Rethinking Dimensional Rationale in Graph Contrastive Learning from Causal Perspective

Qirui Ji[1 3 *], **Jiangmeng Li**[1 2 *†], **Jie Hu**[4], **Rui Wang**[1 2 3], **Changwen Zheng**[1 3], **Fanjiang Xu**[1 3]

[1]Science & Technology on Integrated Information System Laboratory, Institute of Software Chinese Academy of Sciences
[2]State Key Laboratory of Intelligent Game
[3]University of Chinese Academy of Sciences
[4]State Key Laboratory of Computer Science, Institute of Software Chinese Academy of Sciences
{jiqirui2022, jiangmeng2019, wangrui, changwen, fanjiang}@iscas.ac.cn, hujie@ios.ac.cn

Abstract

Graph contrastive learning is a general learning paradigm excelling at capturing invariant information from diverse perturbations in graphs. Recent works focus on exploring the *structural* rationale from graphs, thereby increasing the discriminability of the invariant information. However, such methods may incur in the mis-learning of graph models towards the interpretability of graphs, and thus the learned noisy and task-agnostic information interferes with the prediction of graphs. To this end, with the purpose of exploring the intrinsic rationale of graphs, we accordingly propose to capture the *dimensional* rationale from graphs, which has not received sufficient attention in the literature. The conducted exploratory experiments attest to the feasibility of the aforementioned roadmap. To elucidate the innate mechanism behind the performance improvement arising from the dimensional rationale, we rethink the dimensional rationale in graph contrastive learning from a causal perspective and further formalize the causality among the variables in the pre-training stage to build the corresponding structural causal model. On the basis of the understanding of the structural causal model, we propose the dimensional rationale-aware graph contrastive learning approach, which introduces a learnable dimensional rationale acquiring network and a redundancy reduction constraint. The learnable dimensional rationale acquiring network is updated by leveraging a bi-level meta-learning technique, and the redundancy reduction constraint disentangles the redundant features through a decorrelation process during learning. Empirically, compared with state-of-the-art methods, our method can yield significant performance boosts on various benchmarks with respect to discriminability and transferability. The code implementation of our method is available at https://github.com/ByronJi/DRGCL.

1 Introduction

Graph contrastive learning (GCL) is a general learning paradigm excelling at seeking to understand invariant information from diverse perturbations in graphs (You et al. 2020; Suresh et al. 2021; You et al. 2021; Xia et al. 2022). However, most of these methods focus on building sophisticated data augmentations for GCL, while the intrinsic interpretability in graph representations is not explored, such that the theoretical guarantee for the performance improvement arising from adopting such approaches is insufficient, and the model trained by following these methods may learn stochastic noisy and task-agnostic information, thereby confusing the prediction on downstream tasks. Therefore, the graph rationale exploration is provoked to understand the knowledge driving the model to make certain predictions (Wu et al. 2022), where rationale is a specific subset of graph features, e.g., graph structure, which can guide or explain the model's predictions (Ying et al. 2019). Successes achieved by RGCL (Li et al. 2022c) demonstrate that exploring rationales in graphs can indeed promote the model to learn discriminative representations in GCL. RGCL focuses on exploring the *structural* rationale (SR) from graphs, i.e., the structure containing specific edges or nodes that are correlated with the prediction of graphs. However, the features contained by the nodes or messages passing through the edges may still include certain discriminative information. Thus, arbitrarily removing or assigning weights to the graph structure can undermine the discriminability of the learned representation. Concretely, we raise a crucial question:

"Does there exist a manner to explore the intrinsic rationale in graphs, thereby improving the GCL predictions?"

With the question in mind, we conduct exploratory experiments with GraphCL (You et al. 2020) on the biochemical molecule dataset PROTEINS, and the social network dataset REDDIT-BINARY (RDT-B), where we randomly preserve certain dimensions, i.e., a subset of the representations, while blocking the others. The experimental results are illustrated in Figure 1. We observe from the results and find that the graph representations only preserving specific dimensions indeed achieve better performance than the primitive representations, and such dimensions are treated as *dimensional* rationales (DRs) for the graph. The exploratory experiments jointly prove the existence of DRs and the positive effects of specific DRs in the prediction of GCL. The intuition behind the experimental exploration is that compared with the SR, the DR is intrinsic to the representations learned by GCL methods, which can tackle the long-standing issue of the SR and further achieve the desideratum that jointly preserves the discriminative information and blocks the task-agnostic information of representations. We provide theoret-

*These authors contributed equally.
†Corresponding author.

ical analysis to demonstrate that the proposed DR can degenerate into the conventional SR for GCL.

However, the innate mechanism of the performance improvement brought about by introducing the DR is not sufficiently explored, such that we rethink the DR of graphs from the causal perspective and accordingly develop a structural causal model (SCM). By understanding the SCM, we disclose a counter-intuitive conclusion: the acquired graph DR is determined as a causal confounder in GCL. The reason is that the principle of unsupervised learning incurs the inconsistent variation of the acquired graph DR, such that the acquired rationale may improve or degenerate the model performance on downstream tasks, and inspired by the theory of causal inference (Pearl et al. 2000; Glymour, Pearl, and Jewell 2016), we assert that the acquired graph DR is a causal confounder. The theoretical analysis can further be proved by the empirical evidence. Accordingly, as shown in Figure 1, we observe that the points with different preserved dimension rates are scattered around the baseline dashes, which proves that the inconsistency of the graph DRs may improve or degenerate the model performance.

To this end, we intuitively propose the *Dimensional Rationale-aware Graph Contrastive Learning*, namely *DRGCL*, which initially acquires the graph DRs and further adopts the backdoor adjustment technique (Glymour, Pearl, and Jewell 2016). Specifically, we introduce the learnable graph DR acquiring network, which is trained by adopting a bi-level meta-learning technique. To extend the representation space of the acquired DR, we apply the graph DR redundancy reduction as a regularization term during training. We provide solid theoretical analyses to prove the validity of DRGCL. Empirically, we compare our method with various baselines on benchmark graph datasets, which further demonstrates the effectiveness of DRGCL. **Contributions**:

- We present a heuristic experiment to demonstrate the existence of the graph DR and further provide theoretical analysis to prove that compared with the conventional graph SR, the graph DR is more intrinsic to GCL.
- We formalize the mechanism of introducing DRs by building an SCM and demonstrate that the acquired DR is a causal confounder in GCL with sufficient theoretical and empirical evidence.
- We propose DRGCL to acquire redundancy-against DRs and perform the backdoor adjustment on SCM, thereby consistently improving GCL performance.
- We provide solid theoretical and experimental analyses, which jointly demonstrate the effectiveness of our method in terms of discriminability and transferability.

2 Related Works

2.1 Graph Contrastive Learning

Many methods have been used to study graph-level contrastive learning. GraphCL (You et al. 2020) designs four types of general augmentations for GCL. ADGCL (Suresh et al. 2021) optimizes adversarial graph augmentation strategies to prevent Graph Neural Networks (GNNs) from capturing redundant information. JOAO (You et al. 2021) selects augmentation pairs in GraphCL by an automated approach to solve the trial-and-errors. SimGRACE (Xia et al. 2022) utilizes an original GNN model and its perturbed version as encoders to obtain correlated views further avoiding the cost of trial-and-errors. RGCL (Li et al. 2022c) uses GNNExplainer (Ying et al. 2019) to find invariant SRs to dig discriminative information. Our method applies DRs to our models, which can find more intrinsic discriminability.

2.2 Graph Rationalization

Rationalization in Graphs has two research directions: post-hoc explainability and intrinsic interpretability. Post-hoc explainability uses separate methods (Ying et al. 2019; Tan et al. 2022a) to attribute predictions to the input graph. Intrinsic interpretability integrates a rationalization module, e.g., graph attention (Velickovic et al. 2017; Wu et al. 2022) or graph pooling (Lee, Lee, and Kang 2019; Ranjan, Sanyal, and Talukdar 2020), into the model. The rationalization module employs soft or hard masks on the input graph to guide the model's decisions. While existing methods use SRs from masked subgraphs to train the GNN, our method directly captures DRs within the graph embeddings.

2.3 Causal Inference

Causal inference (Pearl et al. 2000; Glymour, Pearl, and Jewell 2016) has been widely applied in computer science through deconfounding and counterfactual inference. Deconfounding methods (Li et al. 2022b; Gao et al. 2022; Qiang et al. 2022, 2023) estimate the direct causal effect behind confounders. Counterfactual inference (Tan et al. 2022b) aims at finding the smallest change in the input which affects the model's prediction. Our work introduces the dimension confounder in GCL with an SCM. Guided by SCM, we utilize the backdoor adjustment to obtain the direct causal effect between the learned embedding and the predicted label.

