RESEARCH ARTICLE

Invariant Graph Learning Meets Information Bottleneck for Out-of-Distribution Generalization

Wenyu MAO[1], Jiancan WU[2], Haoyang LIU[2], Yongduo SUI[3], Xiang WANG[4]

1 School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China

2 School of Infomation Science and Technology, University of Science and Technology of China, Hefei 230026, China

3 Tencent, Shenzhen 518054, China

4 School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei 230026, China

Invariant Graph Learning Meets Information Bottleneck for Out-of-Distribution Generalization

Wenyu Mao[1], Jiancan Wu (✉)[2], Haoyang Liu[2], Yongduo Sui[3], Xiang Wang[4]

1 School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China

2 School of Infomation Science and Technology, University of Science and Technology of China, Hefei 230026, China

3 Tencent, Shenzhen 518054, China

4 School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei 230026, China

Abstract   Graph out-of-distribution (OOD) generalization remains a major challenge in graph learning since graph neural networks (GNNs) often suffer from severe performance degradation under distribution shifts. Invariant learning, aiming to extract invariant features across varied distributions, has recently emerged as a promising approach for OOD generalization. Despite the great success of invariant learning in OOD problems for Euclidean data (*i.e.,* images), the exploration within graph data remains constrained by the complex nature of graphs. The invariant features at both the attribute and structural levels, combined with the absence of prior knowledge regarding environmental factors, make the invariance and sufficiency conditions of invariant learning hard to satisfy on graph data. Existing studies, such as data augmentation or causal intervention, either suffer from disruptions to invariance during the graph manipulation process or face reliability issues due to a lack of supervised signals for causal parts. In this work, we propose a novel framework, called Invariant Graph Learning based on Information bottleneck theory (InfoIGL), to extract the invariant features of graphs and enhance models' generalization ability to unseen distributions. Specifically, InfoIGL introduces a redundancy filter to compress task-irrelevant information related to environmental factors. Cooperating with our designed multi-level contrastive learning, we maximize the mutual information among graphs of the same class in the downstream classification tasks, preserving invariant features for prediction to a great extent. An appealing feature of InfoIGL is its strong generalization ability without

depending on supervised signal of invariance. Experiments on both synthetic and real-world datasets demonstrate that our method achieves state-of-the-art performance under OOD generalization for graph classification tasks. The source code is available at https://github.com/maowenyu-11/InfoIGL.

Keywords   Graph OOD, Contrastive Learning, Information Bottleneck Theory, Invariant Learning

1   Introduction

Graphs are ubiquitous in the real world, appearing as chemical molecules, recommender systems [1–4], and knowledge graphs, to name a few examples. In recent years, graph neural networks (GNNs) [5, 6] have emerged as a potent representation learning technique for analyzing and making predictions on graphs. Despite significant advancements, most existing GNN approaches rely heavily on the i.i.d. assumption that the distribution of test data is independently and identically distributed to the training data. Such an assumption, however, seldom holds in practice due to environmental asynchrony during data collection, leading to distribution shifts between the training and testing data [7]. In these situations, GNN suffers from severe performance degradation, making graph OOD generalization a significant challenge in graph learning.

Identifying graph features that remain invariant across distribution shifts [8,9] is paramount in overcoming the graph OOD problem. Invariant learning assumes that invariant features sufficiently determine the label while spurious features are affected by task-irrelation environmental factors cross distributions [10]. Two critical conditions of graph
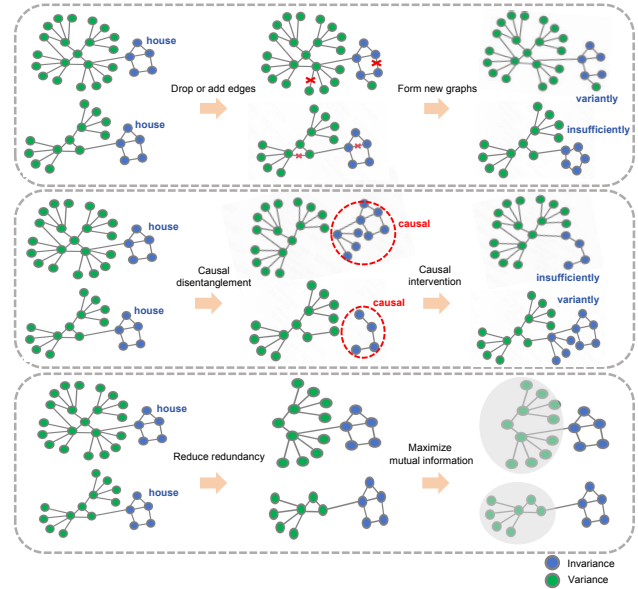
Fig. 1   The comparison of the two branches of existing methods and our method. The upper is data manipulation, which edits nodes or edges and is prone to destroying the invariant parts. The causal disentanglement approach in the middle subfigure separates causal subgraphs and intervenes across various environmental conditions (green patterns), which may fail to identify the causal parts accurately. In contrast, our method in the lower subfigure first removes most of the redundancy to avoid distraction for invariance discovery, facilitating higher identifying accuracy. Then it preserves sufficient predictive information (house) in invariance by maximizing the mutual information, without the use of supervised labels for invariance.

invariance need to be guaranteed in this process [9, 11–13].

- Invariance condition: The graph invariance should exclude spurious features related to environmental factors and maintain robustness across diverse distributions.

- Sufficiency condition: The graph invariance must remain intact and contain sufficient information to predict labels accurately.

Existing studies of invariant graph learning mainly focus on two lines to solve the graph OOD problem: graph manipulation and causal disentanglement approaches. On the one hand, *graph manipulation approaches* [14] typically first generate

diverse augmented data (*e.g.,* adding or removing nodes and edges) to promote the diversity of training distributions, then learn representations consistent on these distributions. This line of research often leads to inappropriate augmentations and destroys the **invariance** accidentally. As shown in the upper subfigure in Figure 1, data augmentation disrupts the invariant parts (the blue patterns) of the original graphs. Consequently, the newly generated graphs are unable to maintain **invariance** or **sufficiently** predict the labels "house". On the other hand, *causal disentanglement methods* [9,12, 15–17] aim to mitigate confounding effects of environments and separate the underlying causal subgraphs according to causal intervention theory [18]. Identifying and separating the causal parts from the non-causal parts is challenging due to two factors. (1) The lack of supervised signals. It is almost impossible to label the causal parts in a graph manually, making the training process for identifying causal subgraphs intractable. (2) The high complexity of the graph. The causal parts may occur in terms of both features (color or properties) and structures (nodes and edges), posing great difficulties in separating the causal parts accurately. As depicted in the middle subfigure of Figure 1, the separated blue pattern is not **invariant** across different green patterns, nor is it **sufficient** for accurately predicting the label "house". Both lines are limited in identifying graph invariance that satisfies the two conditions.

To circumvent the aforementioned limitations, we propose a novel framework **InfoIGL** for invariant graph learning inspired by information bottleneck theory [19], to meet the invariance and sufficiency conditions. The goal is to compress redundant information in graphs to exclude spurious features while maximizing task-relevant information for prediction as invariance, as demonstrated in Figure 1. Specifically, to exclude spurious features, we implement a redundancy filter that assigns minimal invariance scores to redundant information. To preserve sufficient predictive information in invariance, we convert the goal into maximizing the mutual information of graphs from the same class. To fully realize the mutual information maximization, we employ multi-level contrastive learning [20,21] (*i.e.,* semantic- and instance-level), without relying on supervised signals of the invariance. The semantic level focuses on global category semantics that remain stable across complex graphs, while the instance level emphasizes detailed information from local graph samples. This multi-level contrastive learning strategy facilitates a more comprehensive maximization of intra-class mutual information, enhancing the robustness of invariance and satisfying the sufficiency condition.

To prevent contrastive loss from target collapse, where it fails to distinguish between positive and negative samples [22], InfoIGL strengthens instance level contrastive learning with instance constraint and hard negative mining techniques. The outperforming results of extensive experiments on both synthetic and real-world datasets demonstrate the effectiveness of InfoIGL. Our main contributions are summarized as follows.

- To address the graph OOD problem in classification tasks, we propose a novel framework called InfoIGL inspired by information bottleneck theory, which maximizes mutual information of graphs as invariance after compressing the redundant information.
- We incorporate multi-level contrastive learning from both semantic and instance levels to maximize mutual information of graphs in the same class without supervised signals for

invariance.

