing Long, et al. Towards graph contrastive learning: A survey and beyond. *arXiv preprint arXiv:2405.11868*, 2024.

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

Wen-Zhi Li, Chang-Dong Wang, Hui Xiong, and Jian-Huang Lai. Homogcl: Rethinking homophily in graph contrastive learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

Jingyu Liu, Huayi Tang, and Yong Liu. Perfect alignment may be poisonous to graph contrastive learning. In *Forty-first International Conference on Machine Learning*, 2024.

Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and S Yu Philip. Graph self-supervised learning: A survey. *IEEE transactions on knowledge and data engineering*, 35(6): 5879–5900, 2022.

Kaili Ma, Garry Yang, Han Yang, Yongqiang Chen, and James Cheng. Calibrating and improving graph contrastive learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Péter Mernyei and Cătălina Cangea. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Zachary Rabin, Jim Davis, Benjamin Lewis, and Matthew Scherreik. Overfitting in contrastive learning? *arXiv preprint arXiv:2407.15863*, 2024.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Remi Munos, Petar Veličković, and Michal Valko. Bootstrapped representation learning on graphs. In *ICLR 2021 Workshop on Geometrical and Topological Representation Learning*, 2021.

Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Joshua M. Susskind, and Etai Littwin. LiDAR: Sensing linear probing performance in joint embedding SSL architectures. In *The Twelfth International Conference on Learning Representations*, 2024.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Jun Xia, Lirong Wu, Ge Wang, Jintao Chen, and Stan Z Li. Progcl: Rethinking hard negative mining in graph contrastive learning. In *Proceedings of the International Conference on Machine Learning*, 2022.

Rong Yan, Peng Bao, Xiao Zhang, Zhongyi Liu, and Hui Liu. Towards alignment-uniformity aware representation in graph contrastive learning. In *Proceedings of the 17th ACM international conference on web search and data mining*, 2024.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, 2016.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

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

Hengrui Zhang, Qitian Wu, Yu Wang, Shaofeng Zhang, Junchi Yan, and Philip S Yu. Localized contrastive learning on graphs. *arXiv preprint arXiv:2212.04604*, 2022.

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

Yanqiao Zhu, Yichen Xu, Qiang Liu, and Shu Wu. An empirical study of graph contrastive learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021a.

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

Appendix

Table of Contents

A Related Work **14**
- A.1 Graph Contrastive Learning . . . 14
- A.2 Recent Advances Related to Graph Contrastive Overfitting . . . 14

B Additional Experiments **15**
- B.1 Comparison with Learning-Rate Schedules . . . 15
- B.2 Impact of Projection Head . . . 15
- B.3 GCL Methods with Non-InfoNCE Objectives . . . 16
- B.4 Visualizations . . . 17

C Theoretical Analysis **18**
- C.1 Proof of Theorem 1 . . . 18
- C.2 Proof of Theorem 2 . . . 20
- C.3 Proof of Theorem 3 . . . 20
- C.4 Structural Degradation and Preservation . . . 21

D Experimental Details **23**
- D.1 Baselines . . . 23
- D.2 Datasets . . . 24
- D.3 Evaluation Protocol . . . 24
- D.4 Discussion of PARE . . . 25

E Implementation Details **26**
- E.1 Implementation of PARE . . . 26
- E.2 Hyperparameters . . . 27

A Related Work

A.1 Graph Contrastive Learning

Graph contrastive learning (GCL) has become a prominent paradigm for self-supervised graph representation learning, aiming to learn informative representations without relying on task-specific labels (Ju et al., 2024). In general, existing GCL methods differ mainly in how they construct graph views, define contrastive pairs, and formulate the optimization objective. For node-level representation learning, contrast can be performed either across different representation scales or directly between node representations. Early representative methods mainly adopt local–global contrast. For example, Deep Graph Infomax (DGI) (Veličković et al., 2019) maximizes the mutual information between node representations and a global graph summary, while MVGRL (Hassani & Khasahmadi, 2020) extends this idea by constructing multiple structural views based on the original graph and graph diffusion and contrasting node- and graph-level representations across views. These methods established mutual-information maximization as an important principle for self-supervised node representation learning.

A major line of subsequent node-level GCL methods adopts local-level contrast, where different augmented views of the same graph are encoded by a shared GNN and node representations are directly contrasted across views. Among them, GRACE (Zhu et al., 2020) is a representative framework based on the InfoNCE objective (Oord et al., 2018). GRACE generates two graph views through random edge dropping and feature masking, treats representations of the same node across the two views as a positive pair, and regards representations of other nodes as negatives. This instance-discrimination formulation has subsequently become a widely adopted paradigm for node-level GCL. Building upon this framework, GCA (Zhu et al., 2021b) introduces adaptive graph augmentations that assign different perturbation probabilities to edges and feature dimensions according to their importance. Subsequent studies further extend InfoNCE-based GCL from different perspectives. For example, Localized GCL (Zhang et al., 2022) constructs positive pairs from local neighborhoods and introduces a kernelized contrastive objective, while ProGCL (Xia et al., 2022) improves negative sampling by explicitly accounting for false negatives. Collectively, these methods refine graph augmentation, contrastive pair construction, and sampling strategies to better accommodate the relational characteristics of graph data.

Beyond negative-sample-based contrastive objectives, negative-free approaches have also been explored for node-level graph representation learning. BGRL (Thakoor et al., 2021), for example, adapts the bootstrapping principle of BYOL (Grill et al., 2020) to graphs and learns node representations by matching an online encoder with a slowly updated target encoder across augmented graph views. Similarly, CCA-SSG (Zhang et al., 2021) shares the redundancy-reduction principle of Barlow Twins (Zbontar et al., 2021), while deriving its objective from canonical correlation analysis. Despite these alternative formulations, InfoNCE-based methods remain one of the most widely studied families of node-level GCL and form the primary setting considered in this work.

A.2 Recent Advances Related to Graph Contrastive Overfitting

Several recent studies have identified limitations of InfoNCE-based GCL that are closely related to the *contrastive overfitting* phenomenon studied in this work. One line of research focuses on the construction of positive and negative pairs. ProGCL (Xia et al., 2022) observes that similarity-based hard-negative mining can preferentially select false negatives due to graph message passing, and estimates the probability of a negative sample being a true negative for more reliable hard-negative mining. HomoGCL (Li et al., 2023) exploits graph homophily to expand the positive set with neighboring nodes, while GRAPE (Hao et al., 2024) adopts a subspace-preserving perspective to perform expansive and adaptive hard-negative mining. GRACE+ (Chi & Ma, 2024) further constructs positive and negative sampling distributions according to node similarity to alleviate limited positive sampling and false negatives. These methods mainly improve the quality of contrastive pairs and the attainable representation quality of GCL. Another line of work investigates conflicts or discrepancies induced by the contrastive objective itself. PiGCL (He et al., 2024) identifies implicit conflicts between GNN message passing and InfoNCE-based instance discrimination and suppresses conflicting contrastive signals during optimization. ReGCL (Ji et al., 2024) further studies such conflicts from a gradient perspective and mitigates them through gradient-guided structure learning and a reweighted InfoNCE objective. Closely related studies have also questioned whether improved

contrastive optimization necessarily leads to better downstream representations. GRACE-IS (Liu et al., 2024) shows that excessive alignment of positive pairs can impair generalization, while Contrast-Reg (Ma et al., 2023) identifies discrepancies between contrastive-pair optimization and downstream performance and introduces a regularization strategy to improve their consistency.