3 Problem Formulations

3.1 Graph Contrastive Learning

Given a graph $\mathcal{G}$ sampled from the dataset of M graphs, denoted as $\mathcal{G} \in \{\mathcal{G}_m : m \in M\}$, we formulated the augmented graph $\hat{\mathcal{G}}$ by applying the augmentation distribution $\mathcal{T}(\hat{\mathcal{G}}|\mathcal{G})$. During pre-training, we sample a minibatch of N graphs from $\mathcal{G}_m$ and denote it as $\mathcal{G}' = \{\mathcal{G}_n\}_{n=1}^{N}$, where $\mathcal{G}_n$ represents the n-th sample. We perform stochastic data augmentations to transform each sample $\mathcal{G}_n$ into two augmented views $\hat{\mathcal{G}}_{n,i}$ and $\hat{\mathcal{G}}_{n,j}$. Then $\hat{\mathcal{G}}_{n,i}$ and $\hat{\mathcal{G}}_{n,j}$ are fed into a feature extractor to get their feature representations $\boldsymbol{z}_{n,i}$ and $\boldsymbol{z}_{n,j}$. Then a GCL loss function is defined to enforce maximizing the consistency between positive pairs $\boldsymbol{z}_{n,i}$, $\boldsymbol{z}_{n,j}$, such as InfoNCE loss (You et al. 2020; Xia et al. 2022):

$$\mathcal{L}_{IN} = \sum_{n=1}^{N} -\log \frac{\exp\left(d\left(\boldsymbol{z}_{n,i}, \boldsymbol{z}_{n,j}\right)/\tau\right)}{\sum_{n'=1,n'\neq n}^{N} \exp\left(d\left(\boldsymbol{z}_{n,i}, \boldsymbol{z}_{n',j}\right)/\tau\right)}, \quad (1)$$

where τ denotes the temperature parameter and $d(\cdot,\cdot)$ denotes the cosine similarity function.

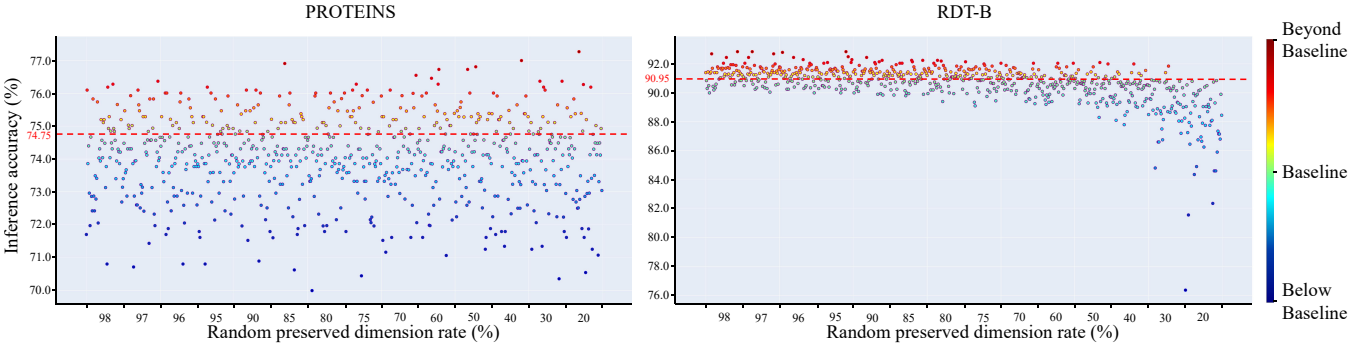

Figure 1: Experimental scatter diagrams obtained by GraphCL with randomly preserving dimensions on PROTEINS and RDT-B datasets. The red dashed lines denote the performance achieved by the primitive representation of GraphCL. The colored scattered points denote the downstream classification performance of embeddings with certain dimensions preserved. Note that the unreserved dimensions are directly valued by 0. The experimental principle emerges from the intuition that the prediction on downstream tasks may be significantly affected if the multi-dimensional representations are perturbed.

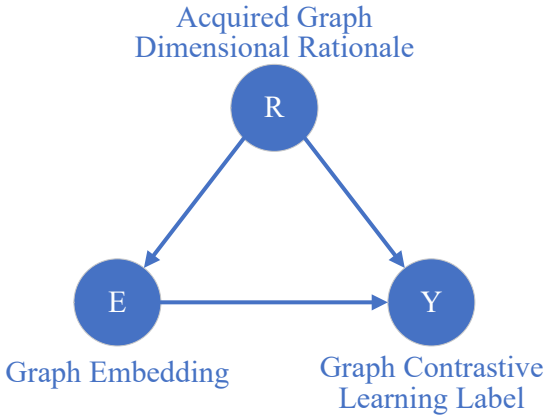

Figure 2: SCM for GCL pretraining.

3.2 Structural Causal Model

The intuition of our work comes from the investigation of the effects of preserving different DRs in the graph embedding, as shown in Figure 1. According to (Wang et al. 2022), the cross-entropy loss can be bounded by the contrastive loss, indicating that we can improve the performance on downstream classification tasks by enhancing the discriminability of representations learned by GCL during pre-training. Thus, improving the quality of the acquired DRs during pre-training derives the indirect promotion of the DRs acquired on downstream tasks. To this end, we establish an SCM to elaborate the causality among the variables in GCL: graph embedding E, acquired graph DR R, and graph contrastive label Y. The SCM is depicted in Figure 2, with each link representing a causality between two variables:

- $E \rightarrow Y$. The graph embedding E can directly affect the graph contrastive label Y.
- $R \rightarrow E$. The acquired graph DR R affects the learned features of the graph embeddings by contributing to the gradient effect of the training of the graph encoder, which further affects the graph embedding E.
- $R \rightarrow Y$. In GCL, the graph contrastive learning label is related to the anchor, such that the acquired graph DR R causally affects the anchor due to the proposed iterative training paradigm of DRGCL.

According to the SCM, the acquired graph DR R is the causal confounder between E and Y due to the causal effects of R towards E and Y.

3.3 Causal Intervention via Backdoor Adjustment

(Pearl et al. 2000) proposes the definition of the backdoor path to demarcate the scope of application of the backdoor criterion. In our SCM, there exists a backdoor path $E \leftarrow R \rightarrow Y$, resulting in the spurious correlation between E and Y. Then, if we use $P(Y|E)$ to measure the causality between E and Y as the approach adopted by the conventional GCL methods, the task-irrelevant features would affect the downstream classification. To eliminate the causal effect of the backdoor path, we can intervene on the variable E and condition on the confounding factor R. The adjustment formula can be written as follows:

$$P(Y|do(E)) = \sum_{r} P(Y|E, R = r)P(R = r), \qquad (2)$$

where r denotes the value of R.

4 Methodology

In this paper, we focus on developing a novel GCL learning framework which is shown in Figure 3.

4.1 Graph Dimensional Rationale Acquiring Network

By following the data augmentations in GraphCL (You et al. 2020), we sample two transformations t_1 and t_2 from the augmentation distribution $\mathcal{T}$ and further obtain two correlated views $\hat{\mathcal{G}}_{n,i}$ and $\hat{\mathcal{G}}_{n,j}$. Then we feed them into the GNN-based encoder $f_\theta(\cdot)$ to extract graph-level representations $\boldsymbol{h}_{n,i}$, $\boldsymbol{h}_{n,j}$. To acquire the DRs from the candidate graphs, we introduce a learnable DR weight, denoted as