- We conduct extensive experiments on diverse benchmark datasets to demonstrate the effectiveness of our proposed framework InfoIGL.

2   Related Work

Invariant learning for graph OOD. Invariant learning [8, 23–25] has been extensively explored to improve generalization performance on graph OOD scenarios, which learns robust representations withstanding distribution shifts. Growing research is concentrating on applying invariant learning strategies to tackle the problem of graph OOD generalization, such as optimization methods [10], causal learning [7, 9, 17], stable learning [15, 16], and data manipulation [26, 27]. Optimization methods design optimization objectives to make the model robust to the shifts in data distribution [10, 28–31]. Causal learning utilizes causal intervention to capture causal features for prediction and ignore noncausal features [9, 12, 17, 32–36]. Stable learning leverages sample reweighting to eliminate spurious correlation and extract stable features across different environments [15, 16]. Data manipulation [13, 14, 26, 37, 38] such as dropEdge [27] randomly drops the edges of graphs to increase the diversity of data distribution. However, many of them overlook the intricacies of graph data and are limited by the lack of theoretical guarantees for direct application.

The information bottleneck theory and boundaries of mutual information. Existing research [39, 40] has delineated the paradigm of acquiring a robust representation through the lens of the information bottleneck theory [19]. This theoretical framework endeavors to optimize the mutual information between the derived representation and

predictive outcomes while concurrently minimizing the mutual information between the representation and the original input. This selective process aims at preserving solely the salient information pertinent to the underlying task at hand. Within the context of distribution shifts [41], the pursuit of a robust representation via information bottleneck theory aligns with the extraction of invariant features [42–44]. Nonetheless, the direct computation of mutual information encounters formidable obstacles when dealing with high-dimensional continuous variables of graphs. Given that the precise calculation of mutual information often proves extraneous to the core objectives, models are tailored towards delineating the boundaries of mutual information [45] for optimization purposes. Poole [45] has expounded upon the intricate interplay between mutual information and its assorted lower bounds, wherein the contrastive loss mechanism emerges as a prevalent methodological choice.

Contrastive learning and OOD generalization. Contrastive learning [46–48] has achieved success in aligning representations by pulling together positive pairs and pushing apart negative pairs. Minimizing the contrastive loss can maximize the mutual information between positive pairs while maximizing that between negative pairs. Such a strategy ensures that the mutual information of inputs from the same class encapsulates information relevant to the target, as highlighted in [49, 50], which is stable in the face of distributional shifts as the invariance. This approach thereby establishes crucial invariance necessary for prediction. Recently, the success of leveraging contrastive learning for domain generalization tasks [22, 51, 52] in the area of computer vision attracts researchers' attention in addressing OOD problems in the graph scenarios. For instance, CIGA [53] applies contrastive learn-

ing after decomposing the graph causality which contains the most information on labels. Unlike previous work, we utilize contrastive learning on the features after reducing redundancy to avoid including spurious features in the invariance. We encourage intra-class compactness and inter-class discrimination [21] with class labels to maximize the predictive information. Additionally, to enhance the robustness of graph invariance and satisfy the sufficiency condition, we conduct contrastive learning from both semantic and instance levels to fully maximize the mutual information between graphs.

3 Preliminaries

3.1 Problem Formulation of graph OOD

Let $\mathbb{G}$ and $\mathbb{Y}$ be the sample space and label space, respectively. We denote a sample graph by $G \in \mathbb{G}$ with the adjacent matrix $\mathbf{A}$ and node feature matrix $\mathbf{X}$. The bold $\mathbf{G}$ and $\mathbf{Y}$ are random variables for graphs and labels. Suppose that $\mathcal{D}_{\text{tr}} = \{(G_i^e, Y_i^e)\}_{e \in \mathcal{E}_{\text{tr}}}$ and $\mathcal{D}_{\text{te}} = \{(G_i^e, Y_i^e)\}_{e \in \mathcal{E}_{\text{te}}}$ be the training and testing dataset, where e donates the environment from training environment sets $\mathcal{E}_{\text{tr}}$ and testing environment sets $\mathcal{E}_{\text{te}}$. The training and testing distributions are often inconsistent due to different environmental factors, *i.e.*, $P(\mathbf{G}^e, \mathbf{Y}^e | e = e_1) \neq P(\mathbf{G}^e, \mathbf{Y}^e | e = e_2)$ with $e_1 \in \mathcal{E}_{\text{tr}}$ and $e_2 \in \mathcal{E}_{\text{te}}$. A graph predictor $f = \theta \circ \Phi : \mathbb{G} \rightarrow \mathbb{Y}$ maps the input graph G to the corresponding label $Y \in \mathbb{Y}$, which can be decomposed into a graph encoder Φ and a classifier θ. Consequently, the aim of generalization for graph OOD is to train models $f = \theta \circ \Phi$ with $\mathcal{D}_{tr}$ to generalize well on unseen distributions $\mathcal{D}_{te}$, which can be formulated as follows:

$$\min_{f} \max_{e \in \mathcal{E}_{\text{te}}} \mathbb{E}_{G^e, Y^e \in \mathcal{D}_{te}} [\mathcal{L}(Y^e, f(G^e))] \qquad (1)$$

where $\mathcal{L}(\cdot, \cdot)$ is the loss function between the ground truth and predicted labels.

3.2 Invariant Learning and Information Bottleneck Theory

The objective of Equation 1 is hard to optimize since the prior knowledge for test environments is not available during the training process. Recent studies [10, 11, 53, 54] focus on invariant learning to solve the problem of OOD. The basic idea is that the variables z in the latent space Z can be partitioned into invariant and spurious parts. They define the invariant features $\mathbf{z}_{inv}$ to be those sufficient for the prediction task [55] while the spurious features $\mathbf{z}_{spu}$ be the task-irrelevant [9, 17] ones related to the environment e. We formulate it as below:

$$\mathbf{Y} \perp e \,|\, \mathbf{z}_{inv}, \quad I(\mathbf{Y}; \mathbf{z}_{spu} | \mathbf{z}_{inv}) = 0, \qquad (2)$$

$$I(\mathbf{Y}; \mathbf{z}_{inv}) = I(\mathbf{Y}; \mathbf{G}) \qquad (3)$$

where $I(\cdot; \cdot)$ represents the mutual information between two variables, $\perp$ denotes independence, Equation 2 denotes the invariance condition while Equation 3 denotes the sufficiency condition. Different from existing work based on data manipulation [14, 26, 27, 37] and causal disentanglement [9, 12, 15–17], we take inspirations from the information bottleneck theory [39] for invariant learning, where the goal for information bottleneck is define by:

$$R_{IB}(\theta) = I(\Phi(\mathbf{G}); \mathbf{Y}) - \beta I(\Phi(\mathbf{G}); \mathbf{G}). \qquad (4)$$

For invariant learning, the goal is to optimize the encoder $\Phi(\mathbf{G})$ by minimizing the mutual information $I(\Phi(\mathbf{G}); \mathbf{G})$ to reduce the redundancy from $\mathbf{z}_{spu}$ while maximizing $I(\Phi(\mathbf{G}); \mathbf{Y})$ to satisfy the sufficiency condition [11] for prediction. Therefore, the encoder $\Phi(\mathbf{G})$ is trained to approximate the invariance features, *i.e.*, $\mathbf{z}_{inv} \approx \Phi(\mathbf{G})$.

3.3 Multi-level Contrastive Learning

In practice, the supervised signals for invariance are intractable to obtain. To facilitate invariant learning, contrastive learning can serve as a practical approximation to identify invariance. To satisfy the sufficiency condition, multi-level contrastive learning from both instance and semantic levels can fully capture the invariant features from complex graph data and enhance the robustness of graph invariance. Specifically, instance-level contrastive learning aims to maintain pairwise similarity within instances shared with the same labels in the classification task [48, 56, 57]. In light of this, the instances with the same label serve as positive samples and are pushed closer in the latent space. In contrast, instances with different labels are negative samples and separated apart in the latent space. Consequently, the objective is

$$\mathcal{L}_{\text{ins}} = \mathbb{E}\left[-\log\frac{\exp(\mathbf{z}\cdot\mathbf{z}_+/\tau)}{\exp(\mathbf{z}\cdot\mathbf{z}_+/\tau) + \sum\exp(\mathbf{z}\cdot\mathbf{z}_-/\tau)}\right]$$
(5)

where $\mathbf{z}$, $\mathbf{z}_+$ and $\mathbf{z}_-$ denote the features of input graphs and their corresponding positive and negative instances, respectively.