Overall, these studies primarily address factors that limit the best achievable representation quality of GCL, rather than the overfitting regime considered here, where continued contrastive optimization after high-quality representations have been learned progressively degrades downstream performance. Moreover, under the converged-training setting considered in this work, experimental results in Section 5 show that these methods remain susceptible to contrastive overfitting.

Beyond graph learning, prior studies have examined the discrepancy between contrastive objectives and downstream representation quality: Bao et al. (2022) theoretically characterize the surrogate gap between contrastive and supervised losses, while Rabin et al. (2024) investigate overfitting in conventional contrastive learning. Our work is also related to unsupervised representation selection in self-supervised learning, where the goal is to assess representation quality without access to downstream labels. α-ReQ (Agrawal et al., 2022) characterizes representation quality through the power-law behavior of the feature covariance eigenspectrum, while RankMe (Garrido et al., 2023) employs the effective rank of learned representations as an unsupervised selection criterion. LiDAR (Thilak et al., 2024) further evaluates representations through the effective rank of the discriminative structure induced by a surrogate self-supervised task. These methods provide label-free criteria for checkpoint or hyperparameter selection, but do not directly explain or prevent the progressive representation degradation caused by excessive GCL optimization. To the best of our knowledge, contrastive overfitting has not been explicitly studied in graph contrastive learning, particularly from the perspective of progressive structural information degradation under continued contrastive optimization.

B Additional Experiments

B.1 Comparison with Learning-Rate Schedules

As shown in Table B.1, simply decaying the learning rate can partially mitigate contrastive overfitting on Cora and CiteSeer, but its effectiveness is inconsistent across datasets. In particular, both linear and cosine schedules substantially improve over constant learning rates on Cora and CiteSeer, while providing no benefit on PubMed. In contrast, PARE consistently achieves the best performance across all three datasets, outperforming the strongest fixed schedule by 14.38, 6.27, and 15.87 percentage points on Cora, CiteSeer, and PubMed, respectively. These results indicate that predetermined learning-rate decay alone is insufficient, and that adapting the update strength to the actual optimization progress is important for robustly mitigating contrastive overfitting. It is worth noting that the compared methods all require predefined scheduling parameters, which are difficult to determine in a strictly unsupervised training setting. In contrast, PARE introduces no additional tunable parameters and demonstrates strong adaptability across different original hyperparameter configurations.

Table B.1: Comparison of learning-rate decay schedules and progress-adaptive regulation.

Dataset	Constant LR	Linear LR	Cosine LR	PARE
Cora	$63.66_{\pm 0.19}$	$68.05_{\pm 0.31}$	$66.89_{\pm 0.38}$	$82.43_{\pm 0.09}$
CiteSeer	$50.82_{\pm 0.34}$	$66.11_{\pm 0.19}$	$64.45_{\pm 0.11}$	$72.38_{\pm 0.07}$
PubMed	$62.33_{\pm 0.13}$	$61.71_{\pm 0.42}$	$60.36_{\pm 1.61}$	$78.19_{\pm 0.19}$

B.2 Impact of Projection Head

Recent contrastive learning methods often append a projection head after the encoder to decouple the representation space used for downstream tasks from the space in which the contrastive objective is optimized (Chen et al., 2020). After training, the projection head is discarded, and the encoder output is directly used as the final representation. Since our theoretical analysis and experiments are built upon the general InfoNCE objective, we further conduct experiments to investigate the impact of the projection head on GCL. As shown in Table B.2, we additionally report results on

Table B.2: Node classification accuracy (%) of GCL methods with projection heads.

Method	Oracle†	Final	+ α-ReQ	+ RankMe	+ LiDAR	+ Contrast-Reg	+ PARE
Cora							
GRACE	$82.21_{\pm 0.05}$	$61.43_{\pm 0.09}$	$\underline{77.39}_{\pm 0.44}$	$64.91_{\pm 0.10}$	$70.10_{\pm 0.10}$	$66.52_{\pm 0.89}$	$\mathbf{79.94}_{\pm 0.23}$
GCA	$80.44_{\pm 0.46}$	$69.57_{\pm 0.46}$	$69.23_{\pm 0.45}$	$70.55_{\pm 0.42}$	$\underline{74.04}_{\pm 0.50}$	$70.32_{\pm 0.32}$	$\mathbf{79.70}_{\pm 0.12}$
PiGCL	$81.59_{\pm 0.11}$	$72.43_{\pm 0.16}$	$77.52_{\pm 0.12}$	$\underline{79.67}_{\pm 0.16}$	$73.09_{\pm 0.16}$	$77.21_{\pm 0.27}$	$\mathbf{80.98}_{\pm 0.10}$
GRAPE	$83.41_{\pm 0.26}$	$74.79_{\pm 0.30}$	$\underline{79.55}_{\pm 0.24}$	$76.35_{\pm 0.47}$	$\underline{79.55}_{\pm 0.24}$	$77.47_{\pm 0.06}$	$\mathbf{82.37}_{\pm 0.22}$
ProGCL	$82.05_{\pm 0.15}$	$70.55_{\pm 0.57}$	$72.93_{\pm 0.14}$	$73.26_{\pm 0.07}$	$\underline{74.93}_{\pm 0.05}$	$70.29_{\pm 0.03}$	$\mathbf{78.33}_{\pm 0.26}$
HomoGCL	$80.24_{\pm 0.43}$	$68.74_{\pm 0.33}$	$\underline{78.71}_{\pm 0.05}$	$73.92_{\pm 0.17}$	$69.25_{\pm 0.30}$	$71.74_{\pm 0.56}$	$\mathbf{79.38}_{\pm 0.19}$
GRACE+	$82.73_{\pm 0.28}$	$79.32_{\pm 0.15}$	$\underline{79.76}_{\pm 0.43}$	$79.25_{\pm 0.19}$	$79.55_{\pm 0.17}$	$78.65_{\pm 0.25}$	$\mathbf{81.97}_{\pm 0.33}$
ReGCL	$81.86_{\pm 0.31}$	$63.52_{\pm 0.55}$	$77.10_{\pm 0.59}$	$\mathbf{80.50}_{\pm 0.24}$	$64.41_{\pm 0.64}$	$76.49_{\pm 0.37}$	$\underline{79.24}_{\pm 0.32}$
GRACE-IS	$81.44_{\pm 0.68}$	$66.05_{\pm 0.38}$	$66.05_{\pm 0.38}$	$66.05_{\pm 0.38}$	$66.74_{\pm 0.28}$	$\underline{73.52}_{\pm 0.28}$	$\mathbf{79.31}_{\pm 0.40}$
CiteSeer							
GRACE	$70.39_{\pm 0.23}$	$66.15_{\pm 0.62}$	$68.10_{\pm 0.33}$	$66.15_{\pm 0.62}$	$68.10_{\pm 0.33}$	$\underline{68.69}_{\pm 0.42}$	$\mathbf{70.22}_{\pm 0.97}$
GCA	$71.71_{\pm 0.25}$	$66.88_{\pm 0.72}$	$\underline{67.25}_{\pm 0.79}$	$\underline{67.25}_{\pm 0.79}$	$67.10_{\pm 0.84}$	$66.69_{\pm 1.80}$	$\mathbf{71.42}_{\pm 0.75}$
PiGCL	$71.90_{\pm 0.19}$	$63.88_{\pm 0.21}$	$\underline{70.84}_{\pm 0.59}$	$\underline{70.84}_{\pm 0.59}$	$64.08_{\pm 0.12}$	$65.26_{\pm 0.24}$	$\mathbf{71.41}_{\pm 0.25}$
GRAPE	$69.99_{\pm 0.16}$	$59.36_{\pm 0.68}$	$\underline{68.65}_{\pm 0.20}$	$61.39_{\pm 0.31}$	$\underline{68.65}_{\pm 0.20}$	$65.84_{\pm 0.63}$	$\mathbf{69.49}_{\pm 0.15}$
ProGCL	$70.74_{\pm 0.28}$	$66.42_{\pm 0.76}$	$66.54_{\pm 0.63}$	$66.54_{\pm 0.63}$	$\underline{66.91}_{\pm 0.58}$	$66.77_{\pm 0.50}$	$\mathbf{69.82}_{\pm 0.39}$
HomoGCL	$68.51_{\pm 0.50}$	$62.49_{\pm 0.40}$	$63.88_{\pm 0.15}$	$63.88_{\pm 0.15}$	$63.88_{\pm 0.15}$	$\underline{65.72}_{\pm 1.16}$	$\mathbf{67.71}_{\pm 0.27}$
GRACE+	$72.26_{\pm 0.45}$	$70.75_{\pm 0.25}$	$71.14_{\pm 0.22}$	$\underline{71.81}_{\pm 0.22}$	$\underline{71.81}_{\pm 0.22}$	$71.13_{\pm 0.68}$	$\mathbf{71.93}_{\pm 0.13}$
ReGCL	$70.63_{\pm 0.26}$	$56.26_{\pm 0.80}$	$67.04_{\pm 0.21}$	$59.42_{\pm 1.19}$	$\mathbf{70.63}_{\pm 0.26}$	$64.80_{\pm 1.01}$	$\underline{67.58}_{\pm 1.39}$
GRACE-IS	$69.83_{\pm 0.37}$	$51.59_{\pm 0.23}$	$51.59_{\pm 0.23}$	$53.23_{\pm 0.22}$	$\underline{61.88}_{\pm 0.21}$	$60.02_{\pm 0.90}$	$\mathbf{68.86}_{\pm 0.49}$