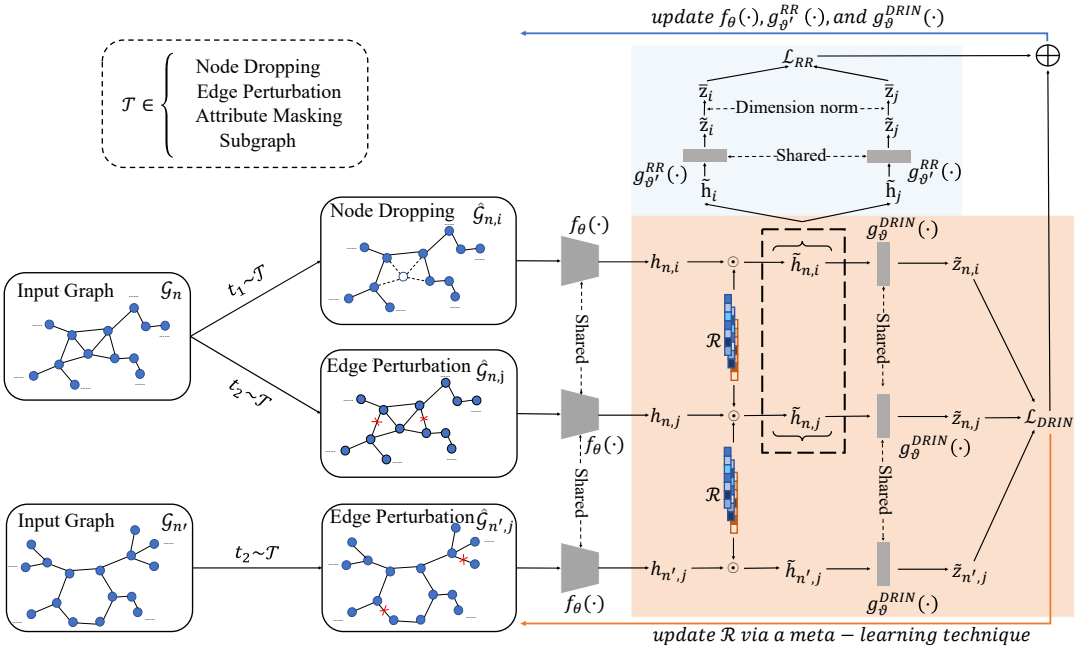

Figure 3: Illustration of DRGCL. The solid blue line pointing backwards represents the regular training step. The solid red line pointing backwards represents the meta-learning step.

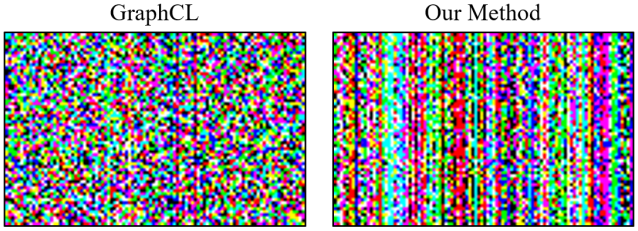

Figure 4: The visualization of the representations learned by GraphCL and our method using the redundancy reduction method on the BBBP dataset, respectively. The learned features are projected into a colored image in RGB format, where different colors represent different types of features. The abscissa axis represents the feature dimensions, and the ordinate axis represents samples of different classes. The greater the color contrast, the lower the dimensional feature similarity. These plots represent the similarity between dimension features with the first 64 samples of BBBP.

$\mathcal{R} = \{\boldsymbol{\omega}_k | k \in [\![1, D]\!]\}$, where D represents the number of graph embedding dimensions, which is treated as containing the shared knowledge with the acquired DR.

$$\tilde{\boldsymbol{h}} = \boldsymbol{h} \odot \mathcal{R}, \tag{3}$$

where $\tilde{\boldsymbol{h}}$ denotes the representation derived by preserving the acquired DR, and $\odot$ represents an element-wise product operation. By utilizing this operation, we obtain $\tilde{\boldsymbol{h}}_{n,i}$ and $\tilde{\boldsymbol{h}}_{n,j}$. Furthermore, we utilize a projection head $g_\vartheta^{DRIN}(\cdot)$ to map the graph representations into a latent space:

$$\tilde{\boldsymbol{z}}_{n,i} = g_\vartheta^{DRIN}(\tilde{\boldsymbol{h}}_{n,i}), \tilde{\boldsymbol{z}}_{n,j} = g_\vartheta^{DRIN}(\tilde{\boldsymbol{h}}_{n,j}). \tag{4}$$

Subsequently, we utilize the DR-aware InfoNCE loss as

$$\mathcal{L}_{DRIN} = \sum_{n=1}^{N} -\log \frac{\exp\left(d\left(\tilde{\boldsymbol{z}}_{n,i}, \tilde{\boldsymbol{z}}_{n,j}\right)/\tau\right)}{\sum_{n'=1, n' \neq n}^{N} \exp\left(d\left(\tilde{\boldsymbol{z}}_{n,i}, \tilde{\boldsymbol{z}}_{n',j}\right)/\tau\right)}. \tag{5}$$

4.2 Graph Dimensional Rationale Redundancy Reduction

From the perspective of information theory, each dimension captures a subset of the information entropy of the graph representation. As depicted in Figure 4, GraphCL encounters the issue of graph dimensional redundancy, which indicates that multiple dimensions in graph embeddings share overlapping information entropy. To tackle the issue, inspired by classical multivariate analysis methods (Hotelling 1992; Zhang et al. 2021), we apply the graph DR redundancy reduction to DRGCL. Following the aforementioned manner to get $\tilde{\boldsymbol{h}}$, we obtain $\tilde{\mathbf{h}}_i$, $\tilde{\mathbf{h}}_j$, which denote the representations from a minibatch of N graphs of two augmented views $\hat{\mathcal{G}}_i$ and $\hat{\mathcal{G}}_j$. Subsequently, we use a specific projection head $g_{\vartheta'}^{RR}(\cdot)$ to map the graph representations into a latent space:

$$\tilde{\mathbf{z}}_i = g_{\vartheta'}^{RR}(\tilde{\mathbf{h}}_i), \tilde{\mathbf{z}}_j = g_{\vartheta'}^{RR}(\tilde{\mathbf{h}}_j). \tag{6}$$

Then we apply an instance-dimensional normalization to ensure each feature dimension has a 0-mean and $1/\sqrt{N}$-standard deviation distribution, which is implemented as:

$$\bar{\mathbf{z}} = \frac{\tilde{\mathbf{z}} - \mu(\tilde{\mathbf{z}})}{\sigma(\tilde{\mathbf{z}}) * \sqrt{N}}. \tag{7}$$

Algorithm 1: The DRGRL training algorithm

Input: Graph dataset $\mathcal{G}_m$ with M graphs, minibatch size N, and a hyper-parameter α.
Initialize The neural network parameters: θ for $f_\theta(\cdot)$, ϑ for $g_\vartheta^{DRIN}(\cdot)$, ϑ' for $g_{\vartheta'}^{RR}(\cdot)$, and $\mathcal{R} = \{\boldsymbol{\omega}_k | k \in [\![1, D]\!]\}$. The learning rates: β_θ and β_ϑ, etc.
repeat
  for t-th training iteration **do**
    Iteratively sample a minibatch $\mathcal{G}'$ with N examples from $\mathcal{G}_m$, $\mathcal{G}' = \{\mathcal{G}_n : n = 1, 2, ...N\}$
    Randomly sample two augmentations t_1, t_2 from $\mathcal{T}$, the augmented views of $\mathcal{G}_n$ can be denoted as $\hat{\mathcal{G}}_{n,i}$ and $\hat{\mathcal{G}}_{n,j}$, the augmented views of $\mathcal{G}'$ can be denoted as $\hat{\mathcal{G}}'$, including $\hat{\mathcal{G}}_i$, $\hat{\mathcal{G}}_j$.
    for $n = 1$ to N **do**
      $\tilde{\boldsymbol{h}}_{n,i} = f_\theta(\hat{\mathcal{G}}_{n,i}) \odot \mathcal{R}$, $\tilde{\boldsymbol{h}}_{n,j} = f_\theta(\hat{\mathcal{G}}_{n,j}) \odot \mathcal{R}$
      $\tilde{\boldsymbol{z}}_{n,i} = g_\vartheta^{DRIN}(\tilde{\boldsymbol{h}}_{n,i})$, $\tilde{\boldsymbol{z}}_{n,j} = g_\vartheta^{DRIN}(\tilde{\boldsymbol{h}}_{n,j})$
    end for
    $\mathcal{L}_{DRIN} = \sum_{n=1}^{N} -\log \frac{\exp\left(d\left(\tilde{\boldsymbol{z}}_{n,i}, \tilde{\boldsymbol{z}}_{n,j}\right)/\tau\right)}{\sum_{n'=1, n' \neq n}^{N} \exp\left(d\left(\tilde{\boldsymbol{z}}_{n,i}, \tilde{\boldsymbol{z}}_{n',j}\right)/\tau\right)}$
    $\tilde{\mathbf{h}}_i = f_\theta(\hat{\mathcal{G}}_i) \odot \mathcal{R}$, $\tilde{\mathbf{h}}_j = f_\theta(\hat{\mathcal{G}}_j) \odot \mathcal{R}$
    $\tilde{\mathbf{z}}_i = g_{\vartheta'}^{RR}(\tilde{\mathbf{h}}_i)$, $\tilde{\mathbf{z}}_j = g_{\vartheta'}^{RR}(\tilde{\mathbf{h}}_j)$
    $\bar{\mathbf{z}}_i = \frac{\tilde{\mathbf{z}}_i - \mu(\tilde{\mathbf{z}}_i)}{\sigma(\tilde{\mathbf{z}}_i) * \sqrt{N}}$, $\bar{\mathbf{z}}_j = \frac{\tilde{\mathbf{z}}_j - \mu(\tilde{\mathbf{z}}_j)}{\sigma(\tilde{\mathbf{z}}_j) * \sqrt{N}}$
    $\mathcal{L}_{RR} = \underbrace{\mathcal{F}(\bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j)}_{\text{invariance term}} + \underbrace{\lambda(\mathcal{F}(\bar{\mathbf{z}}_i^\top \bar{\mathbf{z}}_i, \mathbf{I}) + \mathcal{F}(\bar{\mathbf{z}}_j^\top \bar{\mathbf{z}}_j, \mathbf{I}))}_{\text{decorrelation term}}$
    $\#\ regular\ training\ step,\ fix\ \mathcal{R}$
    $\arg\min_{\theta, \vartheta, \vartheta'} \mathcal{L}_{RR} + \alpha \cdot \mathcal{L}_{DRIN}$
    $\#\ compute\ trial\ weights\ and\ retain\ computational$
    $\#\ graph, fix\ \theta\ and\ \vartheta$
    $\theta_{trial} = \theta - \beta_\theta \nabla_\theta \mathcal{L}_{DRIN}\left(g_\vartheta^{DRIN}\left(f_\theta\left(\hat{\mathcal{G}}'\right) \odot \mathcal{R}\right)\right)$,
    $\vartheta_{trial} = \vartheta - \beta_\vartheta \nabla_\vartheta \mathcal{L}_{DRIN}\left(g_\vartheta^{DRIN}\left(f_\theta\left(\hat{\mathcal{G}}'\right) \odot \mathcal{R}\right)\right)$
    $\#\ meta\ training\ step\ using\ second\ derivative$
    $\arg\min_{\mathcal{R}} \mathcal{L}_{DRIN}\left(g_{\vartheta_{trial}}^{DRIN}\left(f_{\theta_{trial}}\left(\hat{\mathcal{G}}'\right) \odot \mathcal{R}\right)\right)$
  end for
until θ, ϑ, ϑ^{RR}, and $\mathcal{R}$ converge.