Semantics are category centers that represent semantic features for each category [22], obtained through methods such as clustering. Semantic-level contrastive learning aims to compactly embed instances around the corresponding semantic while also refining the semantic to better represent the class by minimizing the loss:

$$\mathcal{L}_{\text{sem}} = \mathbb{E}\left[-\log\frac{\exp(\mathbf{z}\cdot\mathbf{w}_+/\tau)}{\exp(\mathbf{z}\cdot\mathbf{w}_+/\tau) + \sum\exp(\mathbf{z}\cdot\mathbf{w}_-/\tau)}\right]$$
(6)

where $\mathbf{z}$, $\mathbf{w}_+$ and $\mathbf{w}_-$ denote the features of instances and their corresponding positive and negative semantics, respectively. Semantics that are the same

class as the instances $\mathbf{z}$ are positive semantic $\mathbf{w}_+$ while negative semantics $\mathbf{w}_-$ are those from other classes.

Instance-level contrastive methods may overlook global category features and exhibit instability due to the complex nature of different graphs. Conversely, semantic-level contrastive learning may sacrifice the exploration of detailed local features in favor of invariance. Incorporating multi-level contrastive learning [56, 57] can extract invariant features **to the greatest extent**, enhancing the robustness of invariance and satisfying the sufficiency condition.

4 Methodology

To satisfy the invariance and sufficiency conditions of invariant learning, we propose to extract invariant graph representations from the perspective of information bottleneck theory. In this section, we first adapt the goals of information bottleneck theory to invariant learning. Then we introduce a novel framework, called InfoIGL according to it, thus solving the graph OOD problem.

4.1 Rethinking Information Bottleneck Theory for Invariant Graph Learning

According to the information bottleneck theory for invariant learning in Section 3.2, we should minimize the mutual information $I(\Phi(\mathbf{G}); \mathbf{G})$ to the lower bound $I(\mathbf{z}_{inv}; \mathbf{G})$ while maximizing $I(\Phi(\mathbf{G}); \mathbf{Y})$ to the upper bound $I(\mathbf{z}_{inv}; \mathbf{Y})$. However, since the supervised signals for invariance are unrealistic to obtain in practice, we thus return to a surrogate objective for training the encoder $\Phi(\cdot)$ based on the theorem proposed in CNC [52] and CIGA [53]: maximizing the mutual information between samples from the same class can approximate maximizing
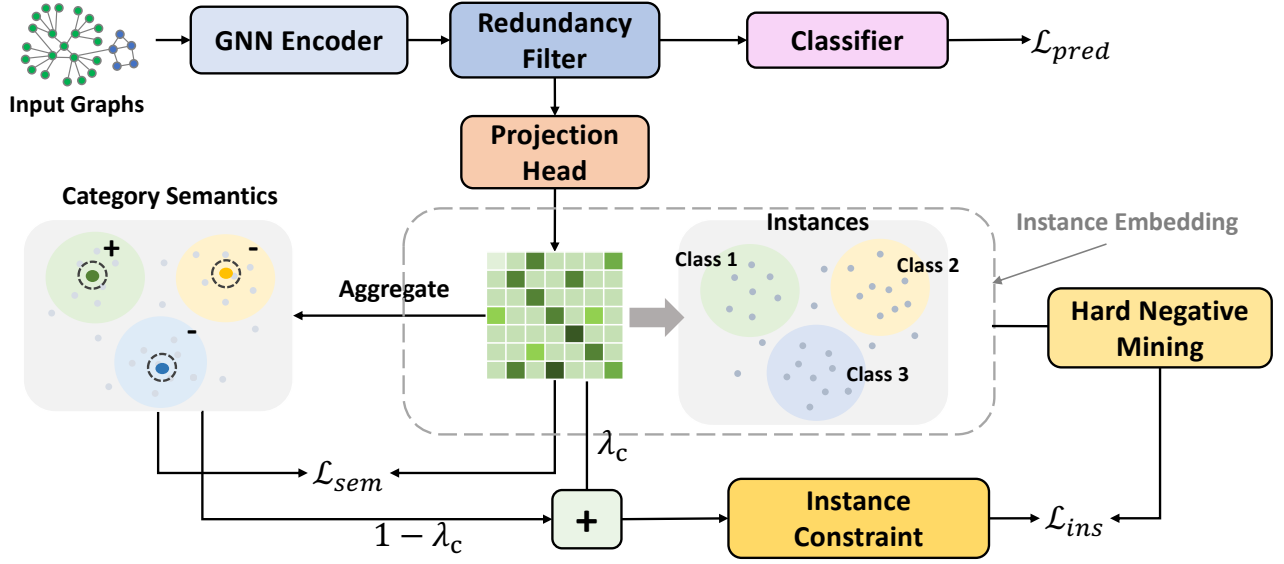
Fig. 2 The overview of proposed InfoIGL framework. The training graphs are fed into the GNN encoder and attention mechanism [17,58]. After being projected to another space, instance embeddings are aggregated to semantics. Then semantic-level and instance-level contrastive learning are optimized jointly, along with instance constraint and hard negative mining to avoid model collapse.

the predictive information for invariance. We formulate it as below:

$$\max I(\Phi(\mathbf{G}); \mathbf{Y}) \rightarrow \max_{\widehat{G}, \widetilde{G}, Y \in D_{tr}} I(\Phi(\widehat{G}); \Phi(\widetilde{G}) \mid Y) \quad (7)$$

To solve the above optimization problem of maximizing $I(\Phi(\widehat{G}); \Phi(\widetilde{G}) \mid Y)$, we resort to the mutual information boundary [45,59] as is shown below. **Mutual information boundary:** Given two variables X, Y, we derive the lower bound of the mutual information $I(X; Y)$ by InfoNCE loss:

$$I(X; Y) \geq \mathbb{E}\left[\frac{1}{K} \sum_{i=1}^{K} \log \frac{\exp(\phi(x_i, y_i))}{\frac{1}{K} \sum_{j=1}^{K} \exp(\phi(x_i, y_j))}\right] \triangleq I_{NCE}$$
(8)

where x_i and y_i denote a positive pair sampled from the joint distribution $P(X, Y)$, while x_i and y_j form a negative pair sampled from the product of marginal

distributions $P(X)P(Y)$, ϕ is the similarity function. This inspires us to leverage contrastive learning to maximize mutual information $I(\Phi(\widehat{G}); \Phi(\widetilde{G}) \mid Y)$.

4.2 Implementation of InfoIGL

Based on Equation 4, InfoIGL first optimizes the encoder with a redundancy filter which minimizes the mutual information $I(\Phi(\mathbf{G}); \mathbf{G})$ by filtering out spurious features $\mathbf{z}_{spu}$. According to Equation 7 and 8, it maximizes the mutual information of graphs from the same class by multi-level contrastive learning as introduced in Section 3.3. Finally, InfoIGL transfers the invariant representation to the downstream task—graph classification. Multi-level contrastive learning can further optimize the redundancy filter, thus satisfying the invariance and sufficiency conditions for invariant learning. The framework is illustrated in Figure 2.