† Oracle results are provided for reference only and are excluded from ranking.

the Cora and CiteSeer datasets, with a projection head added to each GCL method. We observe that GCL methods still exhibit varying degrees of overfitting after introducing a projection head, although the degradation is generally less severe than that reported in Table 2. This suggests that, although the projection head can alleviate contrastive overfitting to some extent, it cannot fully prevent the degradation of encoder representations under excessive optimization; meanwhile, PARE is complementary to the projection-head design and can be seamlessly combined with it to further mitigate such degradation.

B.3 GCL Methods with Non-InfoNCE Objectives

Beyond InfoNCE-based GCL methods, as an exploratory supplementary study, we further evaluate three representative GCL models that adopt non-InfoNCE Objectives and training paradigms: DGI (Veličković et al., 2019), CCA-SSG (Zhang et al., 2021), and BGRL (Thakoor et al., 2021). Their optimization objectives are summarized in Table B.3.

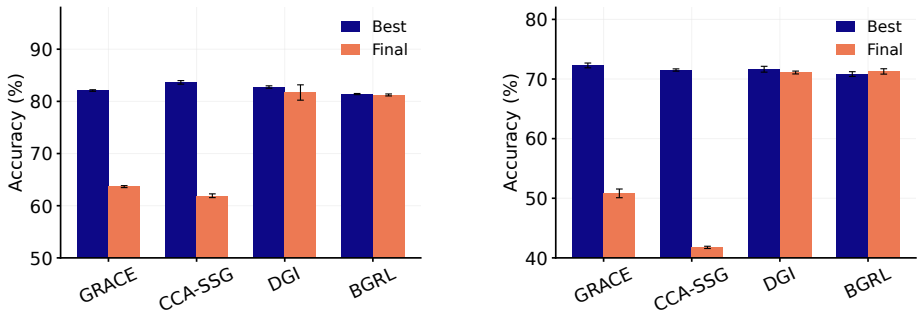
Table B.3: Different GCL objectives.

Method	contrastive objective	contrastive strategy
GRACE	$\frac{1}{2N}\sum_{i=1}^{N}\left[\mathcal{L}(z_i^1, z_i^2) + \mathcal{L}(z_i^2, z_i^1)\right]$	InfoNCE-based
CCA-SSG‡	$\|\mathbf{U}-\mathbf{V}\|^2 + \lambda\left(\left\|\mathbf{U}^\top\mathbf{U}-\mathbf{I}\right\|^2 + \left\|\mathbf{V}^\top\mathbf{V}-\mathbf{I}\right\|^2\right)$	CCA-based
DGI†	$-\frac{1}{N+M}\left[\sum_{i=1}^{N}\mathbb{E}_{\mathcal{G}}\log\mathcal{D}(\mathbf{h}_i,\mathbf{s}) + \sum_{j=1}^{M}\mathbb{E}_{\tilde{\mathcal{G}}}\log\left(1-\mathcal{D}(\tilde{\mathbf{h}}_j,\mathbf{s})\right)\right]$	InfoMax-based
BGRL*	$-\frac{1}{2N}\sum_{i=1}^{N}\left[\cos(\mathbf{p}_i^U, \mathrm{sg}(\mathbf{z}_i^V)) + \cos(\mathbf{p}_i^V, \mathrm{sg}(\mathbf{z}_i^U))\right]$	Bootstrap-based

† $\tilde{\mathcal{G}}$ is the corrupted graph of $\mathcal{G}$.
‡ $\mathbf{U}$ and $\mathbf{V}$ denote the node embeddings learned from two augmented graph views $\mathcal{G}_U$ and $\mathcal{G}_V$, respectively.
* $\mathbf{p}_i^U$ and $\mathbf{p}_i^V$ denote the online predictions of node i from the two graph views, while $\mathbf{z}_i^U$ and $\mathbf{z}_i^V$ denote the corresponding target representations. $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation, and the target encoder is updated as an exponential moving average of the online encoder.

Interestingly, as shown in Figure B.1, the behaviors of these non-InfoNCE methods are closely related to their optimization objectives and training paradigms. CCA-SSG, although not based on InfoNCE, still exhibits clear overfitting. Its objective combines cross-view alignment with feature decorrelation, but does not explicitly constrain graph structural information to be preserved during optimization. Therefore, continued optimization may still improve the self-supervised objective at the expense of structure-relevant information, suggesting that contrastive overfitting is not specific

to InfoNCE. In contrast, DGI and BGRL show little performance degradation. DGI adopts a local-global discrimination objective rather than instance-level local-local separation, and thus does not continuously push individual nodes apart from one another. BGRL further avoids explicit negative separation and employs a slowly updated target encoder with stop-gradient, which may stabilize the representation trajectory during training. These results further suggest that contrastive overfitting is closely related to the optimization pressure toward topology-agnostic representations, rather than being an artifact specific to the InfoNCE formulation.



