Measuring Task Similarity and Its Implication in Fine-Tuning Graph Neural Networks

Renhong Huang[1, 2*†], **Jiarong Xu**[2*‡], **Xin Jiang**[3], **Chenglu Pan**[1], **Zhiming Yang**[2], **Chunping Wang**[4], **Yang Yang**[1]

[1]Zhejiang University,
[2]Fudan University,
[3]Lehigh University,
[4]FinVolution Group
{renh2, chenglupan, yangya}@zju.edu.cn, {jiarongxu, zmyang20}@fudan.edu.cn,
xjiang@lehigh.edu, wangchunping02@xinye.com

Abstract

The paradigm of pre-training and fine-tuning graph neural networks has attracted wide research attention. In previous studies, the pre-trained models are viewed as universally versatile, and applied to a diverse range of downstream tasks. In many situations, however, this practice results in limited or even negative transfer. This paper, for the first time, studies the specific application scope of graph pre-trained models, *i.e.*, the extent to which downstream tasks can benefit from specific pre-training tasks. We find that not all downstream tasks can effectively benefit from a graph pre-trained model. In light of this, we introduce the measure *task consistency* to quantify the similarity between graph pre-training and downstream tasks. This measure assesses the extent to which downstream tasks can benefit from specific pre-training tasks. Moreover, a novel fine-tuning strategy, Bridge-Tune, is proposed to further diminish the impact of the difference between pre-training and downstream tasks. The key innovation in Bridge-Tune is an intermediate step that bridges pre-training and downstream tasks. This step takes into account the task differences and further refines the pre-trained model. The superiority of the presented fine-tuning strategy is validated via numerous experiments with different pre-trained models and downstream tasks.

1 Introduction

The paradigm of pre-training and fine-tuning graph neural networks (GNNs) has recently become an active research area and is able to learn transferable knowledge from graph data without costly labels (Hu et al. 2020b; Liu et al. 2022; Rong et al. 2020; Qiu et al. 2020; Xu et al. 2023; Ma et al. 2023; Xu et al. 2022). This paradigm typically involves two steps: (1) pre-train a GNN encoder on unlabeled graph data via a pre-training task; (2) fine-tune the pre-trained GNN on unseen data so as to benefit different downstream tasks. Such a design hopes to build a one-fits-all model that always benefits the downstream.

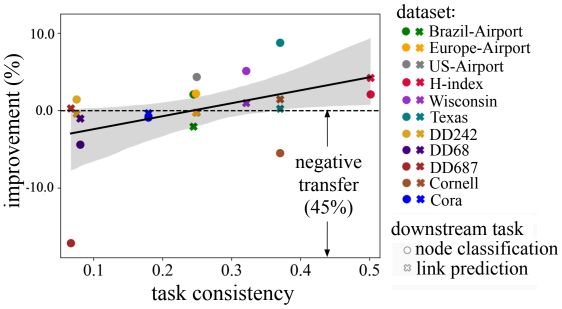


Figure 1: Plot of performance improvement versus the proposed task consistency measure. It shows a clear positive correlation: a larger task consistency implies higher improvement, which in turn suggests that the downstream task can benefit more from the pre-training task. Different points represent improvement on different downstream tasks and different datasets. The black solid line is fitted via linear regression, and the gray shaded area indicates the 95% confidence interval.

However, this ideal expectation is far from the truth in real-world scenarios. As demonstrated in Figure 1, the fine-tuned GCC model (Qiu et al. 2020) suffers from negative transfer in 45.5% of downstream tasks tested, and the given pre-trained model excels in some downstream tasks while underperforms in others. (The improvement result is computed as the relative difference in downstream performance between fine-tuned GCC and GCC learned from scratch.) This undesirable phenomenon is largely attributed to the difference between the pre-training task and the downstream task; as also observed in (Hu et al. 2020b; Lu et al. 2021; Ju et al. 2023).

In view of this, it is then crucial to examine in which case the downstream benefits from the pre-trained model, which in turn asks for a measure to quantify the similarity between a pre-training task and a downstream task. Task similarity has been studied in the literature, but is typically defined based on label distributions (Ganin et al. 2016; Geng 2016; Chen

[*] These authors contributed equally.

[†] This work was done when the author was a visiting student at Fudan University.

[‡] Corresponding author.

et al. 2020). These approaches are not applicable in our case because graph pre-training is conducted on unlabeled data. What's worse, the pre-training and downstream tasks are often defined on different spaces or have distinct objectives, which makes the comparison more difficult.

Considering the above practical needs and challenges, this paper proposes a novel measure of *task consistency* to quantify the similarity between various graph pre-training and downstream tasks within a unified space. Specifically, we introduce a pair-wise label space that is able to encompass different pre-training and downstream tasks even if they are originally defined in different ways. With this novel pair-wise label space, the *task consistency* measure is then proposed to identify those downstream tasks that might benefit from a given graph pre-trained model, as demonstrated in Figure 1.

For those downstream tasks that can potentially make good use of a given pre-trained model, the next question is *how to diminish the impact of task inconsistency* so as to better leverage the knowledge in the pre-trained model. In this case, the proposed task consistency measure is not helpful: The difference between two tasks is intrinsic and is not altered in any learning process. To resolve this difficulty, we introduce the concept of *representation consistency* and a novel fine-tuning strategy *Bridge-Tune*.

The first step is to modify the task consistency measure to take into account the representations learned by the pre-trained model. Such a measure is called *representation consistency*, and can be viewed as a soft version of the task consistency measure. It is able to quantify the contribution of the learned representations to downstream tasks, and is used to guide the refinement of the pre-trained model.

Second, with the proposed representation consistency, we develop a novel fine-tuning strategy, *Bridge-Tune*. The key innovation in Bridge-Tune is an intermediate step between pre-training and downstream. This process is called *pre-trained model refinement*, and aims to maximize the proposed representation consistency. The effectiveness of Bridge-Tune is illustrated in Figure 2. The traditional fine-tuning easily falls into a suboptimal point in the downstream task. In comparison, the pre-trained model refinement step helps find a better starting point for fine-tuning and so Bridge-Tune potentially builds a better model for the downstream task.

Our contributions are summarized as follows.

- **New measure.** We propose a *task consistency* measure to quantify the potential benefits gained by the downstream task from a graph pre-trained model.
- **New method.** We introduce *Bridge-Tune*, a novel fine-tuning strategy. Instead of directly fine-tuning a pre-trained model, Bridge-Tune takes an intermediate step that bridges the pre-training and downstream tasks and refines the model representations.
- **Theoretical guarantees and numerical results.** The effectiveness of Bridge-Tune is verified via theoretical analysis and demonstrated by numerical experiments. In particular, the superiority of Bridge-Tune is justified with different choices of pre-trained models and downstream tasks.

The rest of the paper is organized as follows. The classical paradigm of GNN pre-training and fine-tuning is reviewed in §2. In §3 we introduce the measure of task consistency to quantify the similarity between pre-training and downstream tasks. Then §4 unveils the proposed novel fine-tuning strategy, Bridge-Tune. Numerical experiments in §5 demonstrate the superiority of our approach across various settings.

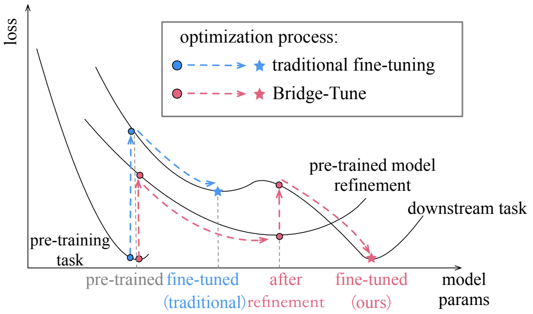


Figure 2: Illustration of the optimization process in traditional fine-tuning (blue) and our fine-tuning strategy (red). The arrow from one (solid) curve to another indicates a change in tasks, and the arrow along one curve represents the optimization process.

2 Preliminaries

In this section, we introduce the basic paradigm of graph pre-training. It typically consists of two steps: pre-training and fine-tuning. First, given a collection of unlabeled graphs G_{pre}, we *pre-train* a generic GNN encoder f_θ by optimizing the self-supervised learning objective $\mathcal{L}_{\text{pre}}$:

$$\theta_{\text{pre-train}} = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}_{\text{pre}}\left(f_\theta; G_{\text{pre}}\right).$$

The learned parameter $\theta_{\text{pre-train}}$ is expected to capture unified and transferable structural patterns in the training graphs. The choice of $\mathcal{L}_{\text{pre}}$ relies on the pre-training task, and this paper focuses on the following three: graph contrastive learning, graph reconstruction and graph context prediction.

Graph Contrastive Learning. The goal of contrastive pre-training task is to capture the similarities (and dissimilarities) between subgraph instances (You et al. 2020; Qiu et al. 2020; Zheng et al. 2022). Specifically, given a subgraph instance ξ_i from an ego network Γ_i centered at the node v_i, we could get its representation $x_i = f(\xi_i)$ via the graph encoder f. The encoder f aims to encourage high similarity between x_i and the representations of another subgraph instance ξ_i^+ which is sampled from the same ego network. This work could be done through optimizing the InfoNCE loss (Oord, Li, and Vinyals 2018): $\mathcal{L}_{\text{pre}} = -\log \frac{e^{x_i^\top f\left(\xi_i^+\right)/\tau}}{e^{x_i^\top f\left(\xi_i^+\right)/\tau}+\sum_{\xi_i'\in\Omega_i^-} e^{x_i^\top f\left(\xi_i'\right)/\tau}}$, where Ω_i^- denotes the collection of subgraph instances that sampled from different ego networks $\Gamma_j\,(j \neq i)$ and τ denotes a pre-defined hyper-parameter. The inner product here denotes the similarity measure between the two subgraph instances.

Graph Reconstruction. Graph autoencoder is another popular approach for GNN pre-training, and utilizes graph reconstruction as self-supervised tasks (Hamilton, Ying, and Leskovec 2017). The main objective of the graph encoder f in graph reconstruction is to encourage high similarity between connected node pairs and low similarity between unconnected node pairs: $\mathcal{L}_{\text{pre}} = -\log \sigma\left(h_u^\top h_v\right) - \log\left(1 - \sigma(h_u^\top h_{v'})\right)$, where v is connected to u but disconnected to v', h_u denotes the representation of node u, and $\sigma(\cdot)$ is the sigmoid function.

Graph Context Prediction. Graph context prediction aims to leverage subgraphs to make predictions about the surrounding graph structures (namely, context graph) (Hu et al. 2020b), by classifying whether a particular neighborhood and a context graph belong to the same node within a K-hop neighborhood. The objective can be formulated as $\mathcal{L}_{\text{pre}} = -\log \sigma(h_v^\top c_v) - \sum_{v' \sim \Omega_v^-} \log\left(1 - \sigma(h_v^\top c_{v'})\right)$, where Ω_v^- is the set of nodes excluding node v, and $h_v^{(K)}$ and c_v are representations of K-hop neighborhood and context graph of node v.

Second, in the fine-tuning stage, the GNN model (initialized with the pre-trained parameters $\theta_{\text{pre-train}}$) is trained on the loss of downstream task $\mathcal{L}_{\text{down}}$ end-to-end together with the classifier on the downstream task. Recent works focus on how to make the most use of the knowledge in pre-trained models during the fine-tuning phase, they can be categorized into parameter regularization (Xuhong, Grandvalet, and Davoine 2018) and representation regularization (Li et al. 2019; Chen et al. 2019; Kou et al. 2020; Flamary et al. 2016; Xu et al. 2020). In the graph domain, some efforts have been also made to develop better fine-tuning strategies. (Zhang et al. 2022) adapts the optimal transport to constrain the fine-tuned model behaviors, which is a kind of representation regularization. (Xia et al. 2022) uses a regularization built on dropout to control the complexity of pre-trained models. Although there are various forms of fine-tuning, it is evident that a gap exists between the learning objectives of the pre-training task and the downstream task.

3 Measure Task Similarity

This section presents a measure to quantify the similarity between pre-training and downstream tasks. We begin to introduce a pair-wise label space to relocate these two tasks in a common space in §3.1, and then our proposed measure of task consistency is presented in §3.2.

3.1 Pair-Wise Label Space

Graph pre-trained models cannot retain competitive performance across all downstream tasks, as significant difference exists between pre-training and downstream tasks. Specifically, the pre-training task typically works on the representation space: By mapping the input data to representations, it attempts to optimize the (dis)agreement of node representations. In comparison, the downstream task is often defined on the label space, aiming to classify the downstream data.

To facilitate the comparison of these tasks, we introduce a new label in both pre-training and downstream tasks. The presented new label is applied to a pair of nodes, and with this definition, the tasks in graph pre-training and downstream can be converted into the same *pair-wise label space*.

Definition 1 (Pair-wise label space). *Given two samples v_i, v_j and their respective labels $y(v_i), y(v_j)$, the label of the node pair (v_i, v_j) is defined as*

$$y^*(v_i, v_j) = \mathbf{1}(y(v_i) = y(v_j)), \tag{1}$$

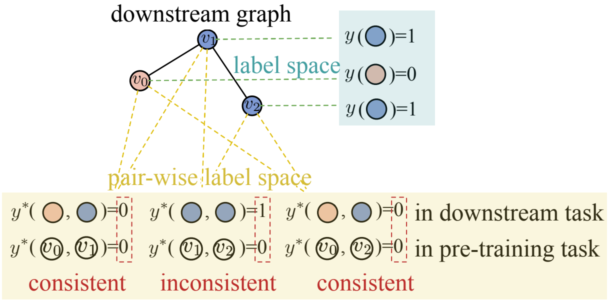


Figure 3: Illustrating example of the pair-wise label space and the measure of task consistency. Consider node classification as a downstream task, originally defined in the label space, a node pair comprising nodes with different labels are labeled as 0 in the pair-wise label space. Take contrastive learning as the pre-training task, a node pair comprising nodes distinct nodes are labeled as 0 in the pair-wise label space. Based on the shared pair-wise label space, the task consistency can be defined to measure the probability of the labels of node pairs being the same in both tasks.

where $\mathbf{1}(\cdot)$ is the indicator function. The pair-wise label space $\boldsymbol{\mathcal{Y}}^*$ consists of all the labels $y^*(v_i, v_j)$.

With Definition 1, a variety of pre-training and downstream tasks can be represented in the pair-wise label space. As an example, node classification in the pair-wise label space sets $y^*(v_i, v_j) = 1$ if v_i and v_j have the same label. The conversion of link prediction into the pair-wise label space is also straightforward: $y^*(v_i, v_j) = 1$ if a link exists between node v_i and v_j, and $y^*(v_i, v_j) = 0$ otherwise.

For pre-training tasks, we study three self-supervised learning approaches and convert them to the pair-wise label space.

- *Graph Contrastive Learning.* Contrastive learning can be viewed as an instance discrimination task, where each instance is treated as a distinct class of its own (Wu et al. 2018). Accordingly, the label of a node pair in contrastive learning can be defined as $y^*(v_i, v_j) = 0$ if $v_i \neq v_j$.
- *Graph Reconstruction.* Graph reconstruction aims to reconstruct the existence of links between node pairs. Hence, the labels of node pairs in this task can be naturally defined as $y^*(v_i, v_j) = 1$ if a link exists between v_i and v_j, and 0 otherwise.
- *Graph Context Prediction.* Graph context prediction is originally a graph-level task, and aims to determine whether a specific neighborhood and a context graph correspond to the same node (Hu et al. 2020b). Roughly speaking, if two nodes are located close enough within K-hops of each other, their neighborhoods and context graphs are considered similar enough. Based on this relaxation, graph context prediction can be converted to a node-level task: $y^*(v_i, v_j) = 1$ if (v_i, v_j) are within K-hops of each other, and $y^*(v_i, v_j) = 0$ otherwise.

Figure 3 presents an illustrating example of converting contrastive learning and node classification into the pair-wise label space.