4.2.1 Reducing Redundancy

We implement a redundancy filter to remove task-irrelevant information from graphs and minimize the mutual information $I(\Phi(\mathbf{G}); \mathbf{G})$ in information bottleneck theory. The filter assigns minimal invariance scores for spurious features, which can be realized through attention mechanism [17,58]. Before applying the redundancy filter, we obtain the representations for graph nodes with GNNs first. We can build our framework on any GNN backbones, and here we take GIN [6] as an example, the node update module is defined as follows:

$$\mathbf{h}_v^{(k)} = \text{MLP}^k((1 + \epsilon^{(k)}) \cdot \mathbf{h}_v^{(k-1)} + \sum_{u \in N(v)} \mathbf{h}_u^{(k-1)}) \quad (9)$$

where $\text{MLP}(\cdot)$ stands for multi-layer perceptron, ϵ is a learnable parameter, $\mathbf{h}_v$ and $\mathbf{h}_u$ separately denote the representations of nodes v and u, $N(v)$ denotes the neighbour nodes of v. Then we assign invariance scores to node v and edge (u, v) through the attention mechanism, which can be obtained as follows:

$$\alpha_v = \text{softmax}(\frac{Q_v K_v^\top}{\sqrt{d_k}})V_v \quad (10)$$

$$\alpha_{uv} = \text{softmax}(\text{LeakyReLU}(\text{MLP}(\mathbf{h}_u\|\mathbf{h}_v))) \quad (11)$$

where $Q_v = \mathbf{h}_v \cdot \mathbf{W}^Q$, $K_v = \mathbf{h}_v \cdot \mathbf{W}^K$, $V_v = \mathbf{h}_v \cdot \mathbf{W}^V$, $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ are trainable parameter matrices, d_k is K_v's dimension, $\|$ is the concatenation operation. After assigning invariance scores for nodes and edges, we encode graphs into graph-level representations $\mathbf{h}_G$:

$$\mathbf{h}_v' = \alpha_v \cdot \mathbf{h}_v, \mathbf{h}_{vu}' = \alpha_{uv} \cdot \mathbf{h}_{uv} \quad (12)$$

$$\mathbf{h}_G = \text{READOUT}(\mathbf{H}_v, \mathbf{H}_{uv}) \quad (13)$$

where $\mathbf{h}_v, \mathbf{h}_{uv}$ are the node and edge embedding obtained from GIN. We denote $\mathbf{H}_v = [..., \mathbf{h}_v', ...]_{v \in \mathcal{V}}^\top$, $\mathbf{H}_{uv} = [..., \mathbf{h}_{uv}', ...]_{u,v \in \mathcal{V}}^\top$, $\mathcal{V}$ is the node set. Through

the redundancy filter, InfoIGL can filter out spurious features in graphs and minimize $I(\Phi(\mathbf{G}); \mathbf{G})$ in Equation 4.

4.2.2 Maximizing Predictive Information

To maximize predictive information $I(\Phi(\mathbf{G}); \mathbf{Y})$ and further optimize the redundancy filter, contrastive learning with supervised class labels provides a practical solution to maximize the mutual information between graphs from the same class based on Equation 7 and 8. To enhance the robustness of graph invariance and satisfy the sufficiency condition, we implement multi-level contrastive learning at both the semantic and instance levels according to section 3.3.

Projection Head. We consider applying contrastive learning in another latent space $\mathbf{z}_G$ that is mapped from $\mathbf{h}_G$ by a projection head [22,47]. We employ a two-layer MLP $f_{\theta_p}(\cdot)$ as the projection head:

$$\mathbf{z}_G = f_{\theta_p}(\mathbf{h}_G). \quad (14)$$

Semantic-level Contrastive Learning. Semantic-level contrastive learning forces the graph embeddings z_G to be closer to their corresponding category semantics, promoting both inter-class separation and intra-class compactness. Since the invariance presents a complex nature across different graphs and is difficult to extract, the clustering semantic features for the graphs are more stable and robust. For semantic-level contrastive learning, we first introduce the cluster center of each class as the corresponding semantic. Formally, we initialize the semantic $\mathbf{w}_c$ as the average semantic representation over examples belonging to class c: $\mathbf{w}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} (\mathbf{z}_{G_i})$, N_c is the number of $\mathbf{z}_G$ with label c in a batch. Then we update the current round $\mathbf{w}_c^{(r)}$

by calculating the similarity between each instance embedding $\mathbf{z}_{G_i}$ and the semantic of last round $\mathbf{w}_c^{(r-1)}$:

$$m_i^{(r)} = \text{softmax}(\text{cosine}(\mathbf{z}_{G_i}, \mathbf{w}_c^{(r-1)}))$$
$$\mathbf{w}_c^{(r)} = \sum_{i=1}^{N_c} m_i^{(r)} \cdot \mathbf{z}_{G_i} \qquad (15)$$

where $\text{cosine}(\cdot)$ denotes the cosine similarity. Then we define the semantic-level contrastive loss $\mathcal{L}_{\text{sem}} :=$

$$-\frac{1}{N^{tr}} \sum_{i=1}^{N^{tr}} \log \frac{\exp(\mathbf{z}_{G_i}^\top \mathbf{w}_c / \tau)}{\exp(\mathbf{z}_{G_i}^\top \mathbf{w}_c / \tau) + \sum_{k=1,k\neq c}^{C-1} \exp(\mathbf{z}_{G_i}^\top \mathbf{w}_k / \tau)} \qquad (16)$$

where N^{tr} denotes the number of graphs in a batch, $\mathbf{w}_c$ denotes the target category semantic of $\mathbf{z}_{G_i}$, C denotes the number of classes, τ is the scale factor.

Instance-level Contrastive Learning. We perform instance-level contrastive learning to fully explore the invariance for prediction. Instance-level contrastive learning aligns the similarity between instances with shared labels, which provides more detailed and informative representations. Specifically, we treat instances with shared labels as positive pairs, while viewing those from different classes as negative pairs. The loss function for instance-level contrastive learning can be defined by $\mathcal{L}_{\text{ins}} :=$

$$-\frac{1}{N^{tr}} \sum_{i=1}^{N^{tr}} \log \frac{\exp(\mathbf{z}_{G_i}^\top \mathbf{z}_{G_+} / \tau')}{\exp(\mathbf{z}_{G_i}^\top \mathbf{z}_{G_+} / \tau') + \sum_{k=1}^{K} \exp(\mathbf{z}_{G_i}^\top \mathbf{z}_{G_k} / \tau')} \qquad (17)$$

where the positive sample $\mathbf{z}_{G_+}$ are randomly selected from graphs that belong to the same class as $\mathbf{z}_{G_i}$, K denotes the number of negative samples for each graph instance. τ' is the scale factor.

Discussion for Instance-level Contrastive Learning. Instance-level contrastive learning prone to suffer from model collapse [60] (*i.e.,* samples are mapped to the same point) due to excessive alignment of positive samples. So we apply instance constraint and hard negative mining to prevent getting stuck in trivial solutions [22].

Instance constraint. Enhancing the uniform distribution of graph embeddings $\mathbf{z}_G$ can prevent model collapse from excessive alignment. Here we achieve it by leveraging the uniformity of semantics $\mathbf{w}_c$, which have been ensured by semantic-level contrastive learning. Thus we utilize instance constraint:

$$\mathbf{z}_G' = \lambda_c \cdot \mathbf{z}_G + (1 - \lambda_c) \cdot \mathbf{w}_c \qquad (18)$$

where $\mathbf{w}_c$ refers to the corresponding semantic belonging to the same class as $\mathbf{z}_G$.

Hard negative mining. Hard negative pair [61,62] can help the network learn a better decision boundary in contrastive learning. To identify hard negative samples for instances $\mathbf{z}_G'$ belonging to class c, we calculate the distance between semantic $\mathbf{w}_c$ and samples from other classes within a batch, then we choose the K nearest ones as hard negative samples $\{\mathbf{z}_{\text{hard}_k}'\}_{k=1}^K$. Then the loss for instance-level contrastive learning can be modified as $\mathcal{L}_{\text{ins}} :=$

$$-\frac{1}{N^{tr}} \sum_{i=1}^{N^{tr}} \log \frac{\exp(\mathbf{z}_{G_i}'^\top \mathbf{z}_{G_+}' / \tau')}{\exp(\mathbf{z}_{G_i}'^\top \mathbf{z}_{G_+}' / \tau') + \sum_{k=1}^{K} \exp(\mathbf{z}_{G_i}'^\top \mathbf{z}_{hard_k}' / \tau')} \qquad (19)$$

4.2.3 Downstream Task and Overall Framework.

To align the invariant graph features to the downstream task of graph classification, we define the loss function for prediction as follows:

$$\mathcal{L}_{\text{pred}} = -\frac{1}{N^{tr}} \sum_{i=1}^{N^{tr}} \mathbf{y}_i^\top \log(\theta(\mathbf{h}_{G_i})) \qquad (20)$$

where $\mathbf{y}_i$ is the label of G_i, θ is the classifier for $\mathbf{h}_{G_i}$.

The overview of our proposed framework is illustrated in Figure 2. The final loss of our method can be given by:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_s \mathcal{L}_{\text{sem}} + \lambda_i \mathcal{L}_{\text{ins}} \qquad (21)$$

where the hyperparameters λ_s, λ_i are scaling weights for each loss, which affect the impact of different modules on the model's results. The overall training procedure of the proposed InfoIGL is summarized in Algorithm 1.

Algorithm 1: InfoIGL Pseudocode.

Input: Training graph dataset $\mathcal{D}^{tr}$,
 hyperparameters $\lambda_s, \lambda_i, \lambda_c$, number of epochs e, batch size b.