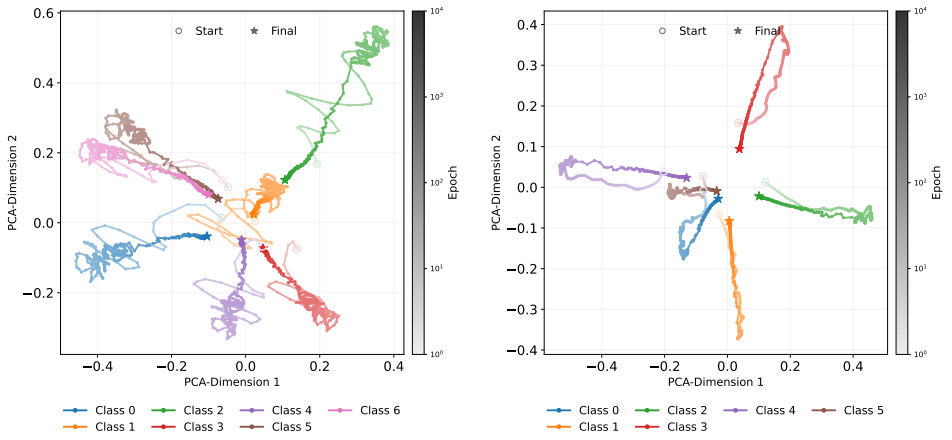
(a) Cora (b) CiteSeer

Figure B.1: Performance degradation under different self-supervised objectives.

B.4 VISUALIZATIONS

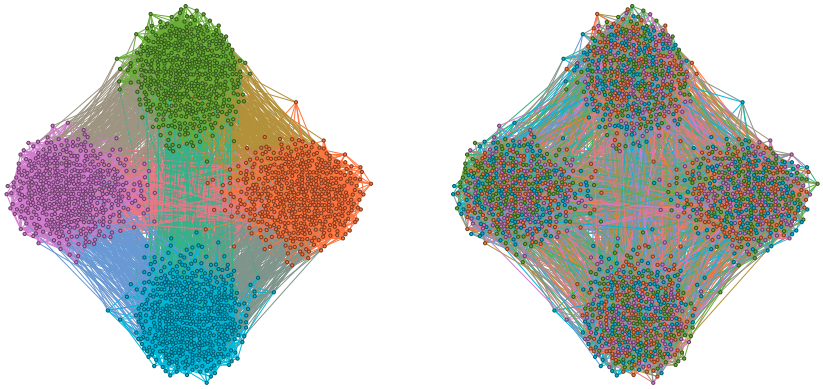
Representation drift trajectories. We further visualize the representation trajectories throughout contrastive optimization on Cora and CiteSeer in Figure B.2. During the early high-return regime, representations from different classes rapidly evolve toward distinct class-specific regions, indicating that contrastive optimization initially helps form a more discriminative embedding structure. As training proceeds into the low-return regime, however, continued representation drift gradually moves the embeddings away from these well-separated configurations, and the final representations become less class-separable than those reached at intermediate stages. This pattern is consistently observed on both datasets, providing an intuitive illustration of contrastive overfitting: prolonged optimization continues to reshape the representation space even after useful class structure has emerged, eventually weakening the previously formed discriminative structure and degrading downstream representation quality.

Synthetic dataset. As shown in Figure B.3, we visualize the graph topology of the SBM dataset. For the SBM graph, we randomly partition the nodes into four communities. The probability of an edge between nodes within the same community is set to 0.02, which is 20 times higher than the inter-community edge probability of 0.001. For each node, we generate a 64-dimensional feature vector from one of four Gaussian distributions with different means and a shared standard deviation of 1. For the **sbm-feature** task, the label of each node is determined by the Gaussian distribution from which its feature vector is sampled, whereas for the **sbm-community** task, the label is determined by its community membership. We randomly split the nodes into 10% for training, 10% for validation, and the remaining 80% for testing.



(a) Cora (b) CiteSeer

Figure B.2: Representation drift trajectories. Node representations are projected into two dimensions using PCA. Each colored trajectory traces the movement of the centroid of one class throughout training.



(a) sbm-community (b) sbm-feature

Figure B.3: Topology visualization of synthetic SBM graphs. Colors denote node classes.

C Theoretical Analysis

We first prove the three main-text theorems. Appendix C.4 then connects the geometry in Theorem 1 and the drift bound in Theorem 3 to quantify structural degradation under continued optimization and structural preservation by PARE.

C.1 Proof of Theorem 1

Theorem 1 (Topology-Independent Regular-Simplex Optimum)**.** *If $N \geq 2$, $d \geq 1$ and $\tau > 0$, then*

$$\mathcal{L}_{\text{con}} \geq \log\left[1 + 2(N-1)\exp\Big(-\frac{N}{(N-1)\tau}\Big)\right] + C_{N,d,\tau},$$

where $C_{N,d,\tau} = e^{-2/\tau}\min\{\frac{1}{\tau}, \frac{1}{2\tau^2}\}\big[\frac{1}{d} - \frac{1}{N-1}\big]_+$ and $[x]_+ = \max\{x, 0\}$. When $d \geq N-1$, the equality holds iff $z_i^1 = z_i^2 = z_i$ and $z_i^\top z_j = -1/(N-1)$ for any $i, j \in [N]$ satisfying $i \neq j$.

Proof. Let $n := 2N$, $U := [(Z^1)^\top, (Z^2)^\top]^\top$, $G := UU^\top$, and let $p(a)$ index the positive partner of u_a. Write

$$F(G) := \frac{1}{n}\sum_{a=1}^{n}\left[\log\sum_{b\neq a} e^{G_{ab}/\tau} - \frac{G_{a,p(a)}}{\tau}\right] = \mathcal{L}_{\mathrm{con}},$$

$$\ell_* := \log\Big[1 + (n-2)e^{-N/((N-1)\tau)}\Big], \qquad c_\tau := e^{-2/\tau}\min\{\tau^{-1}, (2\tau^2)^{-1}\},$$

$$S_* := \frac{NI_N - \mathbf{1}_N\mathbf{1}_N^\top}{N-1}, \qquad G_* := \begin{pmatrix} S_* & S_* \\ S_* & S_* \end{pmatrix}, \qquad H := G - G_*, \qquad v := U^\top\mathbf{1}_n.$$

Then $F(G_*) = \ell_*$, $G_*\mathbf{1}_n = 0$, $H_{aa} = 0$, and $H\mathbf{1}_n = Uv$.