The obtained normalized $\bar{\mathbf{z}}_i$ and $\bar{\mathbf{z}}_j$ are further used to form the redundancy reduction loss for a certain graph as

$$\mathcal{L}_{RR} = \underbrace{\mathcal{F}(\bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j)}_{\text{invariance term}} + \underbrace{\lambda(\mathcal{F}(\bar{\mathbf{z}}_i^\top \bar{\mathbf{z}}_i, \mathbf{I}) + \mathcal{F}(\bar{\mathbf{z}}_j^\top \bar{\mathbf{z}}_j, \mathbf{I}))}_{\text{decorrelation term}}, \quad (8)$$

where $\mathcal{F}(\cdot, \cdot) = \|\cdot - \cdot\|_F^2$, $\|\cdot\|_F^2$ denotes the Frobenius norm and λ is a trade-off hyperparameter. Intuitively, the invariance term makes the embedding invariant to the distortions of a graph by minimizing the difference between two normalized representations. By trying to equate the off-diagonal elements of the auto-correlation matrix of each view's representation to 0, the decorrelation term reduces the redundancy between the representations, thereby avoiding the collapsed trivial solution outputting the same vector for all inputs. In Figure 4, it can be intuitively observed that adopting the redundancy reduction loss, our DRGCL can indeed learn representations with information-decoupled dimensions.

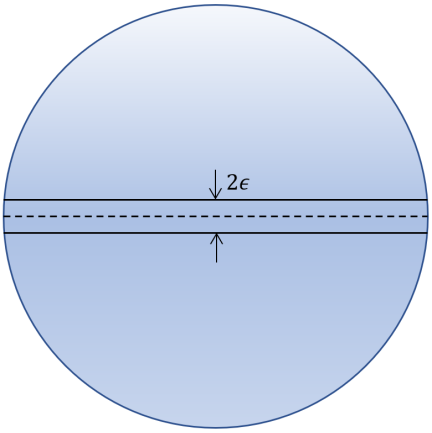

Figure 5: A counter-intuitive high-dimensional phenomenon in the problem of measuring concentration on a sphere. Almost the whole area of a high-dimensional sphere is concentrated in an ϵ-strip around its equator and actually around any great circle.

4.3 Dimensional Rationale-aware Graph Contrastive Learning with Backdoor Adjustment

During pre-training, a conventional training paradigm and a meta-learning training paradigm are iteratively employed. Specifically, we train the encoder $f_\theta(\cdot)$, and the projection heads $g_\vartheta^{DRIN}(\cdot)$ and $g_{\vartheta'}^{RR}(\cdot)$ in a conventional manner, while the DR weight $\mathcal{R}$ is trained by adopting the meta-learning objective. The overall training procedure of DRGCL consists of two steps. In the first training step, we follow the standard contrastive learning approach to train $f_\theta(\cdot)$, $g_\vartheta^{DRIN}(\cdot)$, and $g_{\vartheta'}^{RR}(\cdot)$. This involves jointly minimizing the contrastive loss and the redundancy reduction loss:

$$\mathcal{L}_{DRGCL} = \mathcal{L}_{RR} + \alpha \cdot \mathcal{L}_{DRIN}, \quad (9)$$

where α is a hyper-parameter that governs the trade-off between the two loss components. The second training step is based on meta-learning. We use a second-derivative technique (Liu, Davison, and Johns 2019) to solve a bi-level optimization problem. We encourage $\mathcal{R}$ to re-weight the specific dimensions to preserve task-relevant information, which is regarded as the DR for graph embeddings, so that DRGCL can perform the causal intervention via backdoor adjustment during training. Specifically, $\mathcal{R}$ is updated by computing its gradients with respect to the performance of $f_\theta(\cdot)$ and $g_\vartheta^{DRIN}(\cdot)$. The corresponding performance is measured by using the gradients of $f_\theta(\cdot)$ and $g_\vartheta^{DRIN}(\cdot)$ during the back-propagation of graph contrastive loss. Based on this updating mechanism during pre-training, the iterations of $\mathcal{R}$ can include sufficient values to perform the backdoor adjustment conditional on R with respect to E and Y. Formally, we update the DR weight $\mathcal{R}$ by

$$\arg\min_{\mathcal{R}} \mathcal{L}_{DRIN}\left(g_{\vartheta_{trial}}^{DRIN}\left(f_{\theta_{trial}}\left(\hat{\mathcal{G}}'\right) \odot \mathcal{R}\right)\right), \quad (10)$$

where $\hat{\mathcal{G}}'$, including $\hat{\mathcal{G}}_i$, $\hat{\mathcal{G}}_j$, denotes the augmented views of $\mathcal{G}'$, and $\mathcal{G}'$ is sampled from the graph dataset $\mathcal{G}_m$. θ_{trial} and