Output: graph encoder $\Phi(\cdot)$ (GNN encoder and Attention mechanism), classifier $\theta(\cdot)$.

 1: Initialize graph encoder $\Phi(\cdot)$, Projection head $f_{\theta_p}(\cdot)$, classifier $\theta(\cdot)$.
 2: **for** epoch in $1, 2, ... e$ **do**
 3:     Sample data batches $\mathcal{B} = \mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^K$ with batch size b from $\mathcal{D}^{tr}$.
 4:     **for** $D^i \leftarrow \{(G_k, \mathbf{y}_k)\}_{k=1}^b \subset \mathcal{B}, i \in 1...K$ **do**
 5:         # Compressing redundancy
 6:         Calculate $\mathbf{h}_{G_k} \leftarrow \Phi(G_k)$.
 7:         # Multi-level contrastive learning
 8:         Calculate $\mathbf{z}_{G_k} = f_{\theta_p}(\mathbf{h}_{G_k})$.
 9:         Aggregate $\mathbf{z}_{G_k}$ to $\mathbf{w}_c$.
10:         Calculate the semantic-level contrastive loss $\mathcal{L}_{\text{sem}}$.
11:         Calculate $\mathbf{z}'_{G_k} = \lambda_c \mathbf{z}_{G_k} + (1 - \lambda_c)\mathbf{w}_c$.
12:         Obtain hard negative samples $\mathbf{z}'_{hard}$.
13:         Calculate the instance-level contrastive loss $\mathcal{L}_{\text{ins}}$.
14:         # Transfer to downstream tasks
15:         Calculate the prediction loss $\mathcal{L}_{\text{pred}}$.
16:         $\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_s \mathcal{L}_{\text{sem}} + \lambda_i \mathcal{L}_{\text{ins}}$.
17:     **end for**
18: **end for**
19: Update all the trainable parameters to minimize $\mathcal{L}$

4.3 Time and Space Complexity Analysis

Let N be the number of graphs, n be the average node number per graph, l_G, l_A, l_P and d_G, d_A, d_P be the numbers of layers and the embedding dimensions in the GNN backbone, attention mechanism and projection head, respectively, C be the number of class, and K be the number of hard negative samples per instance. The time complexity of the GNN backbone is $O(Nnl_Gd_G)$. For the attention mechanism, the time complexity is $O(Nnl_Ad_A)$. For the projection head, since it turns from node level to graph level, the time complexity is $O(Nl_Pd_P)$. For semantic-level contrastive learning, the time complexity is $O(NC)$. For instance-level contrastive learning, the time complexity is $O(NK)$. Therefore, the time complexity of the whole model is $O(N(nl_Gd_G + nl_Ad_A + nl_Pd_P) + C + K)$, and the order of magnitude is $O(Nn)$. Similarly, the space complexity is also approximately $O(Nn)$ which is about the same as baselines.

---

5 Experiments

In this section, we conduct extensive experiments on multiple benchmarks to answer the following questions:

- Q1: How effective is InfoIGL compared to existing methods for OOD generalization in graph classification tasks?

- Q2: How do reducing redundancy and maximizing mutual information of graphs work respectively?

- Q3: How do the different levels of contrastive learning impact InfoIGL's performance?

- Q4: How sensitive is the model to the hyperparameters and GNN backbones?

- Q5: Can this model be extended to tackle the graph OOD problem on node classification tasks?

We assess the efficacy of InfoIGL across various out-of-distribution (OOD) graph datasets for graph classification tasks in Q1. To address Q2, we perform an ablation study on the redundancy filter and multi-level contrastive learning. For Q3, we further conduct an ablation study examining the impact of semantic-level and instance-level contrastive learning, respectively. Additionally, we explore the sensitivity of InfoIGL to hyperparameters λ_c, λ_s, and λ_i, as well as to different GNN backbones such as GCN, GIN, and GAT. To validate the scalability of InfoIGL, we extend its application to node classification tasks. Visualization techniques are employed to vividly illustrate the significance of InfoIGL in identifying invariant features.

5.1 Expermental Settings

5.1.1 Datasets and Baselines

We conduct experiments on one synthetic (*i.e.,* Motif) and three real-world (*i.e.,* HIV, Molbbbp, and CMNIST) datasets designed for graph OOD [63] on graph classification tasks, where Motif and CMNIST are evaluated with ROC-AUC while HIV and Molbbbp are evaluated with ACC, following the setting of GOOD [63]. Moreover, to answer Q5 and extend our work on node classification tasks, we conduct experiments on two synthetic datasets (*i.e.,* Cora [64] and Amazon-Photo [65]). The details of the datasets are as follows.

- **Motif [63]** is a synthetic dataset motivated by Spurious-Motif [66], graphs of which are generated by connecting a base graph (wheel, tree, ladder, star, and path) and a motif (house, cycle, and crane). The motifs are invariant for prediction while the base graphs may cause distribution shifts. Here, we use the concept shift of "base" and the covariate shift of "size" to create

testing datasets.

- **HIV [63] and Molbbbp [67]** are small-scale molecular datasets adapted from MoleculeNet [68] in the real world, where atoms serve as nodes and chemical bonds serve as edges. Following the setting of GOOD [63], "scaffold" and "size" are defined as the environmental features to create distribution shifts. Here, we select concept shifts of "size" and "scaffold" for testing.

- **CMNIST [63]** is a real-world dataset created by applying superpixel techniques to handwritten digits, where "color" is defined as the environmental features for distribution shifts. Here we use the covariate shift as the OOD testing dataset. In the covariate shift split, digits are colored with seven different colors, with the first five colors, the sixth color, and the seventh color assigned to the training, validation, and test sets, respectively.

- **Cora [64] and Amazon-Photo [65]** are synthetic datasets for node classification tasks with artificial transformation as distribution shifts, following the setting of EERM [11]. The training, valid, and test datasets are created with covariate shifts, which are split with distinct environment IDs.

Statistics of these datasets are presented in Table 1.

We compare our InfoIGL against diverse graph OOD generalization baselines: **Optimization methods:** ERM, IRM [10], VREx [28], GroupDRO [29], FLAG [30]; **Causal learning:** DIR [9], CAL [17], GREA [12], CIGA [53], Disc [32]; **Stable learning:** StableGNN [15], OOD-GNN [16]; and **Data manipulation method:** GSAT [26], DropEdge [27], M-Mixup [14], G-Mixup [37]. Since the practical implementation of InfoIGL involves graph contrastive learning to maximize mutual information,

Table 1   Statistics of multiple graph OOD datasets: Motif, HIV, Molbbbp, and CMNIST for graph classification tasks, and Cora and Amazon-Photo for node classification tasks. "Train", "Val", and "Test" denote the numbers of graphs in the training set, OOD validation set, and OOD test set, respectively. "Classes" and "Metrics" refer to the number of classes and evaluation metrics used for these datasets. "Shift" indicates the type of distribution shifts.

Dataset	#Train	#Val	#Test	#Classes	#Metrics	#Shift
Motif-size	18000	3000	3000	3	ACC	covariate
Motif-base	12600	6000	6000	3	ACC	concept
HIV-size	14454	9956	10525	2	ROC-AUC	concept
HIV-scaffold	15209	9365	10037	2	ROC-AUC	concept
Molbbbp-size	1631	204	204	2	ROC-AUC	concept
Molbbbp-scaffold	1631	204	204	2	ROC-AUC	concept
CMNIST-color	42000	14000	14000	10	ACC	covariate
Cora	-	-	-	10	ACC	covariate
Amazon-Photo	-	-	-	10	ACC	covariate

we also adopt **classical graph contrastive learning** methods as benchmarks, including CNC [52], GMI [69], Infograph [70], GraphCL [71]. Optimization methods aim to design optimization objectives to enhance the robustness of the model across different environments. Causal learning utilizes the causal theory to extract causal features that play a key role in model prediction and ignore non-causal features. Stable learning is committed to independently extracting stable features across different environments through sample reweighting, Data manipulation is dedicated to generating diverse augmented data to increase the diversity of data distribution. While classical graph contrastive learning methods incorporate self-supervised learning and GNN to enhance the quality of graph representations.