3.2 Task Consistency

After converting pre-training and downstream tasks to the same pair-wise label space, we introduce the measure *task consistency* to quantify the similarity between these tasks.

Definition 2 (Task consistency)**.** *Given a pre-training task $\mathcal{P}$ and a downstream task $\mathcal{D}$, denote by $\mathcal{Y}^*_{\mathcal{D}}$ its pair-wise label space on the downstream task and by $\mathcal{Y}^*_{\mathcal{P}}$ its pair-wise label space on the pre-training task. The task consistency of the downstream task $\mathcal{D}$ with the pre-training task $\mathcal{P}$ is defined by*

$$C_{\mathrm{T}}(\mathcal{D},\mathcal{P}) = \mathbb{P}[y^*_{\mathcal{D}}(\boldsymbol{n}) = y^*_{\mathcal{P}}(\boldsymbol{n})], \tag{2}$$

where $\mathcal{V}_{\mathcal{D}}$ is the space of nodes in the downstream graph, and $\boldsymbol{n} \in \mathcal{V}_{\mathcal{D}} \times \mathcal{V}_{\mathcal{D}}$ is a node pair taken from the downstream graph.

Figure 3 presents an illustrating example of how task consistency is computed when the pre-training task is contrastive learning and the downstream task is node classification. Towards an empirical verification of task consistency, Figure 1 clearly presents a positive correlation between the task consistency and the performance improvement on downstream tasks brought by pre-trained model (see experiment details in Appendix A.2). The larger the task consistency of a downstream task, the more benefit the task can benefit from the pre-trained model. It is also significant that those downstream tasks with low task consistency suffer from negative transfer. This provides us with the rationality to leverage task consistency to determine the extent to which downstream tasks can benefit from specific graph pre-trained models.

The following theorem builds a theoretical connection between the proposed task consistency and the generalization ability from a pre-training task to a downstream task. Its proof is postponed to Appendix A.5.

Theorem 1 (Connection between generalization error and task consistency)**.** *Let $\mathcal{P}$ and $\mathcal{D}$ be the pre-training task and the downstream task, defined on a shared pair-wise label space. Let $C_{\mathrm{T}}(\mathcal{D},\mathcal{P})$ be the task consistency between $\mathcal{P}$ and $\mathcal{D}$, let $\mathcal{S}$ be an infinite hypothesis set, and let $R(s)$ be the generalization error of a hypothesis $s \in \mathcal{S}$ on $\mathcal{D}$. Then, for any $\delta > 0$, the following inequality holds with probability at least $1-\delta$:*

$$R(s) \leq \frac{\log(|\mathcal{S}|/\delta)}{mC_{\mathrm{T}}(\mathcal{D},\mathcal{P})}.$$

4 Bridge-Tune Enhances Downstream

The proposed task consistency measure reveals which downstream tasks can benefit from specific graph pre-trained models. Building upon this finding, this section introduces a novel fine-tuning strategy that aims at maximizing the utilization of pre-trained models to enhance downstream task performance. This is achieved by mitigating the impact of the difference between pre-training and downstream tasks.

Towards this purpose, in §4.1, drawing inspiration from task consistency, we introduce a measure called *representation consistency* to monitor the impacts of model representations on the downstream tasks during the fine-tuning process. Subsequently, we unveil our novel wisdom of fine-tuning strategy, termed *Bridge-Tune*.

4.1 Representation Consistency

To diminish the impact of task inconsistency during fine-tuning, further improvement on the graph pre-trained model should be made. However, the proposed task consistency is an intrinsic property of the two tasks, so cannot help to further the pre-trained model. In view of this, *representation consistency* is proposed to guide the refinement of the pre-train model. It is a soft version of the task consistency measure, and quantifies the contribution of model representations to the downstream.

Definition 3 (Representation consistency)**.** *Given the output representation space $\mathcal{H}$ of the graph pre-training model, the downstream task $\mathcal{D}$, and the pre-training task $\mathcal{P}$, the representation consistency is defined as*

$$C_{\mathrm{R}}(\mathcal{H},\mathcal{D},\mathcal{P}) = \mathbb{E}_{\boldsymbol{n}}\,[\rho\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{D}}(\boldsymbol{n}) = y^*_{\mathcal{P}}(\boldsymbol{n})], \tag{3}$$

where $\boldsymbol{n}$ is a random node pair from $\mathcal{V}_{\mathcal{D}} \times \mathcal{V}_{\mathcal{D}}$, ρ is a binary indicator (with $\rho = 1$ if $y^*_{\mathcal{P}}(\boldsymbol{n}) = y^*_{\mathcal{D}}(\boldsymbol{n}) = 1$ and -1 if $y^*_{\mathcal{P}}(\boldsymbol{n}) = y^*_{\mathcal{D}}(\boldsymbol{n}) = 0$), $h\colon \mathcal{V}_{\mathcal{D}} \times \mathcal{V}_{\mathcal{D}} \to \mathcal{H} \times \mathcal{H}$ is a mapping from node pair to representation pair, and $\mathrm{Sim}(\cdot)$ *is the cosine similarity.*

The introduction of $\mathrm{Sim}(h(\boldsymbol{n}))$ in (3) is inspired by the observation that for a downstream graph, if nodes with similar (dissimilar) representations also exhibit the same (different) labels, fine-tuning is more likely to maximize the potential of the pre-trained model. Moreover, the representation consistency C_{R} can be viewed as a soft version of the task consistency C_{T}: If $y^*_{\mathcal{D}}(\boldsymbol{n}) = y^*_{\mathcal{P}}(\boldsymbol{n})$, then $C_{\mathrm{R}} = \mathbb{E}[\rho\mathrm{Sim}(h(\boldsymbol{n}))] \leq 1 = C_{\mathrm{T}}$. Further discussion on representation consistency, especially its theoretical connection with some other existing measures (*e.g.*, inter-class distance), can be found in Appendix A.5.

4.2 Our Fine-Tuning Strategy

Motivated by the intuition that a larger representation consistency is more desirable for the downstream, we present in this section a novel fine-tuning strategy, Bridge-Tune, which introduces an intermediate task between pre-training and traditional fine-tuning and further improve downstream performance.

Given a graph pre-trained model, Bridge-Tune consists of two stages: (1) *Pre-trained model refinement*: We maximize the empirically computed representation consistency. This stage acts as an intermediate task between pre-training and traditional fine-tuning. (2) *Downstream fine-tuning*: We conduct traditional fine-tuning, in which the graph model is initialized with the learned parameters of the refined pre-trained model.

Two challenges exist during the refinement process. Computation of the empirical representation consistency needs node labels from the downstream task. However, in many cases, part of the downstream labels are inaccessible, making refinement loss computation difficult. Besides the lack of downstream labels, computing refinement loss is expensive as it involves all pairs of nodes. Thus, the computation efficiency is another concern for our Bridge-Tune model. We tackle these two problems below.

Better estimation of refinement loss. During fine-tuning, only part of the label set is accessible (call it $\boldsymbol{Y}_{\mathrm{L}}$); the remaining part $\boldsymbol{Y}_{\mathrm{U}}$ is inaccessible and needs to be predicted. With only the labeled part of the downstream data, the estimation of representation consistency is far from accurate. We now propose an improved approach for estimating refinement loss. The key insight is that if an unlabeled node is predicted by the downstream classifier with high confidence during downstream fine-tuning, its prediction can serve as an addition to enhancing the estimation of refinement loss during pre-trained model refinement. In view of this, the two stages are suggested to be performed in a progressive and iterative way as below.

Step 1 (Pre-trained model refinement). The graph encoder model is further trained to maximize the refinement loss on the downstream graph. The update at t-th iteration is

$$\theta^{(t)}_{\text{refine}} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}_{\text{refine}}\left(f_\theta; \boldsymbol{Y}_{\text{L}} \cup \boldsymbol{Y}^{(t)}_{\text{P}}\right),$$

where $\mathcal{L}_{\text{refine}}$ is the refinement loss (*i.e.*, the computed representation consistency), f is the graph encoder, $\boldsymbol{Y}^{(t)}_{P}$ is the predictions of unlabeled nodes given by the downstream classifier, and we set $\boldsymbol{Y}^{(0)}_{P} = \emptyset$. This optimization process is initialized at $\theta = \theta^{(t)}_{\text{down}}$ at each iteration (see step 2), and we take $\theta^{(0)}_{\text{down}} = \theta_{\text{pre-train}}$.

Step 2 (Downstream fine-tuning). In this step, the graph encoder f is initialized with $\theta^{(t)}_{\text{refine}}$, and trained end-to-end together with the downstream classifier g (parameterized by $\phi^{(t)}_{\text{down}}$) on a downstream task:

$$(\theta^{(t+1)}_{\text{down}}, \phi^{(t+1)}_{\text{down}}) = \underset{\theta,\phi}{\operatorname{argmin}}\, \mathcal{L}_{\text{down}}\left(f_\theta, g_\phi; \boldsymbol{Y}_{\text{L}}\right),$$
$$\boldsymbol{Y}^{(t+1)}_{\text{P}} = g_{\phi^{(t)}_{\text{down}}} \circ f_{\theta^{(t)}_{\text{down}}}(G_{\text{down}}),$$

where $\mathcal{L}_{\text{down}}$ is the loss of the downstream task, and $g \circ f = g(f(\cdot))$ denotes the composition, and $\mathbf{Y}_P$ is the predictions of unlabeled nodes. For efficiency concerns, the optimization process in the first line is initialized at $(\theta, \phi) = (\theta^{(t)}_{\text{refine}}, \phi^{(t)}_{\text{down}})$.

Steps 1 and 2 are performed iteratively. By doing this, pre-trained model refinement and downstream fine-tuning mutually boost the capability of each other, ultimately boosting the downstream performance.

Efficiency improvement. The computation of refinement loss requires all node pairs, thus leading to a high computation cost. To improve efficiency, we propose to sample two specific categories of critical node pairs. (1) The first category involves node pairs in which either one or both nodes possess labels or high-confidence predictions. They are deemed reliable for learning and can accelerate the training process. This set of node pairs forms P_{certain}. (2) The second category includes node pairs where one node is reliable (*i.e.*, labeled or predicted with high confidence) and the other is the downstream classifier uncertain with. In this way, the information in the reliable nodes can diffuse to unlabeled nodes with low prediction confidence. This set of node pairs constitutes $P_{\text{uncertain}}$.

Given the above two node pair sets, we argue that different node pairs should be paid with different attention during the fine-tuning phase proceeds. Initially, the predictions of the downstream classifier may not be accurate enough, in which case we focus on P_{certain}. As the fine-tuning phase proceeds, we can gradually trust the predictions provided by the downstream classifier, and add $P_{\text{uncertain}}$ to estimate refinement loss while paying less attention to labeled node pairs.

In view of this, a time-varying strategy is developed to assign weights to different node pairs. Specifically, the weight of a node pair (v_i, v_j) at t-th iteration is $a_{ij} = \cos(\frac{\pi t}{2T})p_i p_j$, if $(v_i, v_j) \in P_{\text{certain}}$ and $a_{ij} = \sin(\frac{\pi t}{2T})p_i(1-p_j)$, if $(v_i, v_j) \in P_{\text{uncertain}}$, where p_i and p_j are the probabilities of sampling node v_i and v_j respectively, and T is the total number of iterations.

Finally, the refinement loss at t-th iteration is computed as

$$\mathcal{L}_{\text{refine}} = \sum_{(v_i,v_j)\in P_{\text{certain}} \cup P_{\text{uncertain}}} a_{ij}\rho\text{Sim}(f(v_i), f(v_j))\mathbf{1}(y^*_{\mathcal{D}}(v_i, v_j) = y^*_{\mathcal{P}}(v_i, v_j)).$$

Theoretical analysis. Finally, we theoretically demonstrate that pre-trained model refinement can achieve a lower classification loss on the downstream task than traditional fine-tuning.

Theorem 2 (informal). *Under certain theoretical assumptions, the loss for the downstream task $\mathcal{L}_{down}(\theta_{refine}) \leq \mathcal{L}_{down}(\theta_{pre\text{-}train})$, where θ_{refine} is the model parameter after pre-trained model refinement and $\theta_{pre\text{-}train}$ is the pre-trained model parameter.*

The formal statement of Theorem 2 as well as the proof can be found in Appendix A.5.

5 Experiments

In the experiments, we evaluate the performance of Bridge-Tune with different pre-trained models and on different downstream tasks. We present our setup in §5.1, and the comparison results in terms of performance in §5.2. Additional experimental results are presented in Appendix A.4. Our codes are available at https://github.com/zjunet/Bridge-Tune.

5.1 Experimental Setup

Datasets. We use a total of 12 downstream datasets for evaluation: US-Airport, Brazil-Airport, Europe-Airport, H-index, Wisconsin, Texas, Cora, Cornell, DD242, DD68, DD687, and the large-scale dataset Ogbarxiv. Since our focus is not on the pre-training stage, in the experiments we directly use the graph pre-trained models that have already been trained on their corresponding datasets.

Baselines. A total number of 13 baselines are considered in the experiments, and they can be roughly categorized into three groups: naïve fine-tuning, advanced fine-tuning, and prompt-tuning. For **naïve fine-tuning**, we compare with (1) Fine-tune: graph encoder initialized with pre-trained parameters is trained end-to-end with downstream classifier; (2) Freeze: graph encoder's parameters are frozen during fine-tuning; and (3) Rand: graph encoder is learning from scratch. **For advanced fine-tuning**, we compare with (1) L2_penalty, L2_SP and L2_SP_Fisher (Xuhong, Grandvalet, and Davoine 2018): parameter regularization-based models; (2) DELTA, DELTA w/o Att (Li et al. 2019) and GTOT (Zhang et al. 2022): representation regularization-based models; (3)