Dataset	NCI1	PROTEINS	DD	MUTAG	COLLAB	RDT-B	RDT-M5K	IMDB-B	A.R. ↓
GL	-	-	-	81.7 ± 2.1	-	77.3 ± 0.2	41.0 ± 0.2	69.9 ± 1.0	11.0
WL	**80.0** ± 0.5	73.0 ± 0.6	-	80.7 ± 3.0	-	68.8 ± 0.4	46.1 ± 0.2	**72.3** ± 3.4	8.3
DGK	**80.3** ± 0.5	73.3 ± 0.8	-	87.4 ± 0.7	-	78.0 ± 0.4	41.3 ± 0.2	67.0 ± 0.6	8.0
node2vec	54.9 ± 1.6	57.5 ± 3.6	-	72.6 ± 10.0	-	-	-	-	12.3
sub2vec	52.8 ± 1.5	53.0 ± 5.6	-	61.1 ± 15.8	-	71.5 ± 0.4	36.7 ± 0.4	55.3 ± 1.5	13.0
graph2vec	73.2 ± 1.8	73.3 ± 2.0	-	83.2 ± 9.3	-	75.8 ± 1.0	47.9 ± 0.3	71.1 ± 0.5	9.3
InfoGraph	76.2 ± 1.0	74.4 ± 0.3	72.9 ± 1.8	**89.0** ± 1.1	70.7 ± 1.1	82.5 ± 1.4	53.5 ± 1.0	**73.0** ± 0.9	5.8
GraphCL	77.9 ± 0.4	74.4 ± 0.5	**78.6** ± 0.4	86.8 ± 1.3	**71.4** ± 1.2	89.5 ± 0.8	**56.0** ± 0.3	71.2 ± 0.4	5.0
ADGCL	73.9 ± 0.8	73.3 ± 0.5	75.8 ± 0.9	88.7 ± 1.9	**72.0** ± 0.6	**90.1** ± 0.9	54.3 ± 0.3	70.2 ± 0.7	6.1
JOAO	78.1 ± 0.5	74.6 ± 0.4	77.3 ± 0.5	87.4 ± 1.0	69.5 ± 0.4	85.3 ± 1.4	55.7 ± 0.6	70.2 ± 3.1	6.5
JOAOv2	78.4 ± 0.5	74.1 ± 1.1	77.4 ± 1.2	87.7 ± 0.8	69.3 ± 0.3	86.4 ± 1.5	**56.0** ± 0.3	70.8 ± 0.3	5.8
RGCL	78.1 ± 1.0	**75.0** ± 0.4	**78.9** ± 0.5	87.7 ± 1.0	71.0 ± 0.7	**90.3** ± 0.6	**56.4** ± 0.4	71.9 ± 0.9	**3.3**
SimGRACE	**79.1** ± 0.4	**75.3** ± 0.1	77.4 ± 1.1	**89.0** ± 1.3	**71.7** ± 0.8	89.5 ± 0.9	55.9 ± 0.3	71.3 ± 0.8	**3.3**
DRGCL	78.7 ± 0.4	**75.2** ± 0.6	**78.4** ± 0.7	**89.5** ± 0.6	70.6 ± 0.8	**90.8** ± 0.3	**56.3** ± 0.2	**72.0** ± 0.5	**2.8**

Table 1: Unsupervised representation learning classification accuracy (%) on TU datasets (mean 10-fold cross-validation accuracy with 5 runs). A.R denotes the average rank of the results. The top-3 results are highlighted in bold.

Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BACE	AVG.
No Pre-Train	65.8 ± 4.5	74.0 ± 0.8	63.4 ± 0.6	57.3 ± 1.6	58.0 ± 4.4	71.8 ± 2.5	75.3 ± 1.9	70.1 ± 5.4	67.0
AttrMasking	64.3 ± 2.8	**76.7** ± 0.4	**64.2** ± 0.5	**61.0** ± 0.7	71.8 ± 4.1	**74.7** ± 1.4	77.2 ± 1.1	**79.3** ± 1.6	71.1
ContextPred	68.0 ± 2.0	**75.7** ± 0.7	**63.9** ± 0.6	60.9 ± 0.6	65.9 ± 3.8	**75.8** ± 1.7	77.3 ± 1.0	**79.6** ± 1.2	70.9
GraphCL	69.7 ± 0.7	73.9 ± 0.7	62.4 ± 0.6	60.5 ± 0.9	76.0 ± 2.7	69.8 ± 2.7	**78.5** ± 1.2	75.4 ± 1.4	70.8
ADGCL	68.3 ± 1.0	73.6 ± 0.8	63.1 ± 0.7	59.2 ± 0.9	77.6 ± 4.2	**74.9** ± 2.5	75.5 ± 1.3	75.0 ± 1.9	70.9
JOAO	70.2 ± 1.0	75.0 ± 0.3	63.0 ± 0.5	60.0 ± 0.8	**81.3** ± 2.5	71.7 ± 1.4	76.7 ± 1.2	77.3 ± 0.5	71.9
JOAOv2	**71.4** ± 0.9	74.2 ± 0.6	63.2 ± 0.5	60.5 ± 0.7	**81.0** ± 1.6	73.7 ± 1.0	77.5 ± 1.2	75.5 ± 1.3	**72.1**
RGCL‡	**71.4** ± 0.7	**75.2** ± 0.3	63.3 ± 0.2	**61.4** ± 0.6	76.4 ± 3.4	72.6 ± 1.5	**77.9** ± 0.8	76.0 ± 0.8	**71.8**
SimGRACE‡	**71.3** ± 0.9	73.9 ± 0.4	63.4 ± 0.5	60.6 ± 1.0	64.0 ± 1.2	69.4 ± 1.2	75.0 ± 1.1	74.6 ± 0.7	69.0
DRGCL	**71.2** ± 0.5	74.7 ± 0.5	**64.0** ± 0.5	**61.1** ± 0.8	**78.2** $\pm$ **1.5**	73.8 ± 1.1	**78.6** ± 1.0	**78.2** ± 1.0	**72.5**

Table 2: Transfer learning performance on molecular property prediction in ZINC-2M (mean ROC-AUC + std over 10 runs). AVG. denotes the average result in all datasets. ‡ means there exist differences in producing the results. RGCL finetunes ClinTox for 300 epochs and MUV for 50 epochs. For fairness, we reproduce them by finetuning for 100 epochs. SimGRACE only provides the results for BBBP, ToxCast, and SIDER. We provide the results of SimGRACE on other datasets in benchmarks.

ϑ_{trial} denote the $trial$ weights of the encoders and projection heads, respectively, after one gradient update using the contrastive loss defined in Equation 5. The update of these trial weights is formulated as follows:

$$\begin{aligned}\theta_{trial} &= \theta - \beta_\theta \nabla_\theta \mathcal{L}_{DRIN}\left(g_\vartheta^{DRIN}\left(f_\theta\left(\hat{\mathcal{G}}'\right) \odot \mathcal{R}\right)\right), \\ \vartheta_{trial} &= \vartheta - \beta_\vartheta \nabla_\vartheta \mathcal{L}_{DRIN}\left(g_\vartheta^{DRIN}\left(f_\theta\left(\hat{\mathcal{G}}'\right) \odot \mathcal{R}\right)\right),\end{aligned} \tag{11}$$

where β_θ and β_ϑ are learning rates. The intuition behind such a behavior is to leverage the second-derivative trick, which involves computing a derivative over the derivative of the combination θ, ϑ in order to update $\mathcal{R}$. Specifically, we compute the derivative with respect to $\mathcal{R}$ using a retained computational graph of θ, ϑ and then update the DR weight $\mathcal{R}$ by back-propagating this derivative as defined in Equation 10. Intuitively, the initialization of $\mathcal{R}$ is biased. During pre-training, $\mathcal{R}$ is updated per batch over epochs, resulting in the acquirement of local $\mathcal{R}$ with sufficient self-supervision for the current batch. After pre-training, all graphs have gradient contributions to $\mathcal{R}$, thereby achieving the global DR. The two steps for updating $f_\theta(\cdot)$, $g_\vartheta^{DRIN}(\cdot)$, $g_{\vartheta'}^{RR}(\cdot)$ and updating $\mathcal{R}$ are iteratively imposed until convergence. The Algorithm of the training pipeline is detailed in Algorithm 1.

For the fitting on downstream tasks, we utilize the graph DR-aware embeddings for downstream tasks.

5 Theoretical Analyses

5.1 Discussion on Relation between SR and DR

To facilitate comprehension, we recap the necessary preliminaries of GNN as follows. Suppose that $G = (\mathcal{V}, \mathcal{E})$ is a graph instance with the edge set $\mathcal{E}$ and the node set $\mathcal{V}$. The unified GNN framework follows a neighborhood aggregation strategy, where the representation of a node is iteratively updated by aggregating representations of its neighbors (Xu et al. 2019). After undergoing k iterations of aggregation, the representation of a node effectively captures the structural information present within its k-hop network neighborhood. Formally, the k-th layer of a GNN is

$$\begin{aligned}\boldsymbol{a}_v^{(k)} &= \mathrm{AGGREGATE}^{(k)}\left(\left\{\boldsymbol{h}_u^{(k-1)} : u \in \mathcal{N}(v)\right\}\right), \\ \boldsymbol{h}_v^{(k)} &= \mathrm{COMBINE}^{(k)}\left(\boldsymbol{h}_v^{(k-1)}, \boldsymbol{a}_v^{(k)}\right),\end{aligned} \tag{12}$$

Dataset	No PreTrain	AttrMasking	ContextPred	GraphCL	JOAO	JOAOv2	SimGRACE	**DRGCL**
PPI-306K	64.8 ± 2.0	65.2 ± 1.6	64.4 ± 1.3	$\mathbf{67.9} \pm 0.9$	64.4 ± 1.4	63.9 ± 1.6	$\mathbf{70.3} \pm 1.2$	$\mathbf{69.4} \pm 0.4$

Table 3: Transfer leaning performance on protein function prediction in biology PPI-306K dataset.

Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BACE	AVG.
w/o RR & DR	69.7 ± 0.7	73.9 ± 0.7	62.4 ± 0.6	60.5 ± 0.9	76.0 ± 2.7	69.8 ± 2.7	78.5 ± 1.2	75.4 ± 1.4	70.8
w/o RR	69.7 ± 0.7	74.7 ± 0.4	63.6 ± 0.5	59.9 ± 0.4	75.6 ± 3.5	72.2 ± 1.6	76.9 ± 0.8	75.1 ± 0.8	71.0
w/o DR	70.6 ± 0.8	74.3 ± 0.5	63.8 ± 0.5	60.3 ± 0.5	77.4 ± 1.4	73.8 ± 1.1	78.3 ± 1.0	76.8 ± 0.9	71.9
DRGCL	71.2 ± 0.5	74.7 ± 0.5	64.0 ± 0.5	61.1 ± 0.8	78.2 ± 1.5	73.8 ± 1.1	78.6 ± 1.0	78.2 ± 1.0	72.5

Table 4: Ablation study for DRGCL on downstream transfer learning.

where $\mathcal{N}(v)$ is a set of nodes adjacent to v, $\boldsymbol{a}_v^{(k)}$ is an aggregating representations of v's neighbors, $\boldsymbol{h}_v^{(k)}$ is the feature vector of node v at the k-th layer. For graph classification, the READOUT function aggregates node features from the final iteration to obtain the entire graph's representation $\boldsymbol{h}$:

$$\boldsymbol{h} = \text{READOUT}\left(\left\{\boldsymbol{h}_v^{(k)} \mid v \in G\right\}\right), \tag{13}$$

where READOUT can be a simple permutation invariant function such as summation or a more sophisticated graph-level pooling function.

Our DR applies a dimensional weight to the graph representation while the SR concentrates the rationale in message passing or node representation. Suppose the AGGREGATE, COMBINE, and READOUT functions are injective, then obviously the change of nodes is a degeneration or special solution of the graph. For ease of discussion, performing attribute masking on node embeddings is equivalent to setting the weight of corresponding dimensions to zero in the graph embeddings. In addition, as the dimensionality of the representation increases, the representational space of the DR method is expansible. In contrast, SR, being a degenerate form of DR, exhibits a fixed representation space owing to its dependence on the representation space of the underlying graph. Thereby, DR methods can contain more information entropy, which helps the model to acquire more fine-grained and intrinsic rationales of graphs.

5.2 Theoretical Feasibility of the Innate Mechanism of the DR

According to (Wright and Ma 2022), high-dimensional problems can be solved with low dimensions. To understand this, we can obtain inspiration from Figure 5, which is a counter-intuitive high-dimensional phenomenon in the problem of measuring concentration on a sphere (Matousek 2002). Figure 5 depicts an ϵ-strip surrounding the equatorial great circle of the sphere $\mathbb{S}^{n-1}$ in $\mathbb{R}^n$. In this case, the great circle corresponds to the equator, where $x_n = 0$. To ensure that the strip covers a significant portion, let's say 99% of the sphere's area, we have:

$$Area\{\boldsymbol{x} \in \mathbb{S}^{n-1} : -\epsilon \leq x_n \leq \epsilon\} = 0.99 \cdot Area\left(\mathbb{S}^{n-1}\right). \tag{14}$$

Empirical evidence from low-dimensional spheres suggests that a large value of ϵ is necessary. However, a straightforward calculation reveals that as the dimension n increases, ϵ decreases on the order of $n^{-1/2}$. Consequently, as n becomes large, the width of the strip 2ϵ can become arbitrarily small. Consequently, as illustrated in Figure 5, the majority of the sphere's area concentrates around the equator.

By the same token, obtaining discriminative information from high-dimensional graph embeddings can be solved with low dimensions. The process of obtaining the DR can be regarded as detecting the point distribution of the sphere. In extreme cases, only a few dimensions of graph embeddings contribute to the downstream task, i.e., many dimensional weights are approaching 0. Then, our graph representation problem can be equivalent to the problem of measuring concentration on a sphere in Figure 5.

5.3 Guarantees for DRGCL's Effectiveness

Motivated by (Wang et al. 2022; Li et al. 2022a), we provide two Theorems as guarantees for DRGCL's effectiveness in the field of self-supervised graph representation learning. Theorem 5.1 states that reducing the risk of GCL loss can improve the performance on downstream tasks, supporting our intuition to make the model focus on the acquisition of discriminative information by learning a DR weight. Theorem 5.2 states that given the label $\boldsymbol{y}$, the DR-aware representation $\tilde{\boldsymbol{z}}$ has smaller conditional variance than $\boldsymbol{z}$ in conventional GCL. Two Theorems are formulated as follows:

Theorem 5.1. *(Connecting Graph DR-aware Representations to Downstream Cross-Entropy Loss). Under the minimal assumption of GCL, i.e., the graph contrastive label is invariant to the distributions, when $\mathcal{R}$ is optimal, for any $\tilde{\boldsymbol{z}} \in \mathbb{R}$, the cross-entropy loss $\mathcal{L}_{CE}^{\mu}(\tilde{\boldsymbol{z}})$ for downstream classification can be bounded by $\mathcal{L}_{DRIN}(\tilde{\boldsymbol{z}})$:*

$$\begin{aligned}&\mathcal{L}_{DRIN}(\tilde{\boldsymbol{z}}) - \sqrt{\psi\left(\tilde{\boldsymbol{z}}\middle|\boldsymbol{y}\right)} - \frac{1}{2}\psi\left(\tilde{\boldsymbol{z}}\middle|\boldsymbol{y}\right) - Err \\ &\leq \mathcal{L}_{CE}^{\mu}(\tilde{\boldsymbol{z}}) + \log(M/D) \leq \mathcal{L}_{DRIN}(\tilde{\boldsymbol{z}}) + \sqrt{\psi\left(\tilde{\boldsymbol{z}}\middle|\boldsymbol{y}\right)} + Err,\end{aligned} \tag{15}$$

where M is negative samples' quantity, D denotes the representation's dimensionality, $\tilde{\boldsymbol{z}}$ is the DR-aware representation, $\boldsymbol{y}$ is the target label, $\psi\left(\tilde{\boldsymbol{z}}\middle|\boldsymbol{y}\right)$ is the conditional feature variance, and $Err = \mathcal{O}\left(M^{-1/2}\right)$ is the approximation error's order.

Theorem 5.2. *(Guarantees for Reduced Conditional Variance of Graph DR-aware Representations). When $\mathcal{R}$ is opti-*

Fixed $\mathcal{R}$	BBBP	Tox21	ToxCast	SIDER	ClinTox	BACE	AVG.
0.1	69.2 ± 1.0	75.3 ± 0.3	63.3 ± 0.3	60.6 ± 0.7	79.1 ± 1.4	75.3 ± 1.2	70.5
0.3	69.7 ± 0.9	74.1 ± 0.4	63.7 ± 0.4	61.6 ± 0.7	80.8 ± 1.2	76.3 ± 0.9	71.0
0.7	69.7 ± 0.5	75.0 ± 0.4	63.7 ± 0.5	60.0 ± 0.3	79.0 ± 1.9	73.7 ± 1.1	70.2
1.0	70.6 ± 0.8	74.3 ± 0.5	63.8 ± 0.5	60.3 ± 0.5	77.4 ± 1.4	76.8 ± 0.9	70.5
DRGCL	71.2 ± 0.5	74.7 ± 0.5	64.0 ± 0.5	61.1 ± 0.8	78.2 ± 1.5	78.2 ± 1.0	71.2

Table 5: Transfer learning in ZNIC-2M with different fixed $\mathcal{R}$.

mal, for any coupled $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathbb{R}$, given label $\boldsymbol{y}$, the conditional variance of $\tilde{\boldsymbol{z}}$ is reduced:

$$\psi\left(\tilde{\boldsymbol{z}}\middle|\boldsymbol{y}\right) \leq \psi\left(\boldsymbol{z}\middle|\boldsymbol{y}\right), \; yet \; \psi\left(\left(\tilde{\boldsymbol{z}}\right)^k\middle|\boldsymbol{y}\right) \cong \psi\left(\left(\boldsymbol{z}\right)^k\middle|\boldsymbol{y}\right), \quad (16)$$

where $(\cdot)^k$ is a function acquiring k-th dimension vector.