5.1.2   Implementation Details

Our code is implemented based on PyTorch Geometric. For all the experiments, we use the Adam optimizer, where the initial and minimum learning rate are searched within $\{0.01, 0.001, 0.0001\}$ and $\{0.001, 0.00001, 0.000001\}$, respectively. We select embedding dimensions from $\{32, 64, 128, 300\}$ and choose batch sizes from $\{64, 128, 256, 512, 1024\}$. The dropout ratio is searched within $\{0.1, 0.3, 0.5\}$ while $\lambda_c, \lambda_s, \lambda_i$ are searched within $\{0.1, 0.2, ..., 0.9\}$. We adopt grid search to tune the hyperparameters and list the details of hyperparameters for InfoIGL in Table 2.

5.2   Overall Results (Q1)

We train and evaluate our proposed InfoIGL, together with all the baselines, 10 times to obtain the average performance (mean ± standard deviation). It can be observed from Table 3 that optimization methods exhibit stable performance with moderate accuracy and low variance, while causal learning baselines show unstable performance with undulating accuracy and high variance. Besides, stable learning and data manipulation baselines perform relatively poorly compared to other baselines. Additionally, traditional graph contrastive learning

Table 2 The hyperparameters for InfoIGL on different datasets, where "sca" denotes "scaffold".

HP	Motif		HIV		Molbbbp		CMNIST
	size	base	size	sca	size	sca	color
layers	3	3	3	3	2	2	5
emb-dim	64	128	300	128	128	300	32
max-epoch	200	200	100	200	100	100	150
batch size	1024	128	256	1024	64	1024	256
ini-lr	1e-3	1e-3	1e-3	1e-2	1e-2	1e-4	1e-3
min-lr	1e-3	1e-6	1e-6	1e-6	1e-6	1e-6	1e-3
decay	0	1e-1	1e-2	1e-5	1e-5	0	0
λ_c	0.7	0.7	0.7	0.7	0.2	0.7	0.7
λ_s	0.8	0.5	0.5	0.5	0.2	0.2	0.5
λ_i	0.2	0.5	0.5	0.1	0.2	0.2	0.1

methods can partially combat distributional shifts, but their effectiveness is not as strong as InfoIGL since they were not designed specially to extract invariant features. These observations indicate that almost all of the baselines have their limitations for graph OOD generalization. Our proposed framework, InfoIGL, achieves state-of-the-art performance on diverse datasets with low variance, outperforming the strongest baseline by 9.82% on HIV (size) and 12.22% on Motif (size). We conduct multiple tests with p-value< 0.05, demonstrating that the performance improvements of InfoIGL are statistically significant. The relatively weak performance of InfoIGL on HIV (scaffold) can be attributed to the Variability and complexity of molecular structures. These results demonstrate the effectiveness of InfoIGL in extracting stable and invariant graph representations on both concept and covariate shifts for graph classification tasks.

5.3 Ablation Study for Q2

To validate the significance of each task individually, we conduct separate ablation studies on the redundancy filter and contrastive learning. Specifically, we compare InfoIGL with two variants: (1) InfoIGL-R: which includes only the redundancy filter with attention mechanism, and (2) InfoIGL-C, which focuses solely on contrastive learning. As is shown in Table 4, InfoIGL-R outperforms ERM on most of the datasets, demonstrating the effectiveness of reducing redundancy. However, its performance falls short of ERM on Motif-size and HIV-scaffold, which means that the variance can not be identified accurately by compressing redundancy merely, underscoring the significance of optimization from contrastive learning. In contrast, InfoIGL-C yields poorer results than ERM on several datasets, such as Molbbbp and motif-size, shedding light on the negative impact of spurious features and the significance of compressing redundancy based on information bottleneck theory.

5.4 Ablation Study for Q3

We perform an ablation study to analyze the impacts of semantic-level and instance-level contrastive learning respectively. Specifically, we compare InfoIGL with three variants: (1) InfoIGL-N, which does not use contrastive learning; (2) InfoIGL-S, which employs semantic-level contrastive learning only; and (3) InfoIGL-I, which applies instance-level contrastive learning only. We report the results in Table 5. Our observations are as follows: 1) Merely applying semantic-level contrastive learning causes performance degradation on the Molbbbp dataset, which confirms the importance of extracting invariance on local features by instance-level contrastive learning. 2) Applying instance-level con-

Table 3   Performance of different methods on synthetic (Motif) and real-world (HIV, Molbbbp, CMNIST) datasets. The best results are in **bold**, and the runner-up results are underlined.

methods	Motif		HIV		Molbbbp		CMNIST
	size	base	size	scaffold	size	scaffold	color
ERM	$70.75_{\pm0.56}$	$81.44_{\pm0.45}$	$63.26_{\pm2.47}$	$72.33_{\pm1.04}$	$78.29_{\pm3.76}$	$68.10_{\pm1.68}$	$28.60_{\pm1.87}$
IRM	$69.77_{\pm0.88}$	$80.71_{\pm0.46}$	$59.90_{\pm3.15}$	$72.59_{\pm0.45}$	$77.56_{\pm2.48}$	$67.22_{\pm1.15}$	$27.83_{\pm2.13}$
GroupDRO	$69.98_{\pm0.86}$	$81.43_{\pm0.70}$	$61.37_{\pm2.79}$	$\mathbf{73.64_{\pm0.86}}$	$79.27_{\pm2.43}$	$66.47_{\pm2.39}$	$29.07_{\pm3.14}$
VREx	$70.24_{\pm0.72}$	$81.56_{\pm0.35}$	$60.23_{\pm1.70}$	$72.60_{\pm0.82}$	$78.76_{\pm2.37}$	$68.74_{\pm1.03}$	$28.48_{\pm2.87}$
FLAG	$56.26_{\pm3.98}$	$72.29_{\pm1.31}$	$66.44_{\pm2.32}$	$70.45_{\pm1.55}$	$79.26_{\pm2.26}$	$67.69_{\pm2.36}$	$32.30_{\pm2.69}$
DIR	$54.96_{\pm9.32}$	$82.96_{\pm4.47}$	$72.61_{\pm2.03}$	$69.05_{\pm0.92}$	$76.40_{\pm4.43}$	$66.86_{\pm2.25}$	$33.20_{\pm6.17}$
CAL	$66.64_{\pm2.74}$	$81.94_{\pm1.20}$	$\underline{83.33_{\pm2.84}}$	$73.05_{\pm1.86}$	$79.20_{\pm3.81}$	$67.37_{\pm3.61}$	$27.99_{\pm3.24}$
GREA	$\underline{73.31_{\pm1.85}}$	$80.60_{\pm2.49}$	$66.48_{\pm4.13}$	$70.96_{\pm3.16}$	$77.34_{\pm3.52}$	$\underline{69.72_{\pm1.66}}$	$29.02_{\pm3.26}$
CIGA	$70.65_{\pm4.81}$	$75.01_{\pm3.56}$	$65.98_{\pm3.31}$	$64.92_{\pm2.09}$	$76.08_{\pm1.21}$	$66.43_{\pm1.99}$	$23.36_{\pm9.32}$
DisC	$53.34_{\pm13.71}$	$76.70_{\pm0.47}$	$56.59_{\pm10.09}$	$67.12_{\pm2.11}$	$75.68_{\pm3.16}$	$60.72_{\pm0.89}$	$24.99_{\pm1.78}$
GSAT	$64.16_{\pm3.35}$	$\underline{83.71_{\pm2.30}}$	$65.63_{\pm0.88}$	$68.88_{\pm1.96}$	$75.63_{\pm3.83}$	$66.78_{\pm1.45}$	$28.17_{\pm1.26}$
DropEdge	$55.27_{\pm5.93}$	$70.84_{\pm6.81}$	$54.92_{\pm1.73}$	$66.78_{\pm2.68}$	$78.32_{\pm3.44}$	$66.49_{\pm1.55}$	$22.65_{\pm2.90}$
M-Mixup	$67.81_{\pm1.13}$	$77.63_{\pm0.57}$	$64.87_{\pm1.77}$	$72.03_{\pm0.53}$	$78.92_{\pm2.43}$	$68.75_{\pm1.03}$	$26.47_{\pm3.45}$
G-Mixup	$59.92_{\pm2.10}$	$74.66_{\pm1.89}$	$70.53_{\pm2.02}$	$71.69_{\pm1.74}$	$78.55_{\pm4.16}$	$67.44_{\pm1.62}$	$31.85_{\pm5.82}$
OOD-GNN	$68.62_{\pm2.98}$	$74.62_{\pm2.66}$	$57.49_{\pm1.08}$	$70.45_{\pm2.02}$	$79.48_{\pm4.19}$	$66.72_{\pm1.23}$	$26.49_{\pm2.94}$
StableGNN	$59.83_{\pm3.40}$	$73.04_{\pm2.78}$	$58.33_{\pm4.69}$	$68.23_{\pm2.44}$	$77.47_{\pm4.69}$	$66.74_{\pm1.30}$	$28.38_{\pm3.49}$
CNC	$66.52_{\pm3.12}$	$82.51_{\pm1.26}$	$70.68_{\pm2.15}$	$66.53_{\pm2.19}$	$76.19_{\pm3.52}$	$68.16_{\pm1.25}$	$32.41_{\pm1.28}$
GMI	$67.90_{\pm1.46}$	$79.52_{\pm0.45}$	$74.34_{\pm0.55}$	$\underline{73.44_{\pm0.35}}$	$77.67_{\pm0.30}$	$69.38_{\pm1.02}$	$30.24_{\pm5.98}$
InfoGraph	$67.49_{\pm2.54}$	$75.57_{\pm0.88}$	$74.63_{\pm0.80}$	$71.41_{\pm0.82}$	$\underline{80.82_{\pm0.49}}$	$70.39_{\pm1.34}$	$\underline{33.84_{\pm1.52}}$
GraphCL	$66.90_{\pm2.80}$	$74.40_{\pm0.90}$	$77.13_{\pm0.17}$	$72.94_{\pm0.68}$	$80.64_{\pm0.78}$	$69.36_{\pm1.32}$	$32.81_{\pm1.71}$
InfoIGL(ours)	$\mathbf{85.53_{\pm2.37}}$	$\mathbf{92.51_{\pm0.16}}$	$\mathbf{93.15_{\pm0.77}}$	$72.37_{\pm1.63}$	$\mathbf{83.39_{\pm2.76}}$	$\mathbf{77.05_{\pm2.24}}$	$\mathbf{38.93_{\pm1.11}}$
improvemrnt	↑ 12.22%	↑ 8.80%	↑ 9.82%	↓ 1.07%	↑ 2.57%	↑ 7.33%	↑ 5.09%