Model \ Dataset	US-Airport	Brazil-Airport	Europe-Airport	H-index	Wisconsin	Texas	DD242	DD68	DD687	Cornell	Cora	Ogbarxiv
Task consistency	0.249	0.245	0.248	0.501	0.321	0.370	0.075	0.081	0.067	0.370	0.179	0.077
Fine-tune	66.21(4.23)	73.24(11.60)	58.62(7.35)	81.90(1.59)	59.82(7.69)	63.95(8.67)	14.49(2.94)	12.39(3.74)	8.26(4.16)	48.04(5.94)	29.54(1.44)	18.62(1.92)
Freeze	63.78(5.11)	69.31(14.02)	53.43(7.57)	74.88(0.97)	53.39(8.47)	62.31(10.80)	14.79(3.17)	12.77(4.46)	11.31(2.74)	47.57(5.65)	29.98(1.82)	14.56(7.60)
Rand	63.44(3.22)	71.73(13.56)	57.35(4.72)	81.21(1.62)	56.89(8.11)	59.70(12.75)	12.31(2.17)	13.81(3.12)	11.18(3.25)	**50.82(3.27)**	29.80(1.74)	18.61(1.88)
L2_penalty	64.63(2.31)	67.60(11.93)	55.38(5.37)	79.88(0.75)	55.84(7.62)	56.33(3.46)	16.59(2.28)	12.26(2.80)	8.98(3.60)	42.54(10.81)	31.68(0.84)	19.27(0.21)
L2_SP	63.70(3.73)	67.02(8.42)	54.41(5.53)	78.56(1.97)	53.43(5.53)	56.95(8.59)	13.71(2.07)	9.16(2.19)	7.04(2.03)	36.49(8.95)	36.60(3.22)	19.28(0.17)
L2_SP_Fisher	64.71(3.94)	71.25(12.42)	57.14(4.81)	80.32(1.44)	57.71(6.52)	60.76(13.11)	19.32(3.83)	13.54(4.25)	8.97(2.09)	44.30(5.80)	43.28(2.21)	/
DELTA	63.37(3.13)	62.56(11.56)	55.95(6.23)	75.42(1.02)	55.39(4.31)	56.36(9.82)	14.57(2.78)	11.62(2.77)	9.93(3.00)	45.35(5.95)	39.33(2.09)	/
DELTA w/o Att	62.70(1.84)	62.78(1.60)	53.18(7.08)	72.04(2.17)	55.36(5.94)	64.06(6.32)	14.72(1.73)	9.81(3.18)	7.86(3.85)	47.40(9.43)	37.26(1.48)	19.34(0.13)
GTOT	65.82(3.81)	74.60(14.10)	58.43(4.93)	81.00(1.36)	54.14(6.14)	62.95(7.99)	20.33(3.33)	14.45(2.26)	8.28(2.49)	45.91(8.82)	44.13(2.33)	19.32(0.16)
SupCon	66.56(4.18)	76.21(13.15)	59.88(6.67)	81.00(1.47)	60.60(8.20)	67.28(7.83)	14.88(1.77)	9.54(3.69)	7.17(1.71)	37.69(11.29)	35.42(3.13)	19.36(0.31)
L2P	55.90(3.80)	68.43(16.48)	57.51(4.20)	80.82(1.43)	59.37(9.04)	62.32(4.26)	9.15(2.68)	8.59(1.41)	7.26(2.67)	27.89(15.97)	28.97(3.55)	19.13(0.35)
GPPT	64.03(4.13)	65.66(13.82)	53.12(9.00)	74.66(1.90)	47.42(5.22)	57.90(7.46)	13.63(1.31)	10.83(2.98)	6.35(3.20)	43.60(11.57)	35.45(1.43)	19.85(0.19)
GraphPrompt	62.86(5.90)	60.99(16.06)	50.35(10.46)	73.40(1.85)	51.40(7.27)	59.06(7.59)	13.79(1.83)	11.75(3.00)	5.93(2.40)	39.83(8.19)	36.52(1.97)	19.78(0.25)
Bridge-Tune	**68.99(4.96)**	**77.86(13.95)**	**61.88(5.22)**	**82.66(0.96)**	**62.23(8.44)**	**70.00(5.82)**	**22.97(2.64)**	**16.00(5.60)**	**12.82(4.14)**	50.32(9.37)	**44.17(3.37)**	**20.49(4.35)**

Table 1: Micro F1 scores of different fine-tuning strategies on pre-trained GCC model under the downstream task of node classification. The notation "/" means out of memory or no convergence for more than three days. The p-values comparing our model with competitive baseline GTOT are much smaller than 0.05, which indicates our model significantly outperforms baselines.

SupCon (Khosla et al. 2020): a supervised contrastive learning method, which can also be adopted during fine-tuning; (4) L2P (Lu et al. 2021): While not exactly a fine-tuning strategy as it adjusts the pre-training stage to benefit downstream, we include it for a comprehensive comparison; Note that except for GTOT and L2P, the other baselines are originally designed for convolutional neural networks, so we adapt them to our settings by changing the backbone model to the GNN used in our framework. **For prompt-tuning**, we compare with GPPT (Sun et al. 2022a) and GraphPrompt (Liu et al. 2023b).

Settings. We fine-tune on a variety of graph pre-trained models: GCC, GraphCL, EdgePred, and ContextPred. We set the learning rate as 5, 0.1, 0.1, 0.1 when fine-tuning GCC (Qiu et al. 2020), GraphCL (You et al. 2020), EdgePred (Hamilton, Ying, and Leskovec 2017), and ContextPred (Hu et al. 2020b) respectively. We utilize mini-batch training and the batch size is 32. The total iterations of fine-tuning is 30, alternating between one iteration of pre-trained model refinement and one iteration of downstream fine-tuning. When defining P_{certain}, we regard nodes with prediction confidence higher than 0.5 as high-confidence nodes. The numbers reported in all the experiments are the mean and standard deviation over 10 trials. More details can be found in Appendix A.4.

5.2 Experimental Results

Comparison: fine-tuning baselines. Table 1 presents the node classification performance after fine-tuning on the graph pre-trained model GCC. Our model beats the best baseline by an average of +4.68%. In contrast, advanced fine-tuning baselines often cannot help fine-tuning, and even perform worse than directly fine-tuning the models (*i.e.*, the naïve fine-tuning strategies). One possible reason is that most baselines simply focus on regularizing the parameters and representations, and thus cannot essentially diminish the impact of task difference. The unsatisfactory results of SupCon suggest that those methods tailored for supervised learning might not be suitable in "pre-train and fine-tune" paradigm. Recent prompt-tuning approaches also fall short, possibly because they assume pre-training and fine-tuning are conducted on the same dataset.

		Node Classification	**Link Prediction**
GCC	Task consistency	0.249	0.980
	GTOT	65.82(3.81)	90.76(0.12)
	Bridge-Tune	68.99(4.96)	94.97(0.35)
	improvement	4.82% ↑	4.64% ↑
GraphCL	Task consistency	0.249	0.980
	GTOT	59.58(3.13)	63.81(2.37)
	Bridge-Tune	60.50(2.84)	65.71(1.80)
	improvement	1.54% ↑	2.98% ↑
EdgePred	Task consistency	0.242	1.000
	GTOT	61.09(6.63)	62.46(1.87)
	Bridge-Tune	61.34(2.03)	63.15(1.20)
	improvement	0.41% ↑	1.10% ↑
ContextPred	Task consistency	0.292	0.103
	GTOT	60.72(2.58)	64.54(2.52)
	Bridge-Tune	61.86(2.47)	65.29(0.98)
	improvement	1.88% ↑	1.16% ↑

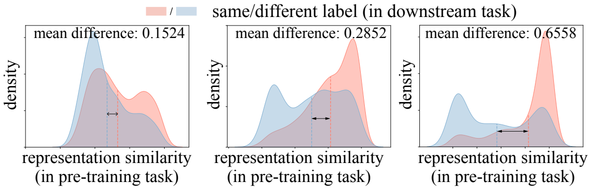
Table 2: Micro F1 on different downstream tasks (in columns), given different pre-trained models (in rows) and on dataset US-Airport.

Comparison: different pre-trained models and downstream tasks. We also conduct experiments using different pre-trained models and on various downstream tasks. Due to space limitations, we only report the comparison between our model and the best baseline GTOT. The results in Table 2 show that our model achieves a significant improvement under various scenarios. More results and details can be found in Appendix A.4.

Ablation study. To demonstrate the effectiveness of each component in our model, we conduct ablation studies on

(1) Bridge-Tune-*a*, which removes the time-varying strategy; (2) Bridge-Tune-P, which does not consider predictions of unlabeled nodes when estimating refinement loss. The Micro F1 score of node classification for Bridge-Tune, Bridge-Tune-*a*, Bridge-Tune-P, fine-tuned on GCC on US-Airport dataset is 68.99%, 67.25% and 68.32% respectively. The superiority of Bridge-Tune compared with Bridge-Tune-*a* and Bridge-Tune-P highlights the importance of time-varying strategy and our estimation of refinement loss.

Case study. To examine whether the proposed pre-trained model refinement can help diminish the impact of the difference between pre-training and downstream tasks, we conduct the following analysis. Figure 4 presents the distribution of pre-trained representation similarity of two nodes in negative pairs. We adopt the pre-trained model GCC on the node classification task on US-Airport. The red (or blue) distribution records the similarity distribution of negative pairs whose two nodes are (or are not) from the same class in downstream task. We first observe a significant difference between pre-training and downstream task: As shown in Figure 4(a), We observe that the similarity distributions of negative pairs within the same class (in red) and across different classes (in blue) in the downstream task are indistinguishable. Then, we can see that our pre-trained model refinement indeed helps diminish the task difference: these two distributions are pulled apart in Figure 4(c). We also note that traditional fine-tuning is less distinctive than ours in diminishing the task difference (see Figure 4(b)). Additional results can be found in Appendix A.4.



(a) After pre-train. (b) After fine-tune. (c) After refinement.

Figure 4: The distribution of representation similarity (cosine similarity) of two nodes in negative pairs. The red (or blue) distribution records the similarity distribution of negative pairs whose two nodes are (or not) from the same class in downstream task. The means are shown in dashed vertical lines.

6 Related Work

Graph fine-tuning strategy. Various fine-tuning strategies have been proposed recently, and most research works concentrate on the image and text domains. These works can be roughly categorized into parameter regularization (Xuhong, Grandvalet, and Davoine 2018) and representation regularization (Li et al. 2019; Kou et al. 2020; Flamary et al. 2016; Xu et al. 2020). However, due to the special structure of graph data, these methods are not directly applicable in the graph domain. Fine-tuning in the graph domain is a promising, yet largely unexplored, research direction. Zhang et al. (2022) adapts representation regularization to graph domain, and the regularizer is inspired by some distances in optimal transport. Though the performance is promising, the use of optimal transport requires a high computational cost, thus not applicable for large-scale models. As another example, Xia et al. (2022) focuses on molecular graphs, and proposes a new regularization tailored to pre-trained molecular model, but this approach can only be applied to molecular graphs.

As a newly developed research direction, prompt-tuning has attracted considerable attention recently. It designs and refines prompts to guide the behavior of pre-trained models towards specific downstream tasks (Liu et al. 2023a). Motivated by this research trend, graph prompt-tuning has received growing research attention. In the context of graph domain, GPPT (Sun et al. 2022a) proposes task token and structure token as prompt template for the node classification applications. GraphPrompt (Liu et al. 2023b) employs a learnable prompt to actively guide downstream tasks using task-specific aggregation. These methods focus on link prediction as the pre-training task and node classification as the downstream task. It is not straightforward to extend these techniques to different tasks, which largely limits their application scope. A very recent paper (Sun et al. 2023) introduces a prompt approach to match various pre-training strategies, but it still lacks explicit consideration of the difference between the pre-training and downstream tasks.

Another line of research, though focusing on the pre-training phase, also attempts to improve downstream performance (Han et al. 2021; Lu et al. 2021). They propose to incorporate auxiliary tasks during pre-training phase so that the pre-trained model is more amenable when adopted to downstream. However, the inclusion of auxiliary tasks potentially compromises the pre-trained model's capacity to generalize across various downstream tasks.

Task similarity. Task similarity refers to the similarity between two machine learning tasks. The research on task similarity was initiated by scholars in the computer vision field, and is used to further improve model performance. Taskonomy (Zamir et al. 2018) delves into the relationship between visual tasks by employing task affinity normalization. Task2vec (Achille et al. 2019) introduces a method to generate task representations. A few follow-up studies apply Task2vec to the graph domain. One such work is GraphGym (You, Ying, and Leskovec 2020), which calculates task similarity by training a collection of anchor models. All the aforementioned works, however, work on supervised tasks, and thus are not applicable to graph pre-training with unlabeled data.

7 Conclusion

We introduce the task consistency measure to quantify the similarity between the graph pre-training and downstream tasks. Such a measure indicates the extent to which a downstream task can benefit from a given pre-training task. Moreover, to diminish the potential impact of task inconsistency, a novel fine-tuning strategy, Bridge-Tune, is proposed, in which the key step aims to mitigate the distinction between the pre-training and downstream tasks. The proposed concepts are theoretically justified, and extensive experiments suggest the superiority of Bridge-Tune on various pre-trained GNN models and downstream tasks.

Acknowledgements

This work is supported by NSFC (No. 62206056, No. 62176233) and the Fundamental Research Funds for the Central Universities.

References

Achille, A.; Lam, M.; Tewari, R.; Ravichandran, A.; Maji, S.; Fowlkes, C. C.; Soatto, S.; and Perona, P. 2019. Task2vec: Task embedding for meta-learning. In *ICCV*, 6430–6439.

Chen, S.; Wang, J.; Chen, Y.; Shi, Z.; Geng, X.; and Rui, Y. 2020. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *CVPR*, 13984–13993.

Chen, X.; Wang, S.; Fu, B.; Long, M.; and Wang, J. 2019. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In *NeurIPS*.

Flamary, R.; Courty, N.; Tuia, D.; and Rakotomamonjy, A. 2016. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 1: 1853–1865.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR*, 17(1): 2096–2030.

Geng, X. 2016. Label distribution learning. *TKDE*, 28(7): 1734–1748.

Gu, Q.; Li, Z.; and Han, J. 2012. Generalized fisher score for feature selection. *arXiv preprint arXiv:1202.3725*.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *NeurIPS*.

Han, X.; Huang, Z.; An, B.; and Bai, J. 2021. Adaptive transfer learning on graph neural networks. In *SIGKDD*, 565–574.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020a. Open graph benchmark: Datasets for machine learning on graphs. In *NeurIPS*, 22118–22133.

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V. S.; and Leskovec, J. 2020b. Strategies for Pre-training Graph Neural Networks. In *ICLR*.

Huang, X.; Yang, Y.; Wang, Y.; Wang, C.; Zhang, Z.; Xu, J.; Chen, L.; and Vazirgiannis, M. 2022. Dgraph: A large-scale financial dataset for graph anomaly detection. In *NeurIPS*, 22765–22777.

Ju, M.; Zhao, T.; Wen, Q.; Yu, W.; Shah, N.; Ye, Y.; and Zhang, C. 2023. Multi-task self-supervised graph neural networks enable stronger task generalization. In *ICLR*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *NeurIPS*, 18661–18673.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of neural network representations revisited. In *ICML*, 3519–3529.

Kou, Z.; You, K.; Long, M.; and Wang, J. 2020. Stochastic normalization. In *NeurIPS*, 16304–16314.

Li, X.; Xiong, H.; Wang, H.; Rao, Y.; Liu, L.; and Huan, J. 2019. Delta: Deep Learning Transfer using Feature Map with Attention for Convolutional Networks. In *ICLR*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. In *ACM Computing Surveys*, volume 55, 1–35. ACM New York, NY.

Liu, S.; Wang, H.; Liu, W.; Lasenby, J.; Guo, H.; and Tang, J. 2022. Pre-training Molecular Graph Representation with 3D Geometry. In *ICLR*.

Liu, Z.; Yu, X.; Fang, Y.; and Zhang, X. 2023b. GraphPrompt: Unifying Pre-Training and Downstream Tasks for Graph Neural Networks. In *WWW*, 417–428.

Lu, Y.; Jiang, X.; Fang, Y.; and Shi, C. 2021. Learning to pre-train graph neural networks. In *AAAI*, 4276–4284.

Ma, R.; Xu, J.; Zhang, X.; Zhang, H.; Zhao, Z.; Zhang, Q.; Huang, X.-J.; and Wei, Z. 2023. One-Model-Connects-All: A Unified Graph Pre-Training Model for Online Community Modeling. In *EMNLP*, 15034–15045.

McCallum, A. K.; Nigam, K.; Rennie, J.; and Seymore, K. 2000. Automating the Construction of Internet Portals with Machine Learning. *Information Retrieval*, 2(3): 127–163.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pei, H.; Wei, B.; Chang, K. C.-C.; Lei, Y.; and Yang, B. 2020. Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:2002.05287*.

Qiu, J.; Chen, Q.; Dong, Y.; Zhang, J.; Yang, H.; Ding, M.; Wang, K.; and Tang, J. 2020. Gcc: Graph contrastive coding for graph neural network pre-training. In *SIGKDD*, 1150–1160.

Ribeiro, L. F.; Saverese, P. H.; and Figueiredo, D. R. 2017. struc2vec: Learning node representations from structural identity. In *SIGKDD*, 385–394.

Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; and Huang, J. 2020. Self-supervised graph transformer on large-scale molecular data. In *NeurIPS*, 12559–12571.

Rossi, R. A.; and Ahmed, N. K. 2015. The Network Data Repository with Interactive Graph Analytics and Visualization. In *AAAI*, 4292–4293.

Sun, M.; Zhou, K.; He, X.; Wang, Y.; and Wang, X. 2022a. Gppt: Graph pre-training and prompt tuning to generalize graph neural networks. In *SIGKDD*, 1717–1727.