Proof. Suppose our redundancy reduction part can best decorrelate the dimensions in graph embeddings, we have

$$\psi\left(\tilde{\boldsymbol{z}}\middle|y\right) \overset{(1)}{=} \sum_{k=1}^{D} \psi\left(\left(\tilde{\boldsymbol{z}}\right)^k\middle|y\right) \quad (17)$$

$$\overset{(2)}{=} \sum_{k=1}^{D} \psi\left(\omega_k \left(\boldsymbol{z}\right)^k\middle|y\right) \quad (18)$$

$$\overset{(3)}{=} \sum_{k=1}^{D} \omega_k^2 \psi\left(\left(\boldsymbol{z}\right)^k\middle|y\right) \quad (19)$$

$$\overset{(4)}{\leq} \sum_{k=1}^{D} \psi\left(\left(\boldsymbol{z}\right)^k\middle|y\right) \overset{(5)}{=} \psi\left(\boldsymbol{z}\middle|y\right), \quad (20)$$

where (1) holds because each dimension is independent of the others; (2) is derived by Equation 3 and Equation 4; (3) is acquired by the property of variance that $\psi(Ax) = A^2\psi(x)$ if A is a random variable; (4) holds because $\omega_k \leq 1$; (5) holds due to Equation 14 in (Wang et al. 2022). □

(Li et al. 2022a) has already proved the equality part of Equation 16. Thus, we further provide the proof for the inequality part in Equation 16 as above. We incorporate Theorem 5.2 into Theorem 5.1 in order to infer an outcome: our methodology can more effectively limit the downstream classification risk. This means the upper and lower limits of supervised cross-entropy loss established by DRGCL are more constrained compared to those obtained through conventional GCL techniques.

6 Experiments

6.1 Experimental Setup

For unsupervised learning, we benchmark DRGCL on eight established datasets in TU datasets (Morris et al. 2020). The baselines include Graphlet Kernel (GL) (Shervashidze et al. 2009), Weisfeiler-Lehman Sub-tree Kernel (WL) (Shervashidze et al. 2011), Deep Graph Kernels (DGK) (Yanardag and Vishwanathan 2015), Node2Vec (Grover and Leskovec 2016), Sub2Vec (Adhikari et al. 2018), Graph2Vec (Narayanan et al. 2017), InfoGraph (Sun et al. 2020), GraphCL (You et al. 2020), ADGCL (Suresh et al. 2021), JOAO (You et al. 2021), RGCL (Li et al. 2022c) and SimGRACE (Xia et al. 2022). For transfer learning, we perform pre-training on ZNIC-2M (Sterling and Irwin 2015) and finetune on eight multi-task binary classification datasets (Wu et al. 2017). The baselines include six of nine methods the same as the ones in unsupervised learning and two different pre-train strategies in (Hu et al. 2020), i.e., attribute masking and context prediction. Furthermore, we evaluated the transferability of our approach on the PPI-306k (Zitnik and Leskovec 2017) dataset. The evaluation protocols for unsupervised and transfer learning follow (Sun et al. 2020; You et al. 2020). The details of our model architectures and corresponding hyper-parameters are summarized in Table 6.

Experiment	Unsupervised learning	Transfer learning
Backbone GNN type	GIN	GIN
Backbone neuron	[32,32,32]	[300,300,300,300,300]
D. R. Gen. neuron	96	300
Projection neuron	[512,512,512]	[300,300]
Pooing type	Global add pool	Global mean pool
Pre-train lr	0.01	0.001
Finetune lr	-	{0.01,0.001,0.0001}
Temperature τ	0.1	0.1
Traning epochs	20	{60,80,100}
Trade-off parameter λ	0.001	0.001
Trade-off parameter α	10	10

Table 6: Model architectures and hyper-parameters.

6.2 Unsupervised Learning

The results of unsupervised graph-level representations for downstream graph classification tasks are shown in Table 1. Our method consistently ranks among the top 3 and achieves the lowest average rank of 2.8, outperforming the SR-based method RGCL and other methods without rationalizations. The findings demonstrate the capability of our method to learn discriminative representations.

6.3 Transfer Learning

The results of transfer learning on ZNIC-2M are presented in Table 2. By utilizing DR to construct embeddings that preserve semantic information, our DRGCL framework achieves top-3 performance on six out of eight datasets and exhibits the highest average accuracy compared to existing baselines. Our method demonstrates superior transferability

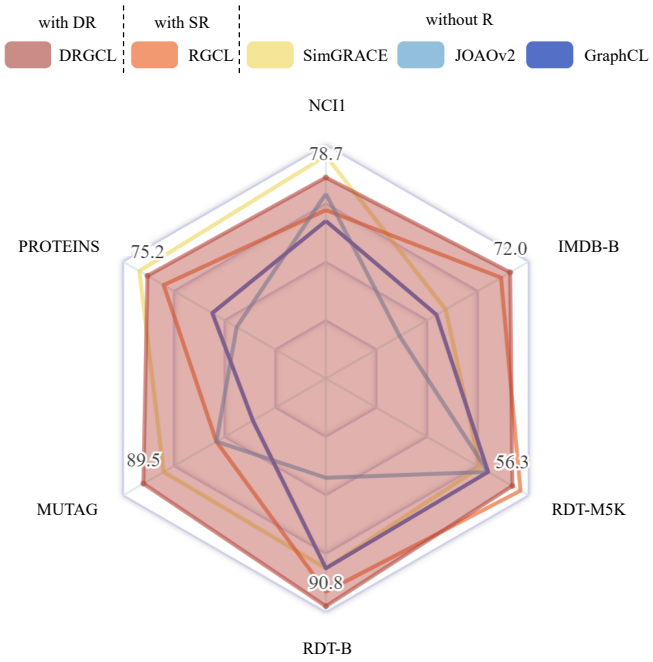

Figure 6: Visualization of unsupervised learning results on six data sets for the top-5 methods. "with DR" denotes our method with DR, "with SR" denotes the SR method RGCL, and "without R" denotes the other methods without using rationale.

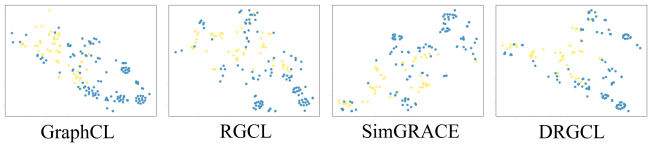

Figure 7: T-SNE visualization of four methods on MUTAG.

compared to other baselines, providing empirical evidence that it can learn more essential rationales in graphs.

The transfer learning results on PPI-306K are shown in Table 3, where our method shows competitive or better transferability than other pre-training schemes.

6.4 Ablation Studies

We conducted ablation studies in transfer learning, shown in Table 4. Our method experiences a decrease of 0.6 in the average score when the DR weight is removed (w/o DR), highlighting the significance of DR. And the decrease of 1.5 in the average score when eliminating the redundancy part (w/o RR) reveals the importance and functionality of RR. It is important to note that the framework GraphCL represents the absence of DR and RR. Notably, both the results of w/o DR and w/o RR outperform GraphCL, emphasizing the positive impact of the two components.

6.5 Transfer Learning with Different Fixed $\mathcal{R}$

In ablation studies, the setting without DR is equivalent to pre-training our model with a fixed dimensional weight, i.e., $\mathcal{R}$, where the meta-learning module is not applied to keep the fixed $\mathcal{R}$ and each dimension of $\mathcal{R}$ is set to 1. In this experiment, we further explore the results of different fixed dimensional weights $\mathcal{R}$ when the DR is not applied. We conducted experiments in transfer learning on ZNIC-2M with fixed $\mathcal{R}$ in $[0.1, 0.3, 0.7, 1.0]$, where each dimension of $\mathcal{R}$ is set the same value. The results of transfer learning on ZNIC-2M with different fixed $\mathcal{R}$ are shown in Table 5. We notice a consistent phenomenon that our DRGCL method with the DR module which updates $\mathcal{R}$ in a meta-learning manner outperforms the other four methods with different fixed dimensional weights. This experiment further proves the significance of the DR.

6.6 Visualization Results

In Figure 6, we visualize the experimental results of the unsupervised learning comparisons. We plot a radar chart with each direction denoting a dataset, the vertexes of the lines denoting the downstream classification results, and the different colors denoting the top-5 methods of unsupervised learning. Note the scale of each direction is different. The visualization results significantly show the performance superiority of the proposed DRGCL over benchmarks. This observation further proves the validity of our findings, i.e., compared with the conventional SR and methods without R, the DR is relatively intrinsic to graphs.

In Figure 7, we utilize T-SNE to visualize four GCL methods on MUTAG in unsupervised learning. In the first three methods, the scatter plots of two different classes exhibit significant overlap, while our proposed method demonstrates reduced overlap, facilitating the identification of a more intuitive hyperplane for their separation. This observation substantiates the superior discriminability of DRGCL.