trastive learning solely falls short of ERM on the datasets of Molbbbp and HIV-scaffold, highlighting its tendency to overlook global features of categories for prediction. 3) InfoIGL with both semantic and instance-level contrastive learning outperforms all of the three variants across diverse datasets, proving that jointly optimizing the two contrastive losses can inspire their individual potentials. The Semantic- and instance-level contrastive learning can promote each other and extract invariance for prediction to the greatest extent.

Additionally, we employ the t-SNE technique to visualize embeddings of graph instances on HIV [63] and Motif [63] datasets (*cf.* Figure 3), where the four variants (InfoIGL-N, InfoIGL-S, InfoIGL-I, and InfoIGL) are compared. The results reveal that compared to InfoIGL-N, the embeddings obtained by InfoIGL-S and InfoIGL-I exhibit a more compact

Table 4 Results of ablation experiments on redundancy filter and contrastive learning. InfoIGL-R is the variant of InfoIGL with merely a redundancy filter, while InfoIGL-C is that with contrastive learning only. The best results are in **bold**. The results of methods that are superior to that of ERM are marked with ↑.

methods	Motif		HIV		Molbbbp		CMNIST
	size	base	size	scaffold	size	scaffold	color
ERM	$70.75_{\pm0.56}$	$81.44_{\pm0.45}$	$63.26_{\pm2.47}$	$72.33_{\pm1.04}$	$78.29_{\pm3.76}$	$68.10_{\pm1.68}$	$28.60_{\pm1.87}$
InfoIGL-R	$69.69_{\pm6.24}$ ↓	$87.14_{\pm0.88}$ ↑	$76.99_{\pm2.55}$↑	$71.56_{\pm1.96}$ ↓	$79.72_{\pm3.50}$ ↑	$74.48_{\pm1.00}$↑	$34.54_{\pm2.11}$↑
InfoIGL-C	$68.01_{\pm2.09}$↓	$86.63_{\pm1.33}$↑	$72.81_{\pm2.92}$↑	$68.02_{\pm2.28}$↓	$75.32_{\pm1.38}$↓	$65.62_{\pm1.07}$↓	$33.40_{\pm2.10}$↑
InfoIGL	$\mathbf{85.53_{\pm2.37}}$↑	$\mathbf{92.51_{\pm0.16}}$↑	$\mathbf{93.15_{\pm0.77}}$ ↑	$\mathbf{72.37_{\pm1.63}}$↓	$\mathbf{83.39_{\pm2.76}}$ ↑	$\mathbf{77.05_{\pm2.24}}$↑	$\mathbf{38.93_{\pm1.11}}$↑

Table 5 Results of ablation experiments on semantic-level and instance-level contrastive learning. InfoIGL-S is the variant of InfoIGL with contrastive learning merely from the semantic level, while InfoIGL-I is that with contrastive learning from the instance level only. The best results are in **bold**. The results of methods that are superior to that of ERM are marked with ↑.

methods	Motif		HIV		Molbbbp		CMNIST
	size	base	size	scaffold	size	scaffold	color
InfoIGL-N	$69.69_{\pm6.24}$	$87.14_{\pm0.88}$	$76.99_{\pm2.55}$	$71.56_{\pm1.96}$	$79.72_{\pm3.50}$	$74.48_{\pm1.00}$	$34.54_{\pm2.11}$
InfoIGL-S	$84.77_{\pm2.10}$↑	$89.93_{\pm0.93}$ ↑	$87.30_{\pm1.21}$ ↑	$72.12_{\pm1.87}$ ↑	$81.97_{\pm1.79}$ ↑	$76.76_{\pm3.66}$ ↑	$37.31_{\pm1.50}$ ↑
InfoIGL-I	$80.05_{\pm2.99}$↑	$90.36_{\pm1.54}$↑	$91.38_{\pm2.38}$↑	$63.70_{\pm5.45}$↓	$74.91_{\pm2.07}$↓	$70.11_{\pm2.02}$↓	$35.67_{\pm1.19}$↑
InfoIGL	$\mathbf{85.53_{\pm2.37}}$↑	$\mathbf{92.51_{\pm0.16}}$↑	$\mathbf{93.15_{\pm0.77}}$↑	$\mathbf{72.37_{\pm1.63}}$↑	$\mathbf{83.39_{\pm2.76}}$↑	$\mathbf{77.05_{\pm2.24}}$↑	$\mathbf{38.93_{\pm1.11}}$↑

clustering pattern, reflecting the efficacy of semantic- and instance-level contrastive learning in aligning shared information, respectively. Furthermore, InfoIGL exhibits the best convergence effect, as its embeddings are more tightly clustered than those produced by the other variants. By promoting inter-class separation and intra-class compactness, the embeddings of graphs share a greater amount of information, indicating that InfoIGL can extract improved invariance within a class. This observation highlights the potential of maximizing predictive information of invariance by maximizing the alignment of samples from the same class.

5.5 Sensitive Analysis (Q4)

To assess the sensitivity of InfoIGL to its hyperparameters, namely λ_c for instance constraint and λ_s and λ_i for contrastive loss respectively, we conduct sensitivity analysis experiments by tuning these hyperparameters within the range of $\{0.1, 0.2, ..., 0.9\}$ under the controlled experimental setting. Specifically, when adjusting a specific hyperparameter, we fix the remaining hyperparameters at the values that yield the best performance. The results are presented in Figure 4. It is noteworthy that InfoIGL is relatively insensitive to λ_c (for instance constraint) and λ_s (for semantic-level contrastive learning) in gray and green curves, as adjusting their values does not cause significant fluctuations