Sun, X.; Cheng, H.; Li, J.; Liu, B.; and Guan, J. 2023. All in One: Multi-Task Prompting for Graph Neural Networks. In *SIGKDD*.

Sun, Y.; Deng, H.; Yang, Y.; Wang, C.; Xu, J.; Huang, R.; Cao, L.; Wang, Y.; and Chen, L. 2022b. Beyond homophily: structure-aware path aggregation graph neural network. In *IJCAI*, 2233–2240.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A. A.; and Hardt, M. 2019. Test-time training for out-of-distribution generalization. In *ICLR*.

Wang, B.; Guo, J.; Li, A.; Chen, Y.; and Li, H. 2021. Privacy-preserving representation learning on graphs: A mutual information perspective. In *SIGKDD*, 1667–1676.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 3733–3742.

Xia, J.; Zheng, J.; Tan, C.; Wang, G.; and Li, S. Z. 2022. Towards effective and generalizable fine-tuning for pre-trained molecular graph models. *bioRxiv*.

Xu, J.; Huang, R.; Jiang, X.; Cao, Y.; Yang, C.; Wang, C.; and Yang, Y. 2023. Better with Less: A Data-Centric Prespective on Pre-Training Graph Neural Networks. In *NeurIPS*.

Xu, J.; Yang, Y.; Chen, J.; Jiang, X.; Wang, C.; Lu, J.; and Sun, Y. 2022. Unsupervised adversarially robust representation learning on graphs. In *AAAI*, 4290–4298.

Xu, R.; Liu, P.; Wang, L.; Chen, C.; and Wang, J. 2020. Reliable weighted optimal transport for unsupervised domain adaptation. In *CVPR*, 4394–4403.

Xuhong, L.; Grandvalet, Y.; and Davoine, F. 2018. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, 2825–2834.

You, J.; Ying, Z.; and Leskovec, J. 2020. Design space for graph neural networks. In *NeurIPS*, 17009–17021.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. In *NeurIPS*, 5812–5823.

Zamir, A. R.; Sax, A.; ; Shen, W. B.; Guibas, L.; Malik, J.; and Savarese, S. 2018. Taskonomy: Disentangling Task Transfer Learning. In *CVPR*. IEEE.

Zhang, J.; Xiao, X.; Huang, L.-K.; Rong, Y.; and Bian, Y. 2022. Fine-tuning graph neural networks via graph topology induced optimal transport. *arXiv preprint arXiv:2203.10453*.

Zheng, Y.; Pan, S.; Lee, V.; Zheng, Y.; and Yu, P. S. 2022. Rethinking and scaling up graph contrastive learning: An extremely efficient approach with group discrimination. In *NeurIPS*, 10809–10820.

A Research Methods

A.1 Notations

The main notations can be found in the following table.

Notation	Description
G_{down}, V, E	The downstream dataset, and corresponding node set and edge set
$\mathcal{P}$, $\mathcal{D}$	The pre-training task and the downstream task
$\mathcal{C}_{\text{T}}$, $\mathcal{C}_{\text{R}}$	The task consistency and representation consistency
$\mathcal{Y}$, y	The label space for downstream task and the its realization
$\mathcal{Y}^*_{\mathcal{P}}$, $y^*_{\mathcal{P}}$	The pair-wise label space on $\mathcal{P}$ and its realization
$\mathcal{Y}^*_{\mathcal{D}}$, $y^*_{\mathcal{D}}$	The pair-wise label space on $\mathcal{D}$ and its realization
$\mathcal{V}$, $\mathcal{H}$	The space of nodes in the downstream graph and the space of node representation
$\boldsymbol{n}$, $h(\boldsymbol{n})$	The random variable of node pair sampled from $\mathcal{V} \times \mathcal{V}$ and its corresponding representation pair
f, θ_{down}, θ_{refine}	The pre-trained GNN encoder, parameters of GNN after optimization of the pre-trained model refinement, parameters of GNN after optimization of downstream tasks
g, ϕ_{down}	The downstream classifier, parameters of classifier after optimization of downstream tasks

Table 3: Description of major notations.

A.2 Implement Details

Statistics for datasets. For pre-training, we use the same datasets as pre-trained models did. For the downstream data, we employ 12 graph datasets from various domains: US-Airport (Ribeiro, Saverese, and Figueiredo 2017), Brazil-Airport (Ribeiro, Saverese, and Figueiredo 2017), Europe-Airport (Ribeiro, Saverese, and Figueiredo 2017), H-index (Qiu et al. 2020), Wisconsin (Pei et al. 2020; Sun et al. 2022b), Texas (Pei et al. 2020), Cora (McCallum et al. 2000), Cornell (Pei et al. 2020), DD242 (Rossi and Ahmed 2015), DD68 (Rossi and Ahmed 2015) and DD687 (Rossi and Ahmed 2015). We also include large scale graph dataset Ogbarxiv (Hu et al. 2020a). Table 4 shows their statistics.

		Type	# Nodes	# Edges	# Classes
Downstream	US-Airport	transportation	1,190	13,599	4
	Brazil-Airport	transportation	131	1,074	4
	Europe-Airport	transportation	399	5,995	4
	H-index	coauthorship	5,000	44,020	2
	Cora	coauthorship	2,708	5,278	7
	Ogbarxiv	coauthorship	169,343	1,166,243	40
	Wisconsin	web	251	466	5
	Texas	web	183	309	5
	Cornell	web	183	280	5
	DD68	others	775	2,093	20
	DD687	others	725	2,600	20
	DD242	others	1,284	3,303	20

Table 4: Statistics for the graph pre-training datasets and the downstream datasets.

Experimental details of Figure 1. We evaluate the performance after fine-tuning and learning from scratch in node classification and link prediction on 11 downstream datasets (except for Ogbarxiv due to time cost). We take GCC as the backbone model, and use GCC pre-trained model released by the official code (https://github.com/THUDM/GCC). We show that the correlation between task consistency and improvement is a strong positive correlation with p-value less than 0.1.

Description of baselines. We compare with the following baselines.

- **L2_penalty** regularizes the model parameters with L2 norm.

- **L2_SP** regularizes the distance between the model parameters and the pre-trained parameters with L2 norm.
- **L2_SP_Fisher** regularizes the distance between the model parameters and the pre-trained parameters with Fisher information matrix. We only implement L2_SP_Fisher to GCC. Parts of model parameter (like projection head) of GraphCL are not shared when fine-tuning, which disables the estimation of the Fisher information matrix.
- **Feature (DELTA w/o Att)** regularizes the distance between the model's hidden layer outputs and the pre-trained hidden layer outputs with L2 norm.
- **DELTA** regularizes the distance between the model's hidden layer outputs and the pre-trained hidden layer outputs with an attention-based L2 norm to preserve the transferable outputs.
- **GTOT** utilizes the node representation between the model and the pre-trained model with graph topology induced optimal transport, which is tailored for graph data.
- **SupCon** extend the self-supervised batch contrastive approach to the fully-supervised setting, which can also pull apart the representation after pre-training.
- **L2P** designs a meta learning strategy for dual adaptation on node-level and graph level to narrow the gap during the pre-training phase.
- **GPPT** designs both task tokens and structure tokens, which are employed to formulate node prompts for node classification applications.
- **GraphPrompt** introduced a unifying framework by mapping diverse tasks onto a common task template.

Implementation details of our model. To evaluate the fine-tuning strategies, we take node classification and link prediction as the downstream task. For node classification, we randomly take 90% of the data for training and the remaining data for testing, following (Qiu et al. 2020). When conducting the link prediction, we use 90% existent links and the same number of nonexistent links as the training set, and the rest 10% existent links and the same number of nonexistent links for testing, following (Wang et al. 2021).

We utilize the pre-trained models GCC, GraphCL, EdgePred, and ContextPred, as released by the original paper. Note that we opt for ContextPred with a value of $K = 2$, as specified in the original paper. We fine-tune our model and all baselines with a logistic classifier on node classification and 3 layers of linear network for link prediction, using the Adam optimizer with a learning rate of 0.005 for 50 epochs. For baselines, we set the regularization coefficient as suggested by the original paper, if the coefficient is not provided, we search it in 0.001, 0.01, 0.1, 1.0 and report the best performance. All experiments are conducted on a single machine of Linux system with an Intel Xeon Gold 5118 (128G memory) and a GeForce GTX Tesla P4 (8GB memory).

We then elaborate on the details of sampling node pairs for constructing P_{certain} and $P_{\text{uncertain}}$ respectively. (1) The first kind of node pair is that whose both or either nodes are labeled or predicted with high confidence. These node pairs are reliable for learning and accelerating the training process. In contrast, node pairs consisting of two unlabeled nodes with low prediction confidence would probably give rise to incorrect pair-wise labels, leading to an unstable training process and poor performance. Specifically, let p_i be the confidence of predicted class for node v_i, and 1 otherwise. For a node pair (v_i, v_j), node v_i and v_j are sampled with the probability p_i and p_j, respectively. The sampled node pairs form the set of the first kind of node pairs P_{certain}. (2) On the other hand, the node pairs whose one node is reliable (*i.e.*, labeled or predicted with high confidence) and the other is the downstream classifier uncertain with. In this way, the information in the reliable nodes can diffuse to unlabeled nodes with low prediction confidence, thus maximizing their influence on the whole graph. For a node pair (v_i, v_j), node v_i and v_j are sampled with the probability p_i and $1 - p_j$, respectively. The sampled node pairs form the set of second kind of node pairs $P_{\text{uncertain}}$. In the implementation, we establish a threshold δ and take nodes with confidence p_i exceeding this threshold δ as those having high confidence.

A.3 Algorithm

The overall algorithm for Bridge-Tune is presented in Algorithm 1. As mentioned in § 4.2, our Bridge-Tune pipeline conducts the following two steps iteratively. (i) The GNN model f is continued training to optimize the pre-trained model refinement objective on the downstream graph (lines 4-6). (ii) The graph encoder and classifier are jointly optimized based on the downstream task (lines y).

Algorithm 1: The pre-trained model refinement.

```
Input: Labeled nodes set $\mathbf{Y}_L$, unlabeled nodes set $\mathbf{Y}_U$, pre-trained parameters $\theta_{\text{pre-train}}$, maximum iteration $T_{\max}$, learning rate $\alpha$,
       threshold $\delta$.
Output: $\theta$ for graph encoder $f$, $\phi$ for downstream classifier $g$,
1 initialization $\mathbf{Y}_P^{(0)} \leftarrow \{\}$
2 initialization $\theta_{\text{down}}^{(0)} \leftarrow \theta_{\text{pre-train}}$
3 for $t = 0, 1, 2, \cdots, T_{max} - 1$ do
4     $\mathbf{Y}_P^{(t)} = g_{\phi_{\text{down}}^{(t)}} \circ f_{\theta_{\text{down}}^{(t)}} (G_{\text{down}})$
5     Update $P_{\text{certain}}, P_{\text{uncertain}}$.
6     Conduct pre-trained model refinement task, resulting $\theta_{\text{refine}}^{(t)}$.
7     Conduct downstream fins-tuning task, resulting $\theta_{\text{down}}^{(t+1)}, \phi_{\text{down}}^{(t+1)}$.
8 end
```

The time complexity of our model mainly focuses on calculating the $\mathcal{L}_{\text{refine}}$. The computation of pairwise labels a_{ij} involves calculations between pairs, and our optimization is adapted in the form of mini-batch optimization. Assuming the batch size is B and the number of downstream nodes is $|V|$, and the overall time complexity is $O(|V|/B \times B^2)$, namely $O(|V|B)$, which is linear to the graph size.

A.4 Additional Experiment Results

Fine-tuned performance on different graph pre-trained models. We first present the fine-tuning results on the graph pre-trained model GCC. Table 5 and Table 6 present the results on node classification and link prediction respectively. Note that GPPT and GraphPrompt can only be adopted to node classification tasks, so we do not compare them when conducting link prediction tasks. We find that our model beats the best baseline in most cases. In contrast, the baseline models often cannot help fine-tuning, and even perform worse than directly fine-tuning the models (*i.e.*, fine-tune). The potential reason might be that most baselines simply focus on regularizing the parameters and representations, and thus cannot essentially diminish the impact of the gap between pre-training and downstream tasks.

		Node Classification											
	Model \ Dataset	US-Airport	Brazil-Airport	Europe-Airport	H-index	Wisconsin	Texas	DD242	DD68	DD687	Cornell	Cora	Ogbarxiv
GCC	Fine-tune	66.21(4.23)	73.24(11.60)	58.62(7.35)	81.90(1.59)	59.82(7.69)	63.95(8.67)	14.49(2.94)	12.39(3.74)	8.26(4.16)	48.04(5.94)	29.54(1.44)	18.62(1.92)
	Freeze	63.78(5.11)	69.31(14.02)	53.43(7.57)	74.88(0.97)	53.39(8.47)	62.31(10.80)	14.79(3.17)	12.77(4.46)	11.31(2.74)	47.57(5.65)	29.98(1.82)	14.56(7.60)
	Rand	63.44(3.22)	71.73(13.56)	57.35(4.72)	81.21(1.62)	56.89(8.11)	59.70(12.75)	12.31(2.17)	13.81(3.12)	11.18(3.25)	**50.82(3.27)**	29.80(1.74)	18.61(1.88)
	L2_penalty	64.63(2.31)	67.60(11.93)	55.38(5.37)	79.88(0.75)	55.84(7.62)	56.33(3.46)	16.59(2.28)	12.26(2.80)	8.98(3.60)	42.54(10.81)	31.68(0.84)	19.27(0.21)
	L2_SP	63.70(3.73)	67.02(8.42)	54.41(5.53)	78.56(1.97)	53.43(5.53)	56.95(8.59)	13.71(2.07)	9.16(2.19)	7.04(2.03)	36.49(8.95)	36.60(3.22)	19.28(0.17)
	L2_SP_Fisher	64.71(3.94)	71.25(12.42)	57.14(4.81)	80.32(1.44)	57.71(6.52)	60.76(13.11)	19.32(3.83)	13.54(4.25)	8.97(2.09)	44.30(5.80)	43.28(2.21)	/
	DELTA	63.37(3.13)	62.56(11.56)	55.95(6.23)	75.42(1.02)	55.39(4.31)	56.36(9.82)	14.57(2.78)	11.62(2.77)	9.93(3.00)	45.35(5.95)	39.33(2.09)	/
	Feature (DELTA w/o Att)	62.70(1.84)	62.78(1.60)	53.18(7.08)	72.04(2.17)	55.36(5.94)	64.06(6.32)	14.72(1.73)	9.81(3.18)	7.86(3.85)	47.40(9.43)	37.26(1.48)	19.34(0.13)
	GTOT	65.82(3.81)	74.60(14.10)	58.43(4.93)	81.00(1.36)	54.14(6.14)	62.95(7.99)	20.33(3.33)	14.45(2.26)	8.28(2.49)	45.91(8.82)	44.13(2.33)	19.32(0.16)
	SupCon	66.56(4.18)	76.21(13.15)	59.88(6.67)	81.00(1.47)	60.60(8.20)	67.28(7.83)	14.88(1.77)	9.54(3.69)	7.17(1.71)	37.69(11.29)	35.42(3.13)	19.36(0.31)
	L2P	55.90(3.80)	68.43(16.48)	57.51(4.20)	80.82(1.43)	59.37(9.04)	62.32(4.26)	9.15(2.68)	8.59(1.41)	7.26(2.67)	27.89(15.97)	28.97(3.55)	/
	GPPT	64.03(4.13)	65.66(13.82)	53.12(9.00)	74.66(1.90)	47.42(5.22)	57.90(7.46)	13.63(1.31)	10.83(2.98)	6.35(3.20)	43.60(11.57)	35.45(1.43)	19.85(0.19)
	GraphPrompt	62.86(5.90)	60.99(16.06)	50.35(10.46)	73.40(1.85)	51.40(7.27)	59.06(7.59)	13.79(1.83)	11.75(3.00)	5.93(2.40)	39.83(8.19)	36.52(1.97)	19.78(0.25)
	Bridge-Tune	**68.99(4.96)**	**77.86(13.95)**	**61.88(5.22)**	**82.66(0.96)**	**62.23(8.44)**	**70.00(5.82)**	**22.97(2.64)**	**16.00(5.60)**	**12.82(4.14)**	50.32(9.37)	**44.17(3.37)**	**20.49(4.35)**

Table 5: Micro F1 scores of different fine-tuning strategies on GCC pre-trained model under the downstream task (node classification). The notation “/” means out of memory or no convergence for more than three days.