For $f(x) := \log\sum_{j=1}^{n-1} e^{x_j/\tau}$, $x \in [-1,1]^{n-1}$, and $h \in \mathbb{R}^{n-1}$, let

$$w_j := \frac{e^{x_j/\tau}}{\sum_k e^{x_k/\tau}} \geq \frac{e^{-2/\tau}}{n-1}, \qquad \bar{h} := \frac{1}{n-1}\sum_j h_j.$$

Then

$$h^\top\nabla^2 f(x)h = \frac{1}{\tau^2}\min_{t\in\mathbb{R}}\sum_j w_j(h_j - t)^2 \geq \frac{e^{-2/\tau}}{\tau^2(n-1)}\sum_j (h_j - \bar{h})^2.$$

Taylor's formula, applied rowwise along $G_* + tH$, gives

$$F(G) - \ell_* \geq DF(G_*)[H] + \frac{e^{-2/\tau}}{2\tau^2 n(n-1)}R,$$

$$R := \sum_a\sum_{b\neq a}\left(H_{ab} - \frac{1}{n-1}\sum_{c\neq a}H_{ac}\right)^2$$

$$= \|H\|_F^2 - \frac{\|Uv\|_2^2}{n-1} \geq 0.$$

Since $G_{a,p(a)} \leq 1$ and $\sum_a\sum_{b\neq a} H_{ab} = \|v\|_2^2$,

$$DF(G_*)[H] = \frac{\|v\|_2^2 + (n-1)\sum_a(1 - G_{a,p(a)})}{n\tau\big(e^{N/((N-1)\tau)} + n - 2\big)}$$

$$\geq \frac{e^{-2/\tau}}{n\tau(n-1)}\|v\|_2^2 \geq \frac{c_\tau}{n(n-1)}\|v\|_2^2.$$

Using

$$\frac{e^{-2/\tau}}{2\tau^2 n(n-1)} \geq \frac{c_\tau}{n^2}, \qquad \|Uv\|_2^2 = \sum_a (u_a^\top v)^2 \leq n\|v\|_2^2,$$

we obtain

$$F(G) - \ell_* \geq \frac{c_\tau}{n(n-1)}\|v\|_2^2 + \frac{c_\tau}{n^2}\left(\|H\|_F^2 - \frac{\|Uv\|_2^2}{n-1}\right)$$

$$\geq \frac{c_\tau}{n^2}\|G - G_*\|_F^2.$$

Since $G \succeq 0$, $\operatorname{tr} G = n$, and $\operatorname{rank} G \leq d$,

$$\|G\|_F^2 \geq \frac{(\operatorname{tr} G)^2}{d} = \frac{n^2}{d}.$$

Moreover,

$$\|G_*\|_F^2 = \frac{n^2}{N-1}, \qquad 0 \preceq G_* \preceq \frac{n}{N-1}I_n, \qquad \operatorname{tr}(GG_*) \leq \frac{n}{N-1}\operatorname{tr} G = \frac{n^2}{N-1}.$$

Therefore,

$$\|G - G_*\|_F^2 = \|G\|_F^2 + \|G_*\|_F^2 - 2\,\mathrm{tr}(GG_*) \geq n^2\left(\frac{1}{d} - \frac{1}{N-1}\right),$$

Consequently,

$$\mathcal{L}_{\mathrm{con}} \geq \ell_* + c_\tau\left[\frac{1}{d} - \frac{1}{N-1}\right]_+.$$

Finally, $c_\tau > 0$ gives

$$F(G) = \ell_* \iff G = G_* \iff \begin{cases} z_i^1 = z_i^2 = z_i, & 1 \leq i \leq N, \\ z_i^\top z_j = -1/(N-1), & i \neq j. \end{cases}$$

Since $G_* \succeq 0$, $\mathrm{diag}(G_*) = \mathbf{1}_n$, and $\mathrm{rank}\, G_* = N-1$, this configuration is realizable if and only if $d \geq N-1$.

□

C.2 Proof of Theorem 2

Theorem 2 (Structure Degeneration Hurts Structure-Dependent Tasks)**.** *If $Y \perp Z$ when given S and U, and $I(Z;S|U) \leq \epsilon$, then $\mathcal{R}^*_{\log}(Z) - \mathcal{R}^*_{\log}(S,U) \geq \Delta_{\mathrm{str}} - \epsilon$.*

Proof. Let $P(\cdot|W)$ denote the true conditional distribution of Y. For every candidate predictor $p(\cdot|W)$,

$$\mathbb{E}[-\log p(Y|W)] = H(Y|W) + \mathbb{E}_W\left[D_{\mathrm{KL}}\left(P(\cdot|W)\,\|\,p(\cdot|W)\right)\right].$$

The KL term is nonnegative and vanishes for $p = P$, so $\mathcal{R}^*_{\log}(W) = H(Y|W)$.

By $Y \perp Z|(S,U)$ and the chain rule for mutual information,

$$\begin{aligned} I(Y;Z|U) &\leq I(Y,S;Z|U) \\ &= I(S;Z|U) + I(Y;Z|S,U) \\ &= I(S;Z|U) \leq \epsilon. \end{aligned}$$

Since conditioning cannot increase the entropy of the discrete label,

$$\begin{aligned} \mathcal{R}^*_{\log}(Z) - \mathcal{R}^*_{\log}(S,U) &= H(Y|Z) - H(Y|S,U) \\ &\geq H(Y|Z,U) - H(Y|S,U) \\ &= H(Y|U) - I(Y;Z|U) - H(Y|S,U) \\ &= \Delta_{\mathrm{str}} - I(Y;Z|U) \\ &\geq \Delta_{\mathrm{str}} - \epsilon. \end{aligned}$$

□

C.3 Proof of Theorem 3

Theorem 3 (Bounded Cumulative Representation Drift by PARE)**.** *Assume that: (i) $\mathcal{L}_{\mathrm{con}}(\theta)$ is a fixed, β-smooth objective bounded below by F_{low}; (ii) $\theta \mapsto Z(\theta)/\sqrt{N}$ is K-Lipschitz for a fixed input. Let $F_t := \mathcal{L}_{\mathrm{con}}(\theta_t)$. Consider the update in Equation (9) with $0 < \eta \leq 1/\beta$ and $B_t = [F_{t-1} - F_t]_+$. For any $s \geq 1$, if $\bar{R}_t \leq R_{\max} < \infty$ for all $t \geq s$, then*

$$\sum_{t=s+1}^{\infty} C_t \leq K\sqrt{2\eta R_{\max}}\left(F_s - F_{\mathrm{low}} + \frac{\bar{B}_s}{2(1-\mu)}\right).$$

Proof. Let $F_t = \mathcal{L}_{\mathrm{con}}(\theta_t)$. Since $0 \leq w_t \leq 1$ and $\eta\beta \leq 1$, smoothness gives

$$F_t - F_{t+1} \geq \eta w_t\left(1 - \frac{\beta\eta w_t}{2}\right)\|\nabla F(\theta_t)\|_2^2 \geq \frac{\eta w_t}{2}\|\nabla F(\theta_t)\|_2^2 \geq 0.$$

Thus $B_{t+1} = F_t - F_{t+1}$. Using Lipschitz continuity and $w_t \le R_{\max}\bar{B}_t$,

$$\begin{aligned} C_{t+1}^2 &\le K^2\eta^2 w_t^2 \|\nabla F(\theta_t)\|_2^2 \\ &\le 2K^2\eta w_t B_{t+1} \\ &\le 2K^2\eta R_{\max}\bar{B}_t B_{t+1}. \end{aligned}$$

The arithmetic–geometric mean inequality yields

$$C_{t+1} \le K\sqrt{\frac{\eta R_{\max}}{2}}\left(\bar{B}_t + B_{t+1}\right).$$

For any $T > s$, the EMA recursion implies

$$\begin{aligned} \sum_{t=s}^{T-1}\bar{B}_t &= \frac{\bar{B}_s - \mu\bar{B}_{T-1}}{1-\mu} + \sum_{t=s+1}^{T-1} B_t \\ &\le \frac{\bar{B}_s}{1-\mu} + F_s - F_T. \end{aligned}$$

Together with $\sum_{t=s}^{T-1} B_{t+1} = F_s - F_T$, this gives

$$\begin{aligned} \sum_{t=s+1}^{T} C_t &\le K\sqrt{2\eta R_{\max}}\left(F_s - F_T + \frac{\bar{B}_s}{2(1-\mu)}\right) \\ &\le K\sqrt{2\eta R_{\max}}\left(F_s - F_{\text{low}} + \frac{\bar{B}_s}{2(1-\mu)}\right). \end{aligned}$$

Letting $T \to \infty$ completes the proof. □

C.4 Structural Degradation and Preservation

In this section, we connect Theorems 1 and 3 through a structural alignment score. The first result bounds the alignment that can remain as the contrastive loss decreases; the second bounds how much alignment PARE can lose after an intermediate training state.