7 Conclusion

In this paper, we elucidate the causal association among graph embeddings, contrastive labels, and graph DRs, subsequently formulating it through the application of a rigorous SCM. To eliminate the task-agnostic information during pre-training, we propose DRGCL as an intuitive approach to adaptively capture DRs in graph embeddings, which introduces a learnable DR weight updated by a bi-level optimization and a graph DR redundancy reduction regularization term implemented. Benefiting from acquiring DR and reducing the redundancy in graph embeddings, our method achieves new state-of-the-art performance compared to various GCL methods on multiple benchmarks.

Limitations and broader impacts. Due to the needing for a bi-level optimization, it will cost more time to train a model with the ability to capture DR-aware representations. Besides, our method can be seen as a plug-and-play layer that can be used with any GCL method on any feature-based dataset. Thus, it's worth exploring the combinations of the rationales during different procedures of GCL, which may be a good research direction next.

Acknowledgements

The authors would like to thank the editors and reviewers for their valuable comments. This work is supported by the National Funding Program for Postdoctoral Researchers, Grant No. GZC20232812, the CAS Project for Young Scientists in Basic Research, Grant No. YSBR-040, the Youth Innovation Promotion Association CAS, No. 2021106, 2022 Special Research Assistant Grant project, No. E3YD5901, and the China Postdoctoral Science Foundation, No. 2023M743639.

References

Adhikari, B.; Zhang, Y.; Ramakrishnan, N.; and Prakash, B. A. 2018. Sub2Vec: Feature Learning for Subgraphs. In Phung, D. Q.; Tseng, V. S.; Webb, G. I.; Ho, B.; Ganji, M.; and Rashidi, L., eds., *Advances in Knowledge Discovery and Data Mining - 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part II*, volume 10938 of *Lecture Notes in Computer Science*, 170–182. Springer.

Gao, H.; Li, J.; Qiang, W.; Si, L.; Xu, B.; Zheng, C.; and Sun, F. 2022. Robust Causal Graph Representation Learning against Confounding Effects. *CoRR*, abs/2208.08584.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Grover, A.; and Leskovec, J. 2016. node2vec: Scalable Feature Learning for Networks. In Krishnapuram, B.; Shah, M.; Smola, A. J.; Aggarwal, C. C.; Shen, D.; and Rastogi, R., eds., *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 855–864. ACM.

Hotelling, H. 1992. Relations between two sets of variates. *Breakthroughs in statistics: methodology and distribution*, 162–190.

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V. S.; and Leskovec, J. 2020. Strategies for Pre-training Graph Neural Networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lee, J.; Lee, I.; and Kang, J. 2019. Self-Attention Graph Pooling. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 3734–3743. PMLR.

Li, J.; Qiang, W.; Zhang, Y.; Mo, W.; Zheng, C.; Su, B.; and Xiong, H. 2022a. MetaMask: Revisiting Dimensional Confounder for Self-Supervised Learning. In *NeurIPS*.

Li, J.; Zhang, Y.; Qiang, W.; Si, L.; Jiao, C.; Hu, X.; Zheng, C.; and Sun, F. 2022b. Disentangle and Remerge: Interventional Knowledge Distillation for Few-Shot Object Detection from A Conditional Causal Perspective. *CoRR*, abs/2208.12681.

Li, S.; Wang, X.; Zhang, A.; Wu, Y.; He, X.; and Chua, T. 2022c. Let Invariant Rationale Discovery Inspire Graph Contrastive Learning. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 13052–13065. PMLR.

Liu, S.; Davison, A. J.; and Johns, E. 2019. Self-Supervised Generalisation with Meta Auxiliary Learning. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 1677–1687.

Matousek, J. 2002. *Lectures on discrete geometry*, volume 212 of *Graduate texts in mathematics*. Springer. ISBN 978-0-387-95373-1.

Morris, C.; Kriege, N. M.; Bause, F.; Kersting, K.; Mutzel, P.; and Neumann, M. 2020. TUDataset: A collection of benchmark datasets for learning with graphs. *CoRR*, abs/2007.08663.

Narayanan, A.; Chandramohan, M.; Venkatesan, R.; Chen, L.; Liu, Y.; and Jaiswal, S. 2017. graph2vec: Learning Distributed Representations of Graphs. *CoRR*, abs/1707.05005.

Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2).

Qiang, W.; Li, J.; Su, B.; Fu, J.; Xiong, H.; and Wen, J. 2023. Meta Attention-Generation Network for Cross-Granularity Few-Shot Learning. *Int. J. Comput. Vis.*, 131(5): 1211–1233.

Qiang, W.; Li, J.; Zheng, C.; Su, B.; and Xiong, H. 2022. Interventional Contrastive Learning with Meta Semantic Regularizer. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 18018–18030. PMLR.

Ranjan, E.; Sanyal, S.; and Talukdar, P. P. 2020. ASAP: Adaptive Structure Aware Pooling for Learning Hierarchical Graph Representations. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 5470–5477. AAAI Press.

Shervashidze, N.; Schweitzer, P.; van Leeuwen, E. J.; Mehlhorn, K.; and Borgwardt, K. M. 2011. Weisfeiler-Lehman Graph Kernels. *J. Mach. Learn. Res.*, 12: 2539–2561.

Shervashidze, N.; Vishwanathan, S. V. N.; Petri, T.; Mehlhorn, K.; and Borgwardt, K. M. 2009. Efficient graphlet kernels for large graph comparison. In Dyk, D. A. V.; and Welling, M., eds., *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, volume 5 of *JMLR Proceedings*, 488–495. JMLR.org.

Sterling, T.; and Irwin, J. J. 2015. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.*, 55(11): 2324–2337.

Sun, F.; Hoffmann, J.; Verma, V.; and Tang, J. 2020. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Suresh, S.; Li, P.; Hao, C.; and Neville, J. 2021. Adversarial Graph Augmentation to Improve Graph Contrastive Learning. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 15920–15933.

Tan, J.; Geng, S.; Fu, Z.; Ge, Y.; Xu, S.; Li, Y.; and Zhang, Y. 2022a. Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning. In Laforest, F.; Troncy, R.; Simperl, E.; Agarwal, D.; Gionis, A.; Herman, I.; and Médini, L., eds., *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, 1018–1027. ACM.

Tan, J.; Geng, S.; Fu, Z.; Ge, Y.; Xu, S.; Li, Y.; and Zhang, Y. 2022b. Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning. In Laforest, F.; Troncy, R.; Simperl, E.; Agarwal, D.; Gionis, A.; Herman, I.; and Médini, L., eds., *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, 1018–1027. ACM.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2017. Graph Attention Networks. *CoRR*, abs/1710.10903.

Wang, Y.; Zhang, Q.; Wang, Y.; Yang, J.; and Lin, Z. 2022. Chaos is a Ladder: A New Theoretical Understanding of Contrastive Learning via Augmentation Overlap. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Wright, J.; and Ma, Y. 2022. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press.

Wu, Y.; Wang, X.; Zhang, A.; He, X.; and Chua, T. 2022. Discovering Invariant Rationales for Graph Neural Networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; and Pande, V. S. 2017. MoleculeNet: A Benchmark for Molecular Machine Learning. *CoRR*, abs/1703.00564.

Xia, J.; Wu, L.; Chen, J.; Hu, B.; and Li, S. Z. 2022. SimGRACE: A Simple Framework for Graph Contrastive Learning without Data Augmentation. In Laforest, F.; Troncy, R.; Simperl, E.; Agarwal, D.; Gionis, A.; Herman, I.; and Médini, L., eds., *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, 1070–1079. ACM.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How Powerful are Graph Neural Networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yanardag, P.; and Vishwanathan, S. V. N. 2015. Deep Graph Kernels. In Cao, L.; Zhang, C.; Joachims, T.; Webb, G. I.; Margineantu, D. D.; and Williams, G., eds., *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, 1365–1374. ACM.

Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; and Leskovec, J. 2019. GNNExplainer: Generating Explanations for Graph Neural Networks. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 9240–9251.

You, Y.; Chen, T.; Shen, Y.; and Wang, Z. 2021. Graph Contrastive Learning Automated. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 12121–12132. PMLR.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph Contrastive Learning with Augmentations. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhang, H.; Wu, Q.; Yan, J.; Wipf, D.; and Yu, P. S. 2021. From Canonical Correlation Analysis to Self-supervised Graph Neural Networks. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 76–89.

Zitnik, M.; and Leskovec, J. 2017. Predicting multicellular function through multi-layer tissue networks. *Bioinform.*, 33(14): i190–i198.