		Link Prediction										
	Model \ Dataset	US-Airport	Brazil-Airport	Europe-Airport	H-index	Wisconsin	Texas	DD242	DD68	DD687	Cornell	Cora
GCC	Fine-tune	88.44(0.14)	80.94(1.53)	82.23(2.68)	90.02(0.68)	71.13(1.05)	74.52(4.16)	99.06(0.38)	98.13(1.07)	99.32(0.22)	96.02(0.79)	97.14(0.23)
	Freeze	93.23(0.09)	82.52(1.28)	86.02(1.84)	95.67(0.07)	68.37(1.21)	65.95(2.28)	98.73(0.46)	97.53(1.34)	97.83(0.13)	95.60(0.64)	94.88(0.11)
	Rand	87.45(0.19)	82.67(1.66)	82.44(2.54)	87.02(0.21)	70.41(0.80)	74.36(3.88)	**99.67(0.21)**	**99.18(0.21)**	99.01(0.18)	94.67(0.86)	97.48(0.20)
	L2_penalty	94.28(0.12)	84.97(1.51)	88.36(2.62)	96.33(0.13)	70.38(1.43)	65.26(5.32)	99.40(0.07)	97.64(1.06)	98.07(0.16)	96.27(0.81)	95.96(0.31)
	L2_SP	93.91(0.12)	82.46(2.01)	87.40(2.32)	97.51(0.53)	70.50(0.98)	68.20(4.36)	98.97(0.03)	97.80(0.95)	98.40(0.18)	96.32(0.66)	96.09(0.40)
	L2_SP_Fisher	90.93(0.13)	81.90(1.48)	80.97(2.84)	98.43(0.42)	70.47(1.33)	78.63(4.09)	99.10(0.21)	97.83(1.03)	99.06(0.15)	95.87(0.72)	97.34(0.25)
	DELTA	94.47(0.17)	85.93(1.57)	84.05(2.44)	98.52(0.13)	71.42(1.29)	71.52(4.35)	99.23(0.06)	98.40(0.89)	99.06(0.16)	96.18(0.80)	97.20(0.22)
	Feature (DELTA w/o Att)	94.03(0.43)	82.17(1.39)	86.42(2.88)	98.04(0.42)	69.89(2.21)	70.88(3.92)	99.30(0.22)	98.19(1.34)	98.82(0.20)	96.17(0.91)	97.16(0.18)
	GTOT	90.76(0.12)	82.65(1.57)	81.47(2.92)	98.58(0.72)	67.73(2.51)	76.80(3.64)	98.27(0.15)	94.86(2.31)	97.82(0.31)	93.79(1.03)	96.78(0.48)
	SupCon	88.62(1.68)	81.83(1.23)	83.05(2.66)	98.56(0.32)	70.89(1.13)	68.57(5.24)	98.53(0.11)	97.80(0.46)	99.42(0.11)	96.18(0.83)	97.13(0.15)
	L2P	92.28(0.33)	75.97(0.05)	80.81(3.07)	96.61(0.48)	68.30(2.35)	70.68(4.27)	99.19(0.19)	98.32(0.61)	98.78(0.39)	91.46(4.02)	97.92(0.42)
	Bridge-Tune	**94.97(0.35)**	**86.13(1.68)**	**89.47(2.71)**	**98.80(0.13)**	**71.90(1.47)**	**78.87(4.89)**	99.61(0.10)	98.56(1.29)	**99.54(0.11)**	**96.34(1.01)**	**97.92(0.29)**

Table 6: AUC scores of different fine-tuning strategies on GCC pre-trained model under the downstream task (link prediction). Due to the large number of node pair samples in Ogbarxiv, which makes the downstream training very slow, we do not report the results.

In the following, we conduct experiments on other pre-trained models, including GraphCL, EdgePred, ContextPred. It is noted that, we only implement L2_SP_Fisher to GCC, because parts of model parameter (like projection head) of GraphCL, EdgePred, ContextPred are not shared when fine-tuning, which disables the estimation of the Fisher information matrix. Besides, SupCon and L2P are not fine-tuning methods indeed, so we do not compare them when utilzing other pre-trained models. Prompt methods such as GraphPrompt and GPPT are only suitable for node classification tasks, so we do not compare them under link prediction tasks.

Table 7 and Table 8 present the fine-tuning results on GraphCL pre-trained model. We can see that our model could beat the baselines in most of the datasets. In addition, some baseline models perform much worse than learning from scratch (*i.e.*, GraphCL rand), which implies that the negative transfer phenomenon would be exacerbated if we do not carefully choose fine-tuning methods. We also note that most of the results are much worse than those on GCC pre-trained model in Table 5, and there even exists a more serious negative transfer phenomenon when fine-tuning on GraphCL. This is because GraphCL model is not designed for cross-domain pre-training, thus presenting limited generalization ability. In such case, our fine-tuning method still achieves comparable results, which verifies the effectiveness of Bridge-Tune.

Table 9, Table 10, Table 11, and Table 12 illustrate the fine-tuning results on the EdgePred and ContextPred pre-trained models on node classification and link prediction. EdgePred and ContextPred are typical examples of graph reconstruction and graph context prediction pre-training tasks. It is evident that our model outperforms the baselines in most cases. Moreover, we can observe that some baseline models

		Node Classification										
	Model \ Dataset	US-Airport	Brazil-Airport	Europe-Airport	H-index	Wisconsin	Texas	DD242	DD68	DD687	Cornell	Cora
GraphCL	Fine-tune	60.00(3.57)	75.55(9.65)	56.84(9.49)	81.48(1.06)	47.86(10.13)	59.53(11.00)	17.04(5.03)	17.82(7.09)	23.08(12.07)	54.20(9.50)	30.24(2.76)
	Freeze	43.19(3.87)	61.15(10.67)	47.38(5.93)	61.50(1.79)	34.22(8.14)	38.24(10.34)	8.26(2.56)	11.21(4.35)	7.17(3.25)	51.99(10.82)	18.24(2.36)
	Rand	59.08(6.19)	74.84(9.03)	57.12(6.19)	80.90(1.03)	47.86(9.81)	59.04(7.33)	17.06(3.55)	17.96(8.09)	22.40(10.74)	54.09(9.10)	30.39(2.67)
	L2_penalty	57.73(2.66)	73.41(10.47)	54.10(6.89)	52.74(1.45)	47.86(8.96)	56.23(9.91)	14.33(3.02)	13.54(2.67)	12.97(2.52)	**55.18(9.09)**	30.21(2.38)
	L2_SP	58.07(3.51)	79.40(9.73)	54.87(6.59)	71.16(3.99)	49.03(8.89)	57.89(7.46)	13.79(3.0)	12.38(2.96)	13.65(2.96)	54.06(9.64)	29.32(2.83)
	DELTA	57.23(3.84)	76.32(8.77)	55.35(8.29)	78.78(1.91)	45.83(6.08)	56.23(9.91)	13.40(2.95)	15.47(2.99)	13.80(4.85)	51.87(7.44)	29.43(3.88)
	Feature (DELTA w/o Att)	58.15(3.27)	74.84(10.26)	56.34(9.50)	67.74(2.11)	44.26(6.98)	55.15(10.98)	14.25(2.62)	11.99(2.41)	14.22(3.06)	53.54(9.84)	29.47(3.17)
	GTOT	59.58(3.13)	78.57(11.88)	56.85(7.78)	80.94(1.41)	48.63(8.46)	55.61(11.22)	15.58(4.44)	18.99(7.34)	22.68(10.91)	53.04(10.86)	30.76(2.99)
	GPPT	56.81(3.69)	76.43(9.63)	58.62(6.72)	64.29(2.98)	47.05(9.73)	57.87(10.00)	14.95(3.13)	12.52(2.24)	13.66(3.76)	54.62(8.52)	30.13(3.14)
	GraphPrompt	57.31(4.22)	74.12(8.20)	**59.13(8.02)**	65.41(2.89)	47.05(9.73)	57.34(10.39)	15.65(2.73)	13.41(2.14)	12.83(3.08)	54.09(9.10)	30.24(3.42)
	Bridge-Tune	**60.50(2.84)**	**80.31(8.42)**	57.33(9.83)	**81.62(1.36)**	**50.09(13.06)**	**59.88(8.04)**	**17.61(3.76)**	**21.53(7.06)**	**25.04(0.14)**	54.65(9.37)	**30.80(3.02)**

Table 7: Micro F1 scores of different fine-tuning strategies on GraphCL pre-trained model under the downstream task (node classification).

		Link Prediction										
	Model \ Dataset	US-Airport	Brazil-Airport	Europe-Airport	H-index	Wisconsin	Texas	DD242	DD68	DD687	Cornell	Cora
GraphCL	Fine-tune	64.97(1.11)	77.05(0.88)	76.17(2.25)	55.52(1.52)	56.36(1.46)	61.39(12.93)	52.12(0.94)	53.59(1.44)	53.13(2.00)	62.60(2.90)	53.46(1.40)
	Rand	64.97(1.23)	76.88(2.43)	76.50(2.73)	55.37(1.50)	56.76(0.71)	64.11(7.23)	52.26(0.98)	53.37(1.56)	52.22(2.68)	63.11(3.36)	53.23(1.41)
	L2_penalty	63.26(0.70)	74.70(2.13)	75.89(1.51)	53.05(0.68)	**57.28(5.74)**	53.68(8.83)	50.77(1.66)	53.22(1.72)	52.73(2.32)	61.69(3.99)	50.14(1.57)
	L2_SP	64.36(1.95)	75.10(1.64)	76.92(0.64)	55.21(1.59)	56.16(0.83)	62.41(10.52)	51.86(0.64)	53.44(0.75)	52.40(2.45)	61.60(5.15)	53.44(1.88)
	DELTA	61.74(2.50)	73.61(0.91)	73.24(0.66)	55.19(1.42)	52.72(2.51)	58.76(9.04)	51.83(2.92)	53.39(1.03)	51.08(1.28)	57.56(1.67)	50.93(2.34)
	Feature (DELTA w/o Att)	61.88(1.55)	72.49(4.25)	71.13(0.47)	53.43(0.87)	53.43(2.36)	61.13(9.03)	51.22(1.82)	**54.60(1.93)**	51.51(1.56)	57.96(4.26)	51.82(1.12)
	GTOT	63.81(2.37)	77.26(1.55)	76.55(2.35)	55.54(1.56)	56.44(1.20)	63.92(8.58)	51.76(0.74)	53.55(1.18)	51.26(1.14)	61.49(3.04)	52.04(1.57)
	Bridge-Tune	**65.71(1.80)**	**77.68(2.47)**	**77.36(1.77)**	**55.78(1.50)**	55.42(2.68)	**64.92(12.66)**	**52.34(1.07)**	53.48(1.81)	**53.28(2.29)**	**65.34(1.05)**	**53.52(1.61)**

Table 8: AUC scores of different fine-tuning strategies on GraphCL pre-trained model under the downstream task (link prediction).

perform significantly worse than learning from scratch when their task consistency is low, such as DD68 and DD687. Such a result further validates the effectiveness of our consistency metric.

Based on the previous experimental results, we can draw the following two conclusions. (1) The downstream task node classification has the highest task consistency measure (averaged over all datasets) with the pre-training task constrastive learning. Experimental results also support this by showing that our proposed fine-tuning yields improvements of 19.01%, 5.07%, 4.24% in contrastive learning, graph reconstruction and graph context prediction, when compared to the method without pre-training. (2) Link prediction has the highest task consistency with graph reconstruction, which is also supported by experimental results.

		Node Classification										
	Model \ Dataset	US-Airport	Brazil-Airport	Europe-Airport	H-index	Wisconsin	Texas	DD242	DD68	DD687	Cornell	Cora
EdgePred	Fine-tune	58.31(3.17)	75.71(10.28)	56.00(5.15)	81.08(1.32)	47.72(11.34)	58.06(12.37)	17.14(3.87)	17.06(7.08)	22.53(9.17)	54.65(8.06)	30.39(2.46)
	Freeze	44.71(3.12)	38.90(13.90)	43.86(5.49)	60.45(2.30)	31.06(11.38)	42.49(14.10)	7.40(1.91)	7.60(3.36)	6.48(2.81)	42.19(11.03)	17.65(2.99)
	Rand	59.08(6.19)	74.84(9.03)	57.12(6.19)	80.90(1.03)	47.86(9.81)	**59.04(7.33)**	17.06(3.55)	17.96(8.09)	22.40(10.74)	54.09(9.10)	30.49(2.67)
	L2_penalty	49.16(4.61)	62.69(17.03)	51.34(6.02)	70.18(2.82)	47.06(8.91)	55.15(11.20)	14.33(3.02)	13.54(2.67)	12.97(2.52)	55.18(9.09)	30.21(2.38)
	L2_SP	58.32(2.82)	71.81(10.11)	57.36(7.74)	63.19(1.22)	48.26(9.05)	55.70(10.49)	14.33(3.26)	14.96(1.97)	13.67(2.98)	55.15(8.21)	30.28(2.38)
	DELTA	58.57(3.96)	77.09(10.33)	56.36(8.51)	78.86(1.93)	48.66(9.11)	55.70(10.20)	15.74(2.81)	16.76(3.34)	14.35(3.31)	55.18(9.09)	30.46(2.69)
	Feature (DELTA w/o Att)	58.91(2.92)	74.84(11.86)	56.59(8.82)	70.14(2.35)	49.06(9.51)	55.70(10.20)	15.03(2.24)	15.61(3.90)	13.67(4.26)	54.62(8.52)	30.50(2.51)
	GTOT	61.09(3.63)	77.25(11.30)	59.62(7.63)	80.70(1.63)	48.26(9.89)	57.87(11.30)	18.24(5.27)	19.25(6.26)	22.53(9.17)	54.06(9.64)	30.91(3.68)
	GPPT	55.63(2.22)	65.66(14.65)	55.07(9.81)	61.51(2.40)	47.86(8.96)	55.15(11.20)	14.25(3.30)	13.54(2.85)	12.15(3.73)	54.40(7.23)	30.21(2.41)
	GraphPrompt	57.73(3.34)	64.89(13.37)	55.08(9.35)	62.29(2.36)	47.46(8.30)	55.70(10.49)	14.17(3.17)	13.02(2.75)	13.65(2.99)	54.62(8.52)	30.17(2.44)
	Bridge-Tune	**61.34(2.03)**	**77.36(8.06)**	**60.00(4.74)**	**82.00(1.45)**	**49.42(6.82)**	56.26(10.90)	**18.94(6.06)**	**20.28(5.63)**	**24.63(13.51)**	**58.07(6.89)**	**31.22(2.59)**

Table 9: Micro F1 scores of different fine-tuning strategies on EdgePred (graph reconstruction) pre-trained model under the downstream task (node classification).