Measuring structural alignment. For $N \ge 3$, consider representations $Z \in \mathbb{R}^{N\times d}$ with unit-norm rows. Fix a graph-dependent matrix $M = M^\top$ with $M_{ii} = 0$, $\sum_{i,j} M_{ij} = 0$, and $\|M\|_F = 1$, and define

$$Q_M(Z) := \left|\left\langle M, \frac{ZZ^\top}{N}\right\rangle_F\right|, \qquad \langle B, C\rangle_F := \operatorname{tr}(B^\top C). \tag{10}$$

For a simple graph containing both edges and non-edges, one choice is

$$M := \frac{A - \bar{a}(\mathbf{1}\mathbf{1}^\top - I)}{\|A - \bar{a}(\mathbf{1}\mathbf{1}^\top - I)\|_F}, \qquad \bar{a} := \frac{\sum_{i\ne j} A_{ij}}{N(N-1)}.$$

Then Q_M measures the magnitude of the association between adjacency and pairwise cosine similarity after removing the graph-density baseline. Its task relevance depends on the structural relation chosen through M; the score is used only for analysis.

Three properties connect this score to the preceding theorems. First, $0 \le Q_M(Z) \le 1$, since $\|ZZ^\top/N\|_F \le 1$. Second, Q_M is invariant to orthogonal rotations and vanishes at the simplex Gram matrix $S_* = (NI - \mathbf{1}\mathbf{1}^\top)/(N-1)$ because $\langle M, S_*\rangle_F = 0$. Third, for normalized Z, Z',

$$|Q_M(Z) - Q_M(Z')| \le \frac{\|ZZ^\top - Z'Z'^\top\|_F}{N} \le 2\frac{\|Z - Z'\|_F}{\sqrt{N}}. \tag{11}$$

The last inequality follows by expanding the Gram-matrix difference and using $\|Z\|_2, \|Z'\|_2 \le \sqrt{N}$.

Corollary 1 (Contrastive convergence limits structural alignment)**.** *Let Z_t^1, Z_t^2 be the normalized representations entering the symmetric GRACE loss, and let Z_t be the normalized fixed-input representation in the same space. Set*

$$\Delta_t := \mathcal{L}_{\text{con}}(Z_t^1, Z_t^2) - \ell_{\text{simp}}, \qquad r_t := \frac{\|Z_t - Z_t^1\|_F}{\sqrt{N}}.$$

With c_τ defined in Appendix C.1,

$$Q_M(Z_t) \le 2\sqrt{\frac{\Delta_t}{c_\tau}} + 2r_t =: a_t. \tag{12}$$

Hence, alignment decreases relative to an intermediate iterate s whenever $t > s$ and $a_t < Q_M(Z_s)$. If $d \ge N-1$, $\Delta_t \to 0$, and $r_t \to 0$, then $Q_M(Z_t) \to 0$.

Proof. The proof of Theorem 1 gives $\|G_t - G_*\|_F \le 2N\sqrt{\Delta_t/c_\tau}$ for the corresponding block Gram matrices. Since $\langle M, S_*\rangle_F = 0$,

$$Q_M(Z_t^1) \le \frac{\|Z_t^1 (Z_t^1)^\top - S_*\|_F}{N} \le \frac{\|G_t - G_*\|_F}{N} \le 2\sqrt{\frac{\Delta_t}{c_\tau}}.$$

Equation (11) adds the discrepancy term $2r_t$. □

If early learning yields $Q_M(Z_s) > Q_M(Z_0)$ and $\limsup_t a_t < Q_M(Z_s)$, the highest alignment along the trajectory occurs at a finite, noninitial iteration: all sufficiently late scores are below $Q_M(Z_s)$. This does not require a unique peak or monotone degradation. When $d < N-1$, however, Theorem 1 gives $\Delta_t \ge C_{N,d,\tau} > 0$, precluding $\Delta_t \to 0$. The finite-iteration bound remains valid but may be too loose to certify degradation.

We next consider preservation relative to structural optima. Let $\mathcal{K}$ be a nonempty compact set of normalized representations containing the PARE trajectory. Define its structural optima and the distance to them, allowing orthogonal changes of coordinates, by

$$\begin{aligned} \mathcal{Z}_Q^* &:= \arg\max_{Z \in \mathcal{K}} Q_M(Z), \\ \rho_Q(Z) &:= \inf_{\substack{Z^* \in \mathcal{Z}_Q^* \\ O^\top O = I_d}} \frac{\|Z - Z^* O\|_F}{\sqrt{N}}. \end{aligned} \tag{13}$$

The optimal set is nonempty by compactness and continuity.

Corollary 2 (Preservation of an intermediate representation)**.** *Under the assumptions of Theorem 3, including a fixed objective and exact-gradient updates, let*

$$\mathcal{D}_s := K\sqrt{2\eta R_{\max}}\left(F_s - F_{\text{low}} + \frac{\bar{B}_s}{2(1-\mu)}\right). \tag{14}$$

For every $s \ge 1$ to which that theorem applies and every $t \ge s$,

$$Q_M(Z_t) \ge Q_M(Z_s) - 2\mathcal{D}_s, \tag{15}$$

$$\rho_Q(Z_t) \le \rho_Q(Z_s) + \mathcal{D}_s. \tag{16}$$

Proof. By Theorem 3,

$$\frac{\|Z_t - Z_s\|_F}{\sqrt{N}} \le \sum_{k=s+1}^{t} C_k \le \mathcal{D}_s.$$

The score bound follows from Equation (11). For the distance bound, apply the triangle inequality to $Z_t - Z^*O = (Z_t - Z_s) + (Z_s - Z^*O)$ and take the infimum over (Z^*, O). □

Thus, a small $\mathcal{D}_s$ limits the loss of alignment already acquired at Z_s. Distance to structural optima is bounded by the initial distance plus subsequent drift. The reference Z_s must belong to the PARE trajectory; the result does not establish that this trajectory reaches a structural optimum.

These conclusions concern the specified geometric score Q_M, which is distinct from the conditional mutual information in Theorem 2; they do not by themselves imply an accuracy guarantee. The loss-dependent bound applies in the space entering the contrastive objective. With a projection head, transferring it to encoder outputs requires an additional relation between the two spaces.