Bridge-Tune performance on large-scale real-world scenario. We evaluate the performance of Bridge-Tune on a real large-scale financial dataset. DGraph-Fin (Huang et al. 2022) represents a realistic user-to-user social network in the financial industry, and its objective is to differentiate Fraud users (Class 1) from Normal users (Class 0) in the context of anomaly detection. To adapt it to the graph pretraining scenario, we divided DGraph-Fin based on time slices. The data that appeared in the first half of the time period is considered as the pretraining dataset. And the data that appeared in the second half of the time period is considered as the downstream dataset. It is divided into training/validation/test

		Link Prediction										
	Model \ Dataset	US-Airport	Brazil-Airport	Europe-Airport	H-index	Wisconsin	Texas	DD242	DD68	DD687	Cornell	Cora
EdgePred	Fine-tune	60.96(1.36)	73.76(3.22)	74.55(3.53)	54.50(1.93)	50.49(6.67)	49.52(12.21)	48.90(0.56)	52.44(1.92)	50.94(0.81)	51.20(4.20)	49.69(1.93)
	Rand	59.64(0.90)	**77.76(2.49)**	74.82(1.49)	54.31(1.38)	53.15(3.81)	53.32(5.51)	**54.31(0.70)**	**57.47(0.78)**	52.17(1.29)	**61.09(1.89)**	51.66(1.41)
	L2_penalty	54.98(1.70)	66.97(1.53)	63.99(3.71)	52.61(0.43)	53.20(4.99)	44.9(11.99)	48.79(0.83)	50.36(2.78)	49.48(1.39)	53.41(4.97)	51.11(0.85)
	L2_SP	60.30(1.12)	73.09(1.63)	72.95(2.42)	53.09(0.93)	52.00(7.39)	50.16(11.68)	48.87(0.44)	52.12(1.73)	51.17(1.11)	51.17(1.11)	49.51(1.82)
	DELTA	55.49(2.38)	67.74(1.56)	68.61(2.97)	52.98(0.73)	52.91(3.84)	52.45(4.72)	48.89(0.54)	49.74(0.64)	50.10(1.02)	54.96(3.48)	51.83(0.66)
	Feature (DELTA w/o Att)	55.90(1.21)	67.97(3.20)	68.62(1.58)	53.36(0.50)	52.42(5.06)	48.88(6.93)	50.38(0.21)	50.43(0.85)	50.24(0.79)	52.55(4.65)	51.18(1.08)
	GTOT	62.46(1.87)	73.41(1.55)	76.19(3.14)	53.54(1.21)	54.27(6.57)	**58.39(5.77)**	49.74(0.59)	52.09(1.80)	50.40(0.88)	57.31(1.21)	50.35(0.32)
	Bridge-Tune	**63.15(1.20)**	77.23(0.98)	**76.84(2.55)**	**55.27(0.85)**	**57.42(4.06)**	57.35(9.75)	53.73(1.19)	55.38(1.21)	**53.33(3.39)**	60.42(3.84)	**51.97(1.52)**

Table 10: AUC scores of different fine-tuning strategies on EdgePred pre-trained model under the downstream task (link prediction).

		Node Classification										
	Model \ Dataset	US-Airport	Brazil-Airport	Europe-Airport	H-index	Wisconsin	Texas	DD242	DD68	DD687	Cornell	Cora
ContextPred	Fine-tune	60.67(4.15)	75.79(13.10)	60.37(6.49)	81.04(1.62)	48.26(9.73)	58.95(11.71)	16.98(3.64)	15.38(5.19)	21.62(12.24)	54.65(8.06)	31.72(2.79)
	Freeze	44.45(3.41)	51.10(11.64)	51.35(10.15)	56.83(1.93)	37.09(14.83)	47.57(10.50)	11.06(2.43)	9.03(2.81)	8.84(3.49)	47.43(13.73)	20.72(3.77)
	Rand	59.08(6.19)	74.84(9.03)	57.12(6.19)	80.90(1.03)	47.86(9.81)	59.04(7.33)	16.96(3.55)	17.96(8.09)	22.40(10.74)	54.09(9.10)	30.39(2.67)
	L2_penalty	49.16(4.71)	70.27(13.81)	49.83(7.00)	52.78(2.38)	47.86(8.41)	55.15(11.20)	14.33(3.02)	13.54(2.67)	12.97(2.52)	55.18(9.09)	30.21(2.38)
	L2_SP	56.64(3.67)	71.81(10.11)	57.36(7.74)	62.29(5.99)	48.26(9.05)	55.70(10.49)	15.58(3.84)	14.96(1.97)	13.67(2.98)	54.62(8.52)	30.28(2.38)
	DELTA	56.39(1.97)	77.09(10.33)	56.36(8.51)	78.86(1.93)	47.86(8.78)	55.70(10.20)	14.88(3.01)	16.76(3.34)	14.35(3.31)	55.18(9.09)	30.13(2.50)
	Feature (DELTA w/o Att)	55.29(3.54)	75.55(11.34)	57.36(7.50)	63.05(5.39)	48.26(8.50)	56.81(11.81)	15.35(2.63)	15.08(3.77)	15.05(4.03)	54.62(8.52)	29.80(2.30)
	GTOT	60.72(2.58)	77.25(11.30)	60.15(4.41)	80.70(1.63)	48.26(9.89)	57.89(11.00)	16.36(3.91)	**18.85(5.73)**	17.13(6.09)	55.18(8.74)	30.98(2.95)
	GPPT	56.13(3.96)	75.60(8.15)	58.87(7.43)	64.53(2.25)	48.26(9.89)	55.70(10.49)	15.65(3.44)	15.09(2.39)	13.25(2.68)	54.62(8.52)	30.24(3.42)
	GraphPrompt	55.88(4.19)	74.89(10.60)	59.63(6.46)	65.13(2.18)	48.46(5.34)	54.34(8.34)	15.58(3.66)	14.83(2.06)	13.80(3.01)	54.62(8.52)	30.28(2.38)
	Bridge-Tune	**61.86(2.47)**	**77.25(9.75)**	**61.00(2.00)**	**81.44(1.68)**	**49.06(9.51)**	**59.21(12.38)**	**18.24(5.27)**	18.37(3.70)	**23.89(9.12)**	**58.07(6.03)**	**31.81(3.23)**

Table 11: Micro F1 scores of different fine-tuning strategies on ContextPred (graph context prediction) pre-trained model under the downstream task (node classification).

		Link Prediction										
	Model \ Dataset	US-Airport	Brazil-Airport	Europe-Airport	H-index	Wisconsin	Texas	DD242	DD68	DD687	Cornell	Cora
ContextPred	Fine-tune	65.23(2.40)	77.59(2.15)	79.17(0.82)	56.27(0.85)	49.76(2.69)	52.08(8.75)	**57.29(5.77)**	52.86(1.59)	52.74(3.69)	62.15(2.62)	50.57(0.99)
	Rand	65.16(0.55)	77.34(2.32)	**80.46(0.93)**	53.27(2.72)	54.32(1.45)	58.26(5.49)	54.07(2.03)	54.19(1.85)	52.67(1.52)	**64.32(2.05)**	**53.07(0.84)**
	L2_penalty	60.31(1.65)	74.26(3.64)	74.16(2.03)	54.16(1.57)	52.79(3.46)	53.45(2.37)	53.12(1.56)	52.39(1.87)	51.18(2.26)	61.38(3.31)	50.09(1.81)
	L2_SP	64.22(1.96)	77.42(2.35)	78.80(1.36)	55.42(0.16)	50.68(1.22)	52.45(8.89)	54.57(4.15)	53.39(2.06)	52.62(3.29)	60.37(1.40)	50.41(0.91)
	DELTA	59.35(1.98)	78.51(2.23)	75.57(1.52)	55.74(3.6)	53.72(1.14)	56.93(3.76)	54.16(2.25)	52.27(1.44)	54.56(0.85)	60.58(3.76)	50.22(2.21)
	Feature (DELTA w/o Att)	62.28(1.55)	77.69(3.21)	76.01(0.81)	53.78(0.73)	56.36(2.78)	56.01(4.60)	53.77(1.09)	52.02(1.36)	54.51(0.84)	58.77(4.05)	50.15(2.01)
	GTOT	64.54(2.52)	**77.95(1.66)**	79.82(0.11)	54.78(1.31)	54.60(3.51)	58.39(2.44)	51.88(1.97)	52.78(1.12)	54.38(2.09)	60.86(2.67)	51.23(1.99)
	Bridge-Tune	**65.29(0.98)**	77.52(3.22)	79.95(0.13)	**57.52(1.32)**	**57.72(2.34)**	**58.45(2.93)**	56.45(1.88)	**54.32(1.15)**	**54.65(1.66)**	63.23(3.02)	53.03(2.31)

Table 12: AUC scores of different fine-tuning strategies on ContextPred pre-trained model under the downstream task (link prediction).

sets by averaging the classes, with a split of 70/15/15.

We implemented GCC, GraphCL, and JOAO methods to pre-train GNN, and compared our method to the naive fine-tuning approach in Table 13. The implementation details of the pre-trained model used in the experiment are consistent with A.2. From Table 13, we can see that our fine-tuning strategy Bridge-Tune performs well and beats its corresponding backbone models by an average of +7.99% even on large-scale datasets with imbalanced label distributions.

Dataset	# Nodes	# Edges	Method	JOAO	GraphCL	GCC
Pre-training	2,416,372	2,172,701	Fine-tune	56.54(0.09)	51.07(0.32)	64.86(2.85)
Downstream	1,310,092	813,166	Bridge-Tune	**58.32(0.48)**	**60.71(0.52)**	**66.13(0.93)**

Table 13: AUC scores of Bridge-Tune performance on different contrastive-based pre-trained model in DGraph-Fin.

Optimization Analysis. Figure 5 shows the optimization process during fine-tuning. *The left figure* presents the downstream loss curve of directly fine-tuning and that of fine-tuning after refinement. We can see that after performing pre-trained model refinement, we obtain a better start point when conducting downstream task, and achieve lower loss than directly fine-tuning. This suggests that pre-trained model refinement can help downstream make full use of the pre-trained models. *The right figure* shows the model parameter similarity between the reference point and the parameter before and after pre-trained model refinement. We take the parameter of the best performing model in Table 1 as the reference point. We find that pre-trained model refinement process brings the model closer to the best performing model's parameter, which verifies the utility of Bridge-Tune. Another interesting observation is that the parameter similarity with the reference point decreases layer by layer. This validates that the general patterns captured by the first few layers of GNNs are indeed helpful for downstream, which can be applied to downstream without many adaptions.

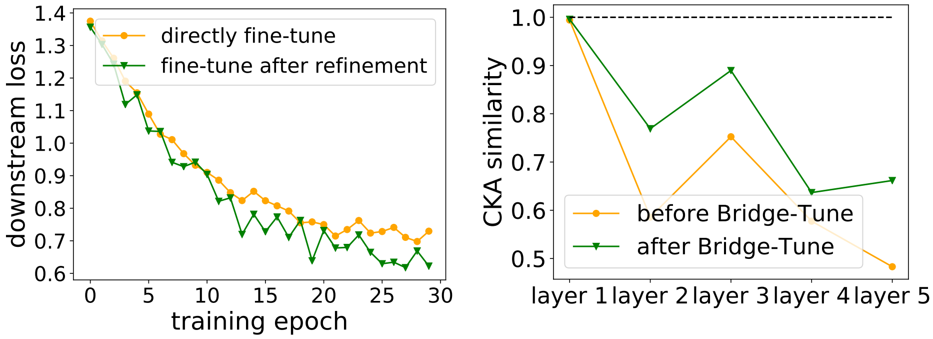


Figure 5: *Left:* The loss curve during fine-tuning phase. *Right:* Parameter similarity between the reference point and parameter before and after pre-trained model refinement in each GNN layer. The similarity is calculated via Centered Kernel Alignment (CKA) similarity (Kornblith et al. 2019).

Evaluation of label-efficiency. In order to more comprehensively demonstrate the label efficiency of our approach, we further investigate the model performance under different label ratios {0.01, 0.05, 0.1, 0.2 ,0.3, 0.4, 0.5, 0.6, 0.7} on node classification and link prediction in the figure below. We find that our model outperforms baselines in most cases, indicating our label-efficiency. Under label ratios {0.05, 0.1, 0.3, 0.5, 0.7}, 56.49%, 58.23%, 62.46%, 64.04%, 65.33% of pseudo labels generated are correct in node classification on US-Airport. In the experiments, when conducting node classification, we assume all links are available. The label ratios of both node classification and link prediction are 90%. The setting is the same as GCC, and more details can be found in §A.2.

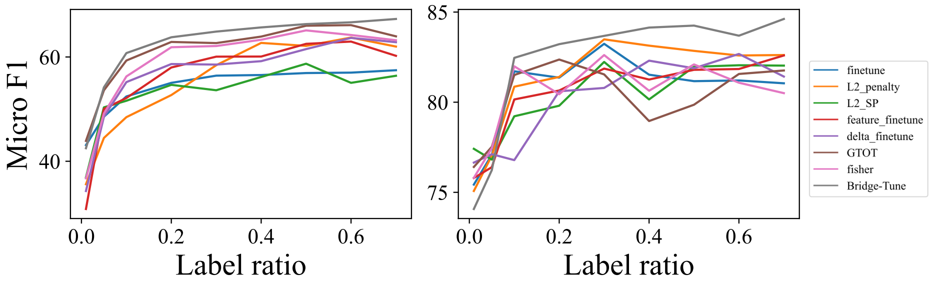


Figure 6: *Left.* Node classification on US-Airport. *Right.* Link prediction on Brazil-Airport. The results are fine-tuned on GCC.

Additional results for case studies. We compare the distribution divergence among datasets and the distributions of representation similarity of two nodes in negative pairs are shown in Figure 7. From the first row, we still find that after pre-training, the similarity distribution of negative pairs with the same class and with different classes is hardly distinguishable in most cases. This suggests the gap between pre-training and downstream task. The final row shows the results after performing the pre-trained model refinement, and we observe a more distinguishable distribution than the previous two cases. It illustrates the effectiveness of Bridge-Tune, but the distributions on some datasets like Wisconsin and H-index are still hard to distinguish. Although, negative transfer phenomenon is alleviated, there is still great potential in distinguishing the distributions of the negative pairs with the same label and different labels.

Effects of hyper-parameters. The hyperparameters learning rate determines the step size at each iteration while moving toward a minimum of a loss function in graph consistency learning, and is thus an important hyperparameter in our proposed strategy. Figure 8 shows the effect of the learning rate on downstream performance. We can see that a too small or too large learning rate could deteriorate the performance, and the optimal performance can be obtained when learning rate is 5.

Comparison: our two-stage optimization vs joint training. We further compare our two-stage approach with the joint training of pre-trained model refinement and downstream fine-tuning (*i.e.*, Joint: pre-trained model refinement + downstream fine-tuning) and co-training pre-trained model refinement with the SSL task (*i.e.*, Joint: SSL + pre-trained model refinement) during the fine-tuning stage in Table 14 . The results

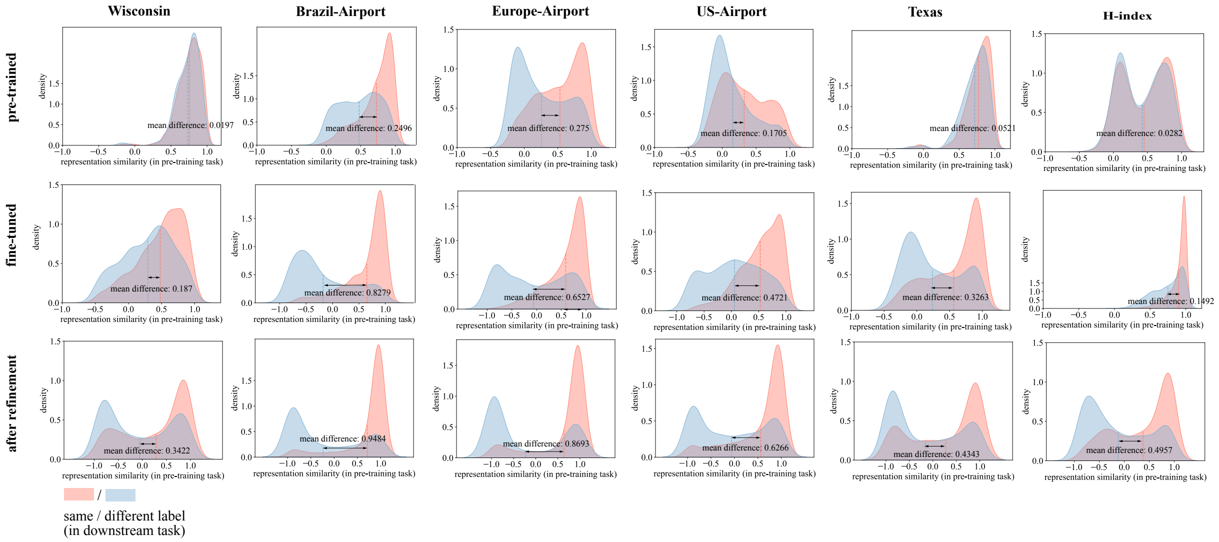
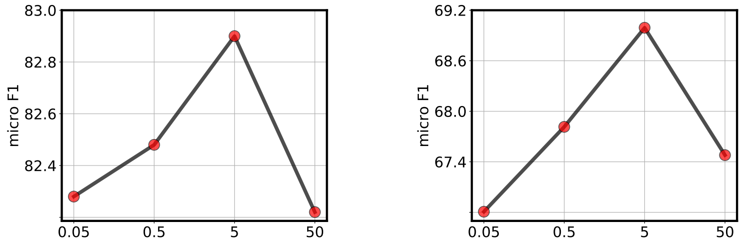


Figure 7: The distribution of representation similarity (cosine similarity) of two nodes in negative pairs. The red (or blue) distribution records the similarity distribution of negative pairs that are (or not) from the same class. *Each row* represents the distribution of representation given by the pre-trained model, directly fine-tuning and after refinement, respectively. *Each column* represents the distribution on different downstream datasets. All the experiments are conducted on GCC pre-trained model on node classification.



(a) H-index

(b) US-Airport

Figure 8: Performance of Bridge-Tune w.r.t varying learning rate (GCC is adopted as the pre-trained model).

indicate the effectiveness of our two-stage approach, while Joint: SSL + pre-trained model refinement achieves a smooth transition to the downstream task often leads to negative transfer effects.

Model \ Dataset	US-Airport	Brazil-Airport	Europe-Airport	H-index	Wisconsin	Texas	DD242	DD68	DD687	Cornell	Cora
GCC fine-tune	66.21(4.23)	73.24(11.60)	58.62(7.35)	81.90(1.59)	59.82(7.69)	63.95(8.67)	14.49(2.94)	12.39(3.74)	8.26(4.16)	48.04(5.94)	29.54(1.44)
Joint: SSL + pre-trained model refinement	61.77(4.82)	63.41(12.58)	54.36(9.09)	76.74(1.81)	53.80(6.10)	54.56(12.85)	14.57(2.43)	10.96(3.47)	7.59(2.16)	44.24(8.52)	33.86(1.61)
Joint: pre-trained model refinement + downstream fine-tuning	62.77(4.05)	67.86(13.36)	53.62(7.76)	76.34 (3.10)	52.63(8.16)	63.30(7.74)	14.95(1.96)	9.38(2.38)	10.19(3.37)	45.35(12.00)	34.79(1.97)
Bridge-Tune	**68.15(4.62)**	**77.86(13.95)**	**61.88(5.22)**	**82.66(0.96)**	**62.23(8.44)**	**70.00(5.82)**	**22.97(2.64)**	**16.00(5.60)**	**12.82(4.14)**	**50.32(9.37)**	**44.17(3.37)**

Table 14: Micro F1 scores of fine-tuning strategies and joint learning strategies on GCC pre-trained model under the downstream task (node classification).

Detailed results of runtime comparison. We compare the training time of baseline models as well as our model in Table 15.

A.5 Theoretical proofs

Proof of Theorem 1. We here provide the detailed proof of Theorem 1 in the main body.

Theorem 1 (Connection between generalization error and task consistency.) Let $\mathcal{P}$ and $\mathcal{D}$ be the pre-training task and the downstream task, defined on a shared pair-wise label space. Let $\mathcal{C}_{\mathrm{T}}(\mathcal{D}, \mathcal{P})$ be the task consistency between $\mathcal{P}$ and $\mathcal{D}$, let $\mathcal{S}$ be an infinite hypothesis set, and let $R(h)$ be the generalization error of a hypothesis $s \in \mathcal{S}$ on $\mathcal{D}$. And m is the number of instances. Then, for any $\delta > 0$, with probability at least

	Fine-tune	GPPT	GraphPrompt	L2_penalty	L2_SP	DELTA	Feature (DELTA w/o Att)	GTOT	Bridge-Tune
pre-trained model refinement	-	-	-	-	-	-	-	-	28.65(1.45)
downstream fine-tuning	150.66(4.47)	41.41(0.54)	45.41(3.89)	174.68(1.48)	177.58(1.11)	208.89(3.38)	208.17(2.74)	478.09(9.26)	144.86(8.65)
total	150.66(4.47)	41.41(0.54)	45.41(3.89)	174.68(1.48)	177.58(1.11)	208.89(3.38)	208.17(2.74)	478.09(9.26)	173.51(8.85)

Table 15: Runtime (sec) comparison during the fine-tuning of GCC for node classification on US-Airport. All models are trained till convergence, where the convergence condition is defined as the point where the increase in accuracy on the training set is less than 0.01.

$1 - \delta$, it follows that

$$R(h) \leq \frac{\log(|\mathcal{S}|/\delta)}{mC_{\mathrm{T}}(\mathcal{D}, \mathcal{P})}.$$

Proof: Fix any $\epsilon > 0$, define $\mathcal{S} = \{h : R(h) > \epsilon\}$. The probability that a hypothesis h in $\mathcal{S}$ is consistent on a training sample S drawn i.i.d. and the task consistency $C_{\mathrm{T}}(\mathcal{D}, \mathcal{P})$ determine the proportion of samples that should have the same pair-wise label. Therefore, it can be bounded as follows:

$$\mathbb{P}\left[\widehat{R}(h) = 0\right] \leq (1-\epsilon)^{mC_{\mathrm{T}}(\mathcal{D},\mathcal{P})}\epsilon^{m(1-C_{\mathrm{T}}(\mathcal{D},\mathcal{P}))}.$$

Thus, by the union bound, the following holds:

$$\begin{aligned}
\mathbb{P}\left[\exists h \in \mathcal{S} : \widehat{R}(h) = 0\right] &= \mathbb{P}\left[\widehat{R}\left(h_1\right) = 0 \vee \cdots \vee \widehat{R}\left(h_{|\mathcal{S}|}\right) = 0\right] \\
&\leq \sum_{\mathcal{S}} (1-\epsilon)^{mC_{\mathrm{T}}(\mathcal{D},\mathcal{P})}\epsilon^{m(1-C_{\mathrm{T}}(\mathcal{D},\mathcal{P}))} \\
&\leq |\mathcal{S}|(1-\epsilon)^{mC_{\mathrm{T}}(\mathcal{D},\mathcal{P})}\epsilon^{m(1-C_{\mathrm{T}}(\mathcal{D},\mathcal{P}))} \\
&\leq |\mathcal{S}|(1-\epsilon)^{mC_{\mathrm{T}}(\mathcal{D},\mathcal{P})} \leq |\mathcal{S}|e^{-m\epsilon C_{\mathrm{T}}(\mathcal{D},\mathcal{P})}.
\end{aligned}$$

Setting the right-hand side to be equal to δ and solving ϵ, which concludes the proof.

Proof of Theorem 2 and convergence of Bridge-Tune. In this section, we present the theoretical analysis of Bridge-Tune. We first prove Theorem 2 in the main body. Then, we theoretically demonstrate that Bridge-Tune can converge to the optimum faster.

Theorem 2 [informal] Under certain theoretical assumptions, the loss for the downstream task $\mathcal{L}_{\text{down}}(\theta_{\text{refine}}) \leq \mathcal{L}_{\text{down}}(\theta_{\text{pre-train}})$, where θ_{refine} is the model parameter after pre-trained model refinement and $\theta_{\text{pre-train}}$ is the pre-trained model parameter.

Theorem 2. Let $\mathcal{L}_{\text{down}}(\theta, \phi)$ denote the downstream loss with model's parameters θ and downstream classifier's parameter ϕ, and $\mathcal{L}_{\text{refine}}(\theta)$ represent the refinement loss. Assume that $\mathcal{L}_{\text{down}}(\theta, \phi)$ is differentiable, convex and β-smooth in θ, and $\|\nabla\mathcal{L}_{\text{down}}(\theta, \phi)\|, \|\nabla\mathcal{L}_{\text{refine}}(\theta)\| \leq M$ for all θ and ϕ. With a fixed learning rate $\eta = \frac{\epsilon}{\beta M^2}$, for every x, y such that $\langle\nabla\mathcal{L}_{\text{down}}(\theta, \phi), \nabla\mathcal{L}_{\text{refine}}(\theta)\rangle > \epsilon$, we have

$$\mathcal{L}_{\text{down}}(\theta, \phi) > \mathcal{L}_{\text{down}}(\theta', \phi), \tag{4}$$

where $\theta' = \theta - \eta\nabla\mathcal{L}_{\text{refine}}(\theta)$ is the pre-trained model refinement with one step of gradient descent.

Proof of Theorem 2. Draw inspiration from (Sun et al. 2019), for any learning rate η, by β-smoothness, we have

$$\begin{aligned}
\mathcal{L}_{\text{down}}(\theta', \phi) &= \mathcal{L}_{\text{down}}\left(\theta - \eta\nabla_\theta\mathcal{L}_{\text{refine}}(\theta), \phi\right) \\
&\leq \mathcal{L}_{\text{down}}\left(\theta, \phi\right) - \eta\left\langle\nabla_\theta\mathcal{L}_{\text{down}}(\theta, \phi), \nabla_\theta\mathcal{L}_{\text{refine}}(\theta)\right\rangle + \frac{\eta^2\beta}{2}\left\|\nabla_\theta\mathcal{L}_{\text{refine}}(\theta)\right\|^2.
\end{aligned} \tag{5}$$

By substituting the η with the below value, we have

$$\eta^* = \frac{\left\langle\nabla_\theta\mathcal{L}_{\text{down}}(\theta, \phi), \nabla_\theta\mathcal{L}_{\text{refine}}(\theta)\right\rangle}{\beta\left\|\nabla_\theta\mathcal{L}_{\text{refine}}(\theta)\right\|^2}.$$

We can deduce that

$$\mathcal{L}_{\text{down}}\left(\theta - \eta^*\nabla_\theta\mathcal{L}_{\text{refine}}(\theta), \phi\right) \leq \mathcal{L}_{\text{down}}(\theta, \phi) - \frac{\left\langle\nabla_\theta\mathcal{L}_{\text{down}}(\theta, \phi), \nabla_\theta\mathcal{L}_{\text{refine}}(\theta)\right\rangle^2}{2\beta\left\|\nabla_\theta\mathcal{L}_{\text{refine}}(\theta)\right\|^2}.$$

Since $\|\nabla\mathcal{L}_{\text{down}}(\theta, \phi)\|, \|\nabla\mathcal{L}_{\text{refine}}(\theta)\| \leq M$, and the inner product is larger than ϵ, we have

$$\mathcal{L}_{\text{down}}(\theta, \phi) - \mathcal{L}_{\text{down}}\left(\theta - \eta^*\nabla_\theta\mathcal{L}_{\text{refine}}(\theta), \phi\right) \geq \frac{\epsilon^2}{2\beta M^2}.$$

Given $\eta = \frac{\epsilon}{\beta M^2}$, we could deduce that when $0 < \eta \le \eta^*$, we have

$$\begin{aligned}
\mathcal{L}_{\text{down}}(\theta', \phi) &= \mathcal{L}_{\text{down}}(\theta - \eta \nabla_\theta \mathcal{L}_{\text{refine}}(\theta), \phi) \\
&= \mathcal{L}_{\text{down}}\left(\left(1 - \frac{\eta}{\eta^*}\right)\theta + \frac{\eta}{\eta^*}\left(\theta - \eta^* \nabla_\theta \mathcal{L}_{\text{refine}}(\theta)\right), \phi\right) \\
&\le \left(1 - \frac{\eta}{\eta^*}\right)\mathcal{L}_{\text{down}}(\theta, \phi) + \frac{\eta}{\eta^*}\mathcal{L}_{\text{down}}\left(\theta - \eta^* \nabla_\theta \mathcal{L}_{\text{refine}}(\theta), \phi\right) \\
&\le \left(1 - \frac{\eta}{\eta^*}\right)\mathcal{L}_{\text{down}}(\theta, \phi) + \frac{\eta}{\eta^*}\left(\mathcal{L}_{\text{down}}(\theta, \phi) - \frac{\epsilon^2}{2\beta G^2}\right) \\
&= \mathcal{L}_{\text{down}}(\theta, \phi) - \frac{\eta}{\eta^*}\frac{\epsilon^2}{2\beta M^2}.
\end{aligned} \tag{6}$$

Due to $\frac{\eta}{\eta^*} > 0$, we have

$$\mathcal{L}_{\text{down}}(\theta, \phi) > \mathcal{L}_{\text{down}}(\theta', \phi). \tag{7}$$

Theorem 3 (Convergence of fine-tuning task)**.** *Let $\mathcal{L}_{down}(\theta, \phi)$ be the downstream loss after fine-tuned with pre-trained model's parameter θ and downstream classifier's parameter ϕ. Given that $\mathcal{L}^*_{down}$ is the optimal downstream loss, we can obtain the upper bound as follows:*

$$\mathcal{L}_{down}(\theta, \phi) - \mathcal{L}^*_{down} < O\left(\mathcal{L}_{down}\left(\theta_0, \phi_0\right) - \mathcal{L}^*_{down}\right), \tag{8}$$

where θ_0, ϕ_0 are the initial parameters of pre-trained model and downstream classifier before fine-tuning, which vary across different approaches. In particular, in our two-stage approach, θ_0, ϕ_0 are the parameters given by the pre-trained model refinement; while in naïve fine-tuning, they refer to the parameters given by the pre-trained model.

Proof of Theorem 3. Let $\mathcal{L}_{\text{down}}(\theta, \phi)$ be a β-smooth, μ-strongly convex function for $\mu > 0$. Since $\mathcal{L}_{\text{down}}(\theta, \phi)$ is β-smooth, we could derive that

$$\frac{1}{2\beta}\|\nabla \mathcal{L}_{\text{down}}(\theta, \phi)\|_2^2 \le \mathcal{L}_{\text{down}}(\theta, \phi) - \mathcal{L}^*_{\text{down}}$$

We can also prove that

$$\frac{1}{2\mu}\|\nabla \mathcal{L}_{\text{down}}(\theta, \phi)\|_2^2 \ge \mathcal{L}_{\text{down}}(\theta, \phi) - \mathcal{L}^*_{\text{down}}$$

Putting this all together,

$$\begin{aligned}
\mathcal{L}_{\text{down}}(\theta_{k+1}, \phi_{k+1}) - \mathcal{L}^*_{\text{down}} &\le \mathcal{L}_{\text{down}}\left(\theta_k, \phi_k\right) - \mathcal{L}^*_{\text{down}} - \frac{1}{2\beta}\|\nabla \mathcal{L}_{\text{down}}\left(\theta_k, \phi_k\right)\|_2^2 \\
&\le \mathcal{L}_{\text{down}}\left(\theta_k, \phi_k\right) - \mathcal{L}^*_{\text{down}} - \frac{\mu}{\beta}\left(\mathcal{L}_{\text{down}}\left(\theta_k, \phi_k\right) - \mathcal{L}^*_{\text{down}}\right) \\
&= \left(1 - \frac{\mu}{\beta}\right)\left(\mathcal{L}_{\text{down}}\left(\theta_k, \phi_k\right) - \mathcal{L}^*_{\text{down}}\right).
\end{aligned}$$

So, applying this bound repeatedly gives us

$$\mathcal{L}_{\text{down}}(\theta_k, \phi_k) - \mathcal{L}^*_{\text{down}} \le \left(1 - \frac{\mu}{\beta}\right)^k \left(\mathcal{L}_{\text{down}}\left(\theta_0, \phi_0\right) - \mathcal{L}^*_{\text{down}}\right). \tag{9}$$

For simplicity, we denote the downstream loss after k-step optimization $\mathcal{L}_{\text{down}}(\theta_k, \phi_k)$ as $\mathcal{L}_{\text{down}}(\theta, \phi)$. Therefore, we obtain the following inequality.

$$\mathcal{L}_{\text{down}}(\theta, \phi) - \mathcal{L}^*_{\text{down}} \le O\left(\mathcal{L}_{\text{down}}\left(\theta_0, \phi_0\right) - \mathcal{L}^*_{\text{down}}\right). \tag{10}$$

By combining Theorem 2 and Theorem 3, we can deduce that our two-stage approach has a stronger guarantee of converging to the optimum of the downstream task compared to naïve fine-tuning, since $\mathcal{L}_{\text{down}}\left(\theta_0, \phi_0\right) \ge \mathcal{L}_{\text{down}}\left(\theta'_0, \phi_0\right)$.