D EXPERIMENTAL DETAILS

D.1 BASELINES

In this section, we provide brief descriptions of the comparison methods and InfoNCE-based GCL variants used in our experiments. We first introduce three unsupervised representation selection methods and one GCL regularization method used for comparison.

- **α-ReQ** (Agrawal et al., 2022) evaluates representation quality from the eigenspectrum of the empirical feature covariance matrix. It fits the eigenspectrum with a power-law distribution and uses its decay coefficient α as a label-free indicator of downstream representation quality. Code link: https://github.com/BARL-SSL/reptrix
- **RankMe** (Garrido et al., 2023) measures representation quality using the effective rank of the learned representation matrix, computed from the entropy of its normalized singular-value distribution. It provides a simple label-free criterion for model and hyperparameter selection. Code link: https://github.com/BARL-SSL/reptrix
- **LiDAR** (Thilak et al., 2024) improves rank-based representation evaluation by distinguishing informative from uninformative representation directions. It constructs a surrogate self-supervised task and evaluates the effective rank of the corresponding Linear Discriminant Analysis (LDA) structure as an indicator of representation quality. Code link: https://github.com/BARL-SSL/reptrix
- **Contrast-Reg** (Ma et al., 2023) studies the discrepancy between contrastive optimization and downstream performance, where reducing the losses of certain contrastive pairs may adversely affect downstream tasks. It introduces a plug-in regularizer to calibrate contrastive learning and improve the consistency between contrastive loss reduction and downstream representation quality. Code link: https://github.com/MaKaili/Contrast-Reg

We further consider nine representative node-level GCL methods based on the InfoNCE objective.

- **GRACE** (Zhu et al., 2020) constructs two stochastic graph views through edge dropping and feature masking and learns node representations by maximizing the agreement of the same node across views while separating different nodes with an InfoNCE-based objective. Code link: https://github.com/CRIPAC-DIG/GRACE
- **GCA** (Zhu et al., 2021b) extends GRACE with adaptive graph augmentations that preserve important graph structures and attributes. It assigns perturbation probabilities to edges and feature dimensions according to topology- and feature-based importance scores. Code link: https://github.com/CRIPAC-DIG/GCA
- **PiGCL** (He et al., 2024) identifies implicit conflicts between GNN message passing and InfoNCE-based instance discrimination. It dynamically detects conflicting contrastive signals from representation gradients and excludes them from the contrastive objective during training. Code link: https://github.com/hedongxiao-tju/PiGCL
- **ReGCL** (Ji et al., 2024) analyzes the conflict between GNN message passing and contrastive learning from a gradient perspective. It mitigates this conflict through gradient-guided structure learning and a gradient-weighted InfoNCE objective. Code link: https://github.com/RingBDStack/ReGCL
- **ProGCL** (Xia et al., 2022) addresses false negatives in hard-negative mining by estimating the probability that each negative sample is a true negative. It models the similarity distribution of negative pairs with a Beta Mixture Model and incorporates the estimated probabilities into hard-negative weighting or mixing. Code link: https://github.com/junxia97/ProGCL
- **GRACE+** (Chi & Ma, 2024) improves positive and negative pair construction by modeling anchor-aware sampling distributions according to node similarity, thereby increasing the likelihood of informative positive samples while reducing the influence of false negatives. Code link: https://github.com/frankhlchi/SimEnhancedGCL
- **HomoGCL** (Li et al., 2023) explicitly exploits graph homophily to expand the positive set beyond augmented views of the anchor node. It incorporates neighboring nodes as additional

positives and assigns neighbor-specific significance weights to their contributions. Code link: https://github.com/wenzhilics/HomoGCL

- **GRAPE** (Hao et al., 2024) performs expansive and adaptive hard-negative mining based on the subspace-preserving principle. It identifies hard negatives beyond local message-passing neighborhoods and adaptively adjusts the hard-negative set during training. Code link: https://github.com/zz-haooo/WWW24-GRAPE
- **GRACE-IS** (Liu et al., 2024) is motivated by the observation that perfect positive-pair alignment can reduce the contrastive loss while impairing downstream generalization. It combines information- and spectrum-based augmentation strategies to avoid excessive alignment and improve the quality of the learned representations. Code link: https://github.com/somebodyhh1/GRACEIS

D.2 Datasets

In this section, we give brief introductions of the datasets used in this paper. Table D.1 shows detailed information of each dataset.

Table D.1: Dataset statistics.

Dataset	#Nodes	#Edges	#Attributes	#Classes
Cora	2,708	10,556	1,433	7
CiteSeer	3,327	9,228	3,703	6
PubMed	19,717	88,651	500	3
Am-Photo	7,650	238,163	745	8
Am-Computer	13,752	491,722	767	10
Wiki-CS	11,701	431,726	300	10

- **Cora** (Yang et al., 2016): A citation network where each node represents a scientific publication in the field of machine learning, and edges denote citation relationships between papers. Each publication is described by a sparse bag-of-words feature vector derived from its abstract, and is categorized into one of seven predefined research topics.
- **CiteSeer** (Yang et al., 2016): A citation network composed of scientific publications in the field of computer science. Similar to Cora, nodes represent documents and edges represent citation links. Each document is represented by a sparse bag-of-words vector of its content.
- **PubMed** (Yang et al., 2016): A citation network of biomedical research papers from the PubMed database. Each node corresponds to a paper, and edges indicate citation links. Node features are TF-IDF weighted word vectors based on the paper abstracts.
- **Am-Photo** (Shchur et al., 2018): A network of co-purchase relationships constructed from Amazon, where nodes represent products and edges indicate that two products are frequently bought together. Each node is associated with a sparse bag-of-words feature vector derived from product reviews and is labeled according to its category.
- **Am-Computer** (Shchur et al., 2018): A co-purchase network constructed from Amazon, where nodes represent products in the Computers category and edges indicate that two products are frequently purchased together. Each node is associated with a sparse bag-of-words feature vector derived from product reviews and is labeled according to its product category.
- **Wiki-CS** (Mernyei & Cangea, 2020): A reference network derived from Wikipedia, where nodes represent computer science-related articles and edges denote hyperlinks between them. Each node is assigned one of ten class labels, corresponding to distinct subfields within computer science. Node features are computed by averaging the pre-trained GloVe word embeddings of the words appearing in the respective article.

D.3 Evaluation Protocol

Backbone configuration. For each backbone–dataset pair, we follow the original implementation and tune the backbone hyperparameters within the search spaces in Table E.1 to reproduce the

reported performance, accounting for differences in hardware and software environments. We assess reproduction quality through labeled downstream evaluations of embeddings obtained throughout training. Once determined, the configuration is fixed and shared by the original backbone and its PARE variant, including the encoder architecture, projection head, graph augmentations, contrastive objective, optimizer configuration, and backbone hyperparameters. PARE requires no additional hyperparameter search: its exponential moving average coefficient $\mu = 0.9$ and numerical stability constant $\epsilon = 10^{-12}$ are fixed across all backbones and datasets.

In the main experiments, the original backbone and its PARE and Contrast-Reg variants are trained for $T = 10{,}000$ epochs. **Final**, **PARE**, and **Contrast-Reg** report results obtained from their respective final checkpoints. PARE computes its regulation weights exclusively from contrastive loss reduction and representation drift. Once the backbone configuration is fixed, neither node labels nor downstream evaluation scores are used to regulate encoder updates, determine the training duration, or select the checkpoint reported for any of these methods.

Checkpoint selection. **α-ReQ**, **RankMe**, and **LiDAR** operate on the same original backbone training trajectory, without modifying the backbone's training procedure. We compute their label-free scores at every epoch, and each method selects one checkpoint according to its respective selection criterion. The reported accuracy is obtained by evaluating that checkpoint using the frozen-encoder protocol below. Downstream labels and classification scores are not used for these selections. For **Oracle**, we evaluate the frozen embeddings from every epoch of the original backbone and report the highest downstream classification test accuracy along its trajectory. Oracle therefore uses label information to identify the best-performing epoch. It provides a retrospective reference for attainable backbone performance and reproduction quality, and is excluded from the ranking of unsupervised methods.

Downstream evaluation and reporting. For every evaluated checkpoint, we freeze the encoder, extract node embeddings, and train a logistic regression classifier using the training labels. Validation labels are used for classifier selection, and test labels are used to measure classification accuracy. Classifier training does not update the encoder; downstream validation selects the classifier operating on fixed embeddings, without changing the encoder checkpoint.

We follow the standard public splits for Cora, CiteSeer, and PubMed (20 training nodes per class, 500 validation nodes, and 1,000 test nodes), and random 10%/10%/80% splits for Am-Photo, Am-Computer, and Wiki-CS following Zhu et al. (2021b). All compared methods use the same downstream evaluation protocol. We report the mean test accuracy and standard deviation over 10 random seeds.

D.4 Discussion of PARE

Progress-adaptive regulation. PARE can be interpreted as a progress-adaptive regulation of the effective learning rate, where the base learning rate η is dynamically modulated by the regulation weight w_t in Equation (9). The key distinction from conventional learning-rate schedules lies in how the update strength is determined. Predetermined schedules adjust the learning rate as a predefined function of training time, whereas PARE determines w_t from the observed optimization progress, jointly considering the recent contrastive gain and the representation movement required to achieve it. Therefore, PARE does not aim to replace the optimizer or modify the contrastive objective, but instead provides a plug-in mechanism for adapting the effective update strength according to the evolving training dynamics. The comparison with fixed linear and cosine decay schedules in Table B.1 further shows that such progress-adaptive regulation is more robust than predetermined decay under contrastive overfitting.

Loss convention and hyperparameter setting. The progress statistics used by PARE are computed from the normalized mean contrastive objective used consistently throughout this work. For different GCL backbones, their original losses are converted to the same mean-reduced convention before computing the contrastive gain and marginal return. Accordingly, the scale of the progress statistics is defined with respect to this common loss convention. Moreover, "no additional hyperparameter tuning" refers specifically to the parameters introduced by PARE: the backbone configuration is first determined following its original implementation and is then kept fixed for both the original backbone and its PARE variant. The EMA coefficient $\mu = 0.9$ and numerical stability constant $\epsilon = 10^{-12}$ are fixed across all backbones and datasets. Thus, PARE does not use downstream labels

or validation performance to regulate encoder updates, determine the training duration, or select the reported checkpoint.

Theorem 3 provides a trajectory-level characterization of the regulation principle under standard smoothness and Lipschitz assumptions. Its purpose is to isolate how progress-adaptive suppression can bound additional representation movement once optimization enters a low-gain and low-return regime. Practical GCL training may additionally involve stochastic graph augmentations and different optimizer implementations; we therefore view the theorem as a theoretical characterization of the regulation mechanism rather than an exact description of every practical optimization trajectory. The empirical results across different GCL backbones, training horizons, and base learning rates complement this analysis and demonstrate the robustness of PARE under practical training settings.

E Implementation Details

E.1 Implementation of PARE

As shown in Algorithm 1, we provide a PyTorch-style pseudocode of PARE. Each backbone and its PARE variant share the same architecture, augmentations, objective, and optimizer configuration. At each training iteration, PARE computes the regulation weight w_t from the observed contrastive gain and representation drift, and uses it to scale the encoder update. In stochastic GCL training, consecutive loss differences may contain augmentation-induced fluctuations; EMA smoothing is used to estimate their local trend. Therefore, integrating PARE does not require modifying the underlying contrastive objective or model architecture. We keep all hyperparameters the same between a GCL backbone and its variant equipped with PARE. The hyperparameters μ and ϵ are fixed across all methods and datasets. PARE operates on the normalized mean contrastive objective defined in Equation (2); all backbone losses are converted to the same mean-reduced convention before computing the progress statistics.

Algorithm 1 PyTorch-style pseudocode of PARE.

```
# mu = 0.9, eps = 1e-12 (mu and eps are fixed for all methods and datasets)

B_ema, C_ema = 0, 0
L_prev, Z_prev = None, None

for t in range(T):
   # standard GCL forward
   Z = encoder(G)
   L = contrastive_loss(G)

   if t > 0:
      # estimate optimization gain and representation drift
      B = max(L_prev - L, 0)
      C = norm(Z - Z_prev) / sqrt(N)

      # exponential moving average
      B_ema = mu * B_ema + (1 - mu) * B
      C_ema = mu * C_ema + (1 - mu) * C

      # progress-adaptive regulation
      R = B_ema / (C_ema + eps)
      w = min(B_ema * R, 1)
   else:
      w = 1

   # regulated encoder update
   theta -= eta * w * grad(L, theta)

   L_prev, Z_prev = L, Z.detach()
```

E.2 HYPERPARAMETERS

As shown in Table E.1, due to differences in hardware and deep learning libraries, we tune the hyperparameters of each GCL method following the original papers to reproduce their reported performance. PARE requires no additional hyperparameter tuning.

Table E.1: Hyperparameter setting.

Category	**Hyperparameter**	**Search space**
Shared hyperparameters	learning rate	$\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$
	hidden dimension	$\{128, 256, 512\}$
	projection dimension	$\{0, 64, 128, 256, 512\}$
	number of GNN layers	$\{2\}$
	weight-decay coefficient	$\{10^{-5}\}$
	edge-drop rates $(p_{e,1}, p_{e,2})$	$\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$
	feature-drop rates $(p_{f,1}, p_{f,2})$	$\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$
	temperature τ	$\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5\}$
	training epochs	$10,000$
GRACE+	positive threshold τ_{pos}	$\{0.1, 1, 10\}$
	negative threshold τ_{neg}	$\{10^{-3}, 0.1, 10\}$
	feature weight λ_f	$\{0.5, 0.6, 0.7, 0.9\}$
HomoGCL	number of clusters K	$\{5, 10, 30\}$
	homophily coefficient α	$\{1.0\}$
PiGCL	negative-sample ratio	$\{0.05, 0.15, 0.5\}$
	negative-matrix update interval	$\{5, 10\}$
GRAPE	update interval	$\{2, 5, 10\}$
	regularization coefficients (λ, μ)	$\lambda \in \{1, 100\}, \mu = 0.5$
	classifier learning rate and epochs	$\eta_C = 0.1, T_C = 10$
	propagation coefficient η	$\{1000\}$
	threshold θ	$\{1.0\}$
	number of propagation hops k	$\{2, 3, 4, 5\}$

We conduct all experiments on an NVIDIA RTX 5090 GPU.