Theoretical Connection between Representation Consistency and other metrics. In this section, we further explore the connection between the representation consistency and other metrics. Here, we first analyze the connection between representation consistency and the mean difference of distribution in Figure 4. An intuitive idea that comes to mind is that we can utilize the representation consistency as a metric to boost downstream performance if it is beneficial. Given the representation consistency, we are curious whether a larger representation consistency could indeed benefit the downstream performance. Motivated by Figure 4, we think that a better downstream performance can be achieved by pulling apart the representation similarity distributions of negative pairs (in pre-training task) with the same label and different labels (*i.e.*, the mean difference annotated in Figure 4). Therefore, we here establish a theoretical connection between the representation consistency and the mean difference of the above-mentioned distributions.

Theorem 4 (Connection between representation consistency and mean difference of distributions in Figure 4)**.** *Let $\mathcal{C}_\mathrm{R}(\mathcal{H},\mathcal{D},\mathcal{P})$ be the representation consistency. Given that the pre-training task is contrastive learning, we have*

$$\frac{1}{2(|V|-1)}\Delta' + \frac{|V|-5}{|V|-1} \geq \mathcal{C}_\mathrm{R}(\mathcal{H},\mathcal{D},\mathcal{P}) \geq \frac{1}{2(|V|-1)}\Delta' - \frac{|V|-5}{|V|-1}, \tag{11}$$

*where Δ' represents the mean difference in two representation similarity distributions of negative pairs (in pre-training task), whose two nodes are the same class and different classes in the downstream task, respectively (*i.e*, the mean difference annotated in Figure 4), and $|V|$ denotes the number of nodes in the downstream graph.*

Proof of Theorem 4. Let $\mathcal{C}_\mathrm{R}(\mathcal{H},\mathcal{D},\mathcal{P})$ be the representation consistency. $\Delta' = \mathbb{E}_{\boldsymbol{n}}[\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{P}}(\boldsymbol{n}) = 0, y^*_{\mathcal{D}}(\boldsymbol{n}) = 1] - \mathbb{E}_{\boldsymbol{n}}[\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{P}}(\boldsymbol{n}) = 0, y^*_{\mathcal{D}}(\boldsymbol{n}) = 0]$, where $\boldsymbol{n} \in \mathcal{V} \times \mathcal{V}$. Given that the pre-training task is contrastive learning, we can deduce that

$$\begin{aligned}\mathcal{C}_\mathrm{R}(\mathcal{H},\mathcal{D},\mathcal{P}) =&\mathbb{E}_{\boldsymbol{n}}[\rho\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{P}}(\boldsymbol{n}) = y^*_{\mathcal{D}}(\boldsymbol{n})]\\ =&\mathbb{E}_{\boldsymbol{n}}[\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{P}}(\boldsymbol{n}) = y^*_{\mathcal{D}}(\boldsymbol{n}), y^*_{\mathcal{P}}(\boldsymbol{n}) = 1]\\ &- \mathbb{E}_{\boldsymbol{n}}[\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{P}}(\boldsymbol{n}) = y^*_{\mathcal{D}}(\boldsymbol{n}), y^*_{\mathcal{P}}(\boldsymbol{n}) = 0].\end{aligned} \tag{12}$$

Since contrastive learning takes two different (same) nodes as a negative (positive) pair, we have

$$\mathbb{E}_{\boldsymbol{n}}[\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{P}}(\boldsymbol{n}) = 1, y^*_{\mathcal{P}}(\boldsymbol{n}) = 1] = |V|/\binom{|V|}{2} = 2/(|V|-1),$$

$$\begin{aligned}\mathbb{E}_{\boldsymbol{n}}[\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{P}}(\boldsymbol{n}) = 0]) \geq&\mathbb{E}_{\boldsymbol{n}}[-1|y^*_{\mathcal{P}}(\boldsymbol{n}) = 0])\\ =&-(1-|V|/\binom{|V|}{2}) = \frac{|V|-3}{|V|-1}.\end{aligned}$$

Accordingly, we can obtain

$$\begin{aligned}\mathcal{C}_\mathrm{R}(\mathcal{H},\mathcal{D},\mathcal{P}) =&\mathbb{E}_{\boldsymbol{n}}[\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{P}}(\boldsymbol{n}) = y^*_{\mathcal{D}}(\boldsymbol{n}), y^*_{\mathcal{P}}(\boldsymbol{n}) = 1]\\ &- \mathbb{E}_{\boldsymbol{n}}[\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{P}}(\boldsymbol{n}) = y^*_{\mathcal{D}}(\boldsymbol{n}), y^*_{\mathcal{P}}(\boldsymbol{n}) = 0]\\ =&\frac{2}{|V|-1}\mathbb{E}_{\boldsymbol{n}}[\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{D}}(\boldsymbol{n}) = 0]\\ &- \frac{|V|-3}{|V|-1}\mathbb{E}_{\boldsymbol{n}}[\mathrm{Sim}(h(\boldsymbol{n}))|y^*_{\mathcal{D}}(\boldsymbol{n}) = 1]\\ \geq&\frac{1}{2(|V|-1)}\Delta' - \frac{|V|-5}{|V|-1},\end{aligned} \tag{13}$$

where $\boldsymbol{n} \in \mathcal{V} \times \mathcal{V}$. In Eq. (12), Δ' represent the expectation difference on the marginal distribution of representation similarity of the same label pair and the different label pair when $y^*_{\mathcal{P}}(\boldsymbol{n}) = 0$.

Besides, we can further deduce the upper bound for representation consistency.

$$\begin{aligned}\mathcal{C}_\mathrm{R}(\mathcal{H},\mathcal{D},\mathcal{P}) =&\mathbb{E}_{\boldsymbol{n}}[Sim(h(\boldsymbol{n}))|y^*_{\mathcal{P}}(\boldsymbol{n}) = y^*_{\mathcal{D}}(\boldsymbol{n}), y^*_{\mathcal{P}}(\boldsymbol{n}) = 1]\\ &- \mathbb{E}_{\boldsymbol{n}}[Sim(h(\boldsymbol{n}))|y^*_{\mathcal{P}}(\boldsymbol{n}) = y^*_{\mathcal{D}}(\boldsymbol{n}), y^*_{\mathcal{P}}(\boldsymbol{n}) = 0]\\ =&\frac{2}{|V|-1}\mathbb{E}_{\boldsymbol{n}}[Sim(h(\boldsymbol{n}))|y^*_{\mathcal{D}}(\boldsymbol{n}) = 0]\\ &- \frac{|V|-3}{|V|-1}\mathbb{E}_{\boldsymbol{n}}[Sim(h(\boldsymbol{n}))|y^*_{\mathcal{D}}(\boldsymbol{n}) = 1]\\ \leq&\frac{1}{2(|V|-1)}\Delta' + \frac{|V|-5}{|V|-1},\end{aligned} \tag{14}$$

The above theorem suggests that a high representation consistency cannot be achieved without a large difference between the representation similarity distributions of negative pairs with the same label and different labels. This provides clues on how to improve the downstream performance.

Then, we analyze the connection between the representation consistency and inter-class distance (Gu, Li, and Han 2012). We find that the lower bound of representation consistency can be expressed by inter-class distance.

Theorem 5 (Theoretical connection between representation consistency and inter-class distance.)**.** *Let $G_{down} = \{V, E\}$ denote the downstream graph. $\mathcal{C}_\mathrm{R}(\mathcal{H},\mathcal{D},\mathcal{P})$ is the representation consistency, and d_{inter} is the inter-class distance. The lower bound of $\mathcal{C}_\mathrm{R}(\mathcal{H},\mathcal{D},\mathcal{P})$ presents an expression of d_{inter},* i.e.*,*

$$\mathcal{C}_\mathrm{R}(\mathcal{H},\mathcal{D},\mathcal{P}) \geq \frac{1}{4(|V|-1)}d_{inter} - \frac{|V|-5}{|V|-1}.$$

Proof. Let $G_\mathrm{down} = \{V, E\}$ denote the downstream graph. Each node v_i is associated with node representation $h(v_i)$ from model and label $y(v_i) \in \mathcal{Y} = \{c_0, \ldots, c_N\}$, where $\mathcal{Y}$ is the label space and N is the number of classes in $\mathcal{Y}$. The set of nodes with the label $c \in \mathcal{Y}$ is denoted as V_c. We assume that the representation of the node in V_c obeys the multivariate normal distribution, that is the node representation $h_c \sim \mathcal{N}(\mu_c, \Sigma_c)$. In this way, d_{inter} can be expressed as $\sum_{c \neq c' \& c, c' \in \mathcal{Y}} |(\mu_c - \mu_{c'})^{\mathrm{T}}(\mu_c - \mu_{c'})|$.

Now we discuss the expectation of cosine similarity between nodes of different labels c and c' $(c \neq c')$. Accordingly, $h_c \sim \mathcal{N}(\mu_c, \Sigma_c)$ and $h_{c'} \sim \mathcal{N}(\mu_{c'}, \Sigma_{c'})$. To simplify the derivation, all the representation vectors are normalized to 1. In this way, we have $||h_c|| = 1$ and $||\mu_c|| = |\mathbb{E}(h_c)| \leq \mathbb{E}(||h_c||) = 1$ for Jensen inequality. The derivation is shown below:

$$\begin{aligned} \mathbb{E}(\mathrm{Sim}(h_c, h_{c'})) =& \mathbb{E}(h_c^{\mathrm{T}}\, h_{c'}/||h_c||\,||h_{c'}|| = \mathbb{E}(h_c^{\mathrm{T}}\, \mathbf{I}\, h_{c'}) \\ =& \mu_c^{\mathrm{T}}\, \mathbf{I}\, \mu_{c'} + Tr(\mathbf{I} * \mathbf{Cov}(h_c, h_{c'})) \\ =& \mu_c^{\mathrm{T}}\, \mu_{c'} + Tr(\mathbf{Cov}(h_c, h_{c'})). \end{aligned} \tag{15}$$

The similarity divergence $D(c, c')$ between same label pair (c, c) and different label pair (c, c') can be expressed as follows:

$$\begin{aligned} D(c, c') =& |\mathbb{E}(Sim(h_c, h_{c'})) - \mathbb{E}(Sim(h_c, h_c))| \\ =& |\mu_c^{\mathrm{T}} \mu_{c'} + Tr(\mathbf{Cov}(h_c, h_{c'})) \\ & - \mu_c^{\mathrm{T}} \mu_c - tr(\mathbf{Cov}(h_c, h_c))| \\ =& |\mu_c^{\mathrm{T}}(\mu_{c'} - \mu_c) + Tr(\mathbf{Cov}(h_c, h_{c'}) - \mathbf{Cov}(h_c, h_c))|. \end{aligned} \tag{16}$$

The similarity divergence object should also be symmetrical that $D(c, c')$ and $D(c', c)$ should be the same. Therefore, the symmetric similarity divergence D' is shown as follows:

$$\begin{aligned} D'(c, c') =& 1/2(D(c, c') + D(c', c)) \\ =& 1/2(|\mu_c^{\mathrm{T}}(\mu_{c'} - \mu_c) + Tr(\mathbf{Cov}(h_c, h_{c'}) - \mathbf{Cov}(h_c, h_c))| \\ & + |\mu_{c'}^{\mathrm{T}}(\mu_c - \mu_{c'}) + Tr(\mathbf{Cov}(h_{c'}, h_c) - \mathbf{Cov}(h_{c'}, h_c'))|) \\ =& 1/2(|\mu_c^{\mathrm{T}}(\mu_{c'} - \mu_c) + Tr(\mathbf{Cov}(h_c, h_{c'}) - \mathbf{Cov}(h_c, h_c))| \\ & + |\mu_{c'}^{\mathrm{T}}(\mu_c - \mu_{c'}) + Tr(\mathbf{Cov}(h_{c'}, h_c) - \mathbf{Cov}(h_{c'}, h_{c'}))|) \\ \geq& 1/2|(\mu_c - \mu_{c'})^{\mathrm{T}}(\mu_c - \mu_{c'}) \\ & + Tr(-2\mathbf{Cov}(h_c, h_{c'}) + \Sigma_c + \Sigma_{c'})|. \end{aligned} \tag{17}$$

For simplicity, we just ignore the Σ part since the feature is extracted from the same model and the correlation between each dimension of feature is similar. In this way, $\Sigma_c = \Sigma_{c'} = \mathbf{Cov}(h_c, h_{c'})$,

$$\begin{aligned} D'(c, c') =& 1/2(D(c, c') + D(c', c)) \\ \geq& 1/2|(\mu_c - \mu_{c'})^{\mathrm{T}}(\mu_c - \mu_{c'}) - Tr(2\mathbf{Cov}(h_c, h_{c'}) + \Sigma_c + \Sigma_{c'})| \\ \geq& 1/2|(\mu_c - \mu_{c'})^{\mathrm{T}}(\mu_c - \mu_{c'})|. \end{aligned} \tag{18}$$

The combination of different label pairs that denotes the divergence of cosine similarity which is the same as the mean difference Δ' is as follows:

$$\begin{aligned} \Delta' =& \sum_{c, c' \in \mathcal{Y}} D'(c, c') \\ =& \sum_{c, c' \in \mathcal{Y} \& c = c'} D'(c, c') + \sum_{c, c' \in \mathcal{Y} \& c \neq c'} D'(c, c') \\ \geq& 1/2 \sum_{c, c' \in \mathcal{Y} \& c \neq c'} |(\mu_c - \mu_{c'})^{\mathrm{T}}(\mu_c - \mu_{c'})| = d_{\text{inter}}. \end{aligned} \tag{19}$$

By Theorem 4 proved above, $\mathcal{C}_{\mathrm{R}}(\mathcal{H}, \mathcal{D}, \mathcal{P}) \geq \frac{1}{2(|V|-1)}\Delta' - \frac{|V|-5}{|V|-1}$. And the left hand side is the objective for mean difference. Thus, we can get that:

$$\begin{aligned} \mathcal{C}_{\mathrm{R}}(\mathcal{H}, \mathcal{D}, \mathcal{P}) \geq& \frac{1}{2(|V|-1)}\Delta' - \frac{|V|-5}{|V|-1} \\ \geq& \frac{1}{4(|V|-1)} d_{\text{inter}} - \frac{|V|-5}{|V|-1}. \end{aligned} \tag{20}$$

From Eq. (20), we can gain insight that the left hand side is the objective for Bridge-Tune which is representation consistency. The right hand side is the inter-class distance between each label, which means the inter-class distance can also be feasible to improve the representation consistency by optimizing the lower bound.