(a) HIV
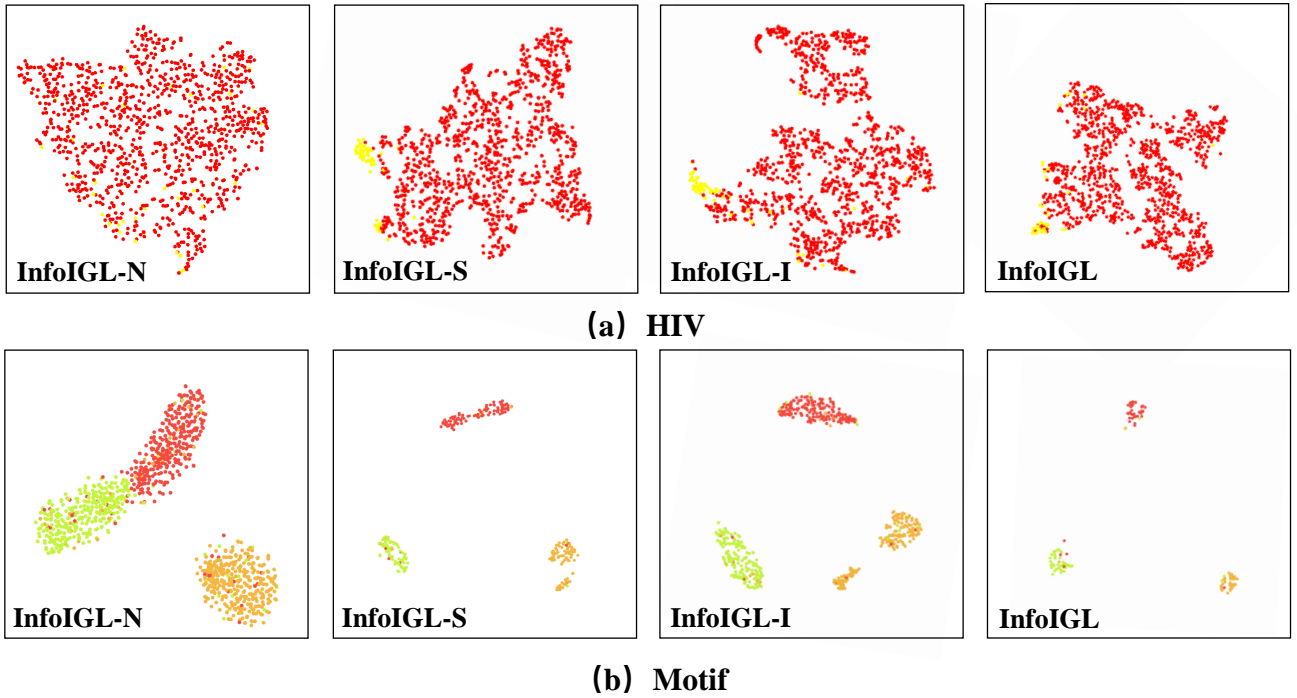

(b) Motif

Fig. 3 The t-SNE visualizations for different levels of contrastive learning.

in InfoIGL's performance. While λ_i (for instance-level contrastive learning) has a significant impact on the InfoIGL and requires fine-tuning. We analyze the reasons behind these phenomena carefully.

- The significant sensitivity of λ_i demonstrates that the instance-level contrastive learning is fragile when applied to complex graph instances. The hyperparameter λ_i needs careful finetuning to accurately identify the invariance of the graphs, due to the intricate nature of graph data.

- The minor influence of λ_s demonstrates that the semantic-level contrastive learning is stable and not sensitive to this hyperparameter. This method effectively learns semantic-level invariant features, benefiting from the robustness of global category semantics.

- The relative smooth curve for λ_c confirms that constraining instance-level representations with robust category semantics can facilitate smoother training. Introducing stable category informa-

tion into complex data instances can enhance prediction accuracy.

Furthermore, the results demonstrate the negative impact of excessively large values for λ_i in InfoIGL. For instance, the performance of InfoIGL on Motif (size) and CMNIST (color) drops significantly when λ_i exceeds 0.6. By comparing the performance of InfoIGL across different datasets under different hyperparameter settings, we can identify the optimal hyperparameters for each dataset. For example, λ_s ranging from 0.3 to 0.7 and λ_c ranging from 0.5 to 0.8 are more suitable hyperparameter values.

We also conduct experiments to evaluate how sensitive is InfoIGL to the choice of graph neural network architectures (GCN, GIN, and GAT). The results are listed in Table 6. As shown in Table 6, InfoIGL-GCN, InfoIGL-GIN, and InfoIGL-GAT are competent on Motif and CMNIST datasets while they far surpass the baseline ERM. The re-
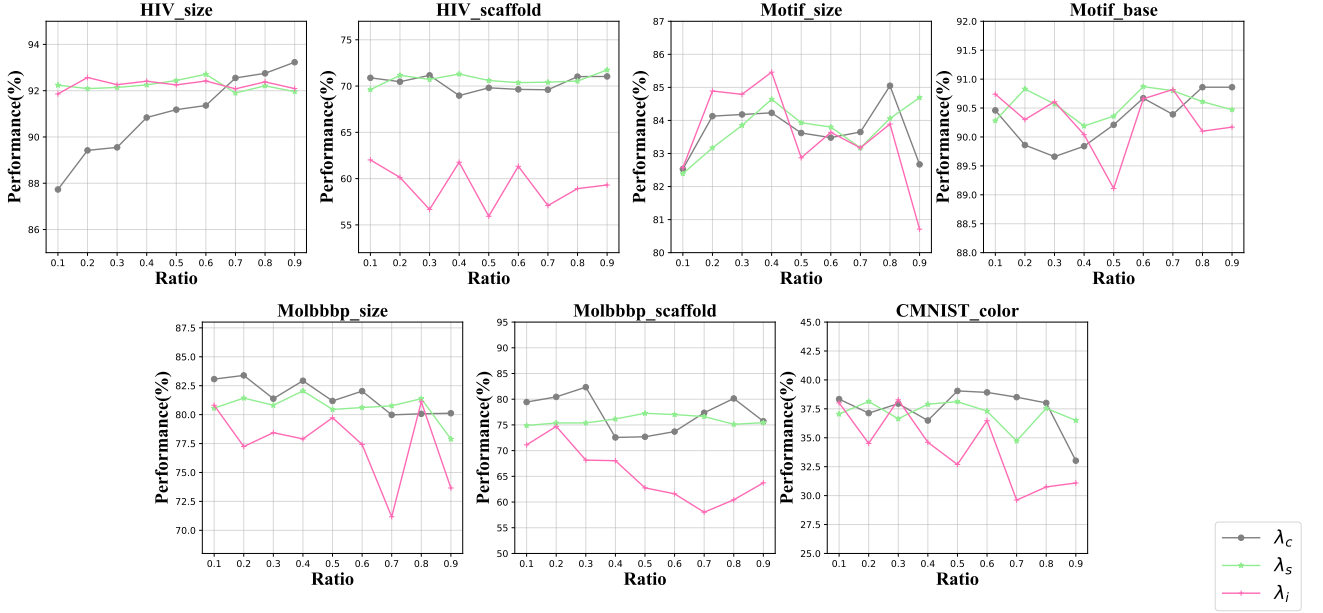
Fig. 4 Sensitivity analysis of hyperparameters $\lambda_c, \lambda_s, \lambda_i$

sults demonstrate the effectiveness of our method, irrespective of the choice of GNN backbones.

Table 6 Results of experiments with different backbones, including GCN, GIN, and GAT.

methods	Motif		CMNIST
	size	base	color
ERM	$70.75_{\pm0.56}$	$81.44_{\pm0.45}$	$28.60_{\pm1.87}$
InfoIGL-GCN	$86.53_{\pm2.15}$	$91.56_{\pm0.91}$	$38.30_{\pm0.76}$
InfoIGL-GIN	$85.53_{\pm2.37}$	$92.51_{\pm0.16}$	$38.93_{\pm1.11}$
InfoIGL-GAT	$84.66_{\pm1.23}$	$90.32_{\pm1.45}$	$37.51_{\pm2.07}$

5.6 Scalability Analysis (Q5)

To extend our method on solving the OOD problems of node classification tasks, we follow the setting of EERM [11] and implement InfoIGL on the dataset of Cora and Amazon-Photo. The performance of InfoIGL on node classification tasks is listed in Table 7. InfoIGL outperforms ERM and EERM [11] on both of Cora and Amazon-photo

datasets, achieving notable performance improvements on different backbones (including GCN, GAT, and SGC).

5.7 Visualization of the Invariance

To validate the effectiveness of our method in extracting invariant features from graphs of the same class, we visually examine a selection of randomly sampled graphs during training to illustrate the instantiated invariance obtained by InfoIGL. We employ a GIN-based encoder and apply InfoIGL specifically on motif-size graphs. The visualizations are presented in Figure 5. The original graphs are depicted on the left side of the grids, where the green regions represent motifs (such as a house, crane, or cycle) that remain invariant across graphs within the same class. Conversely, the yellow portions indicate the base graph (such as a wheel, tree, star, or path) whose size may cause a distribution shift. The ground-truths of invariance are labeled by the dataset creator. On the right side of the grids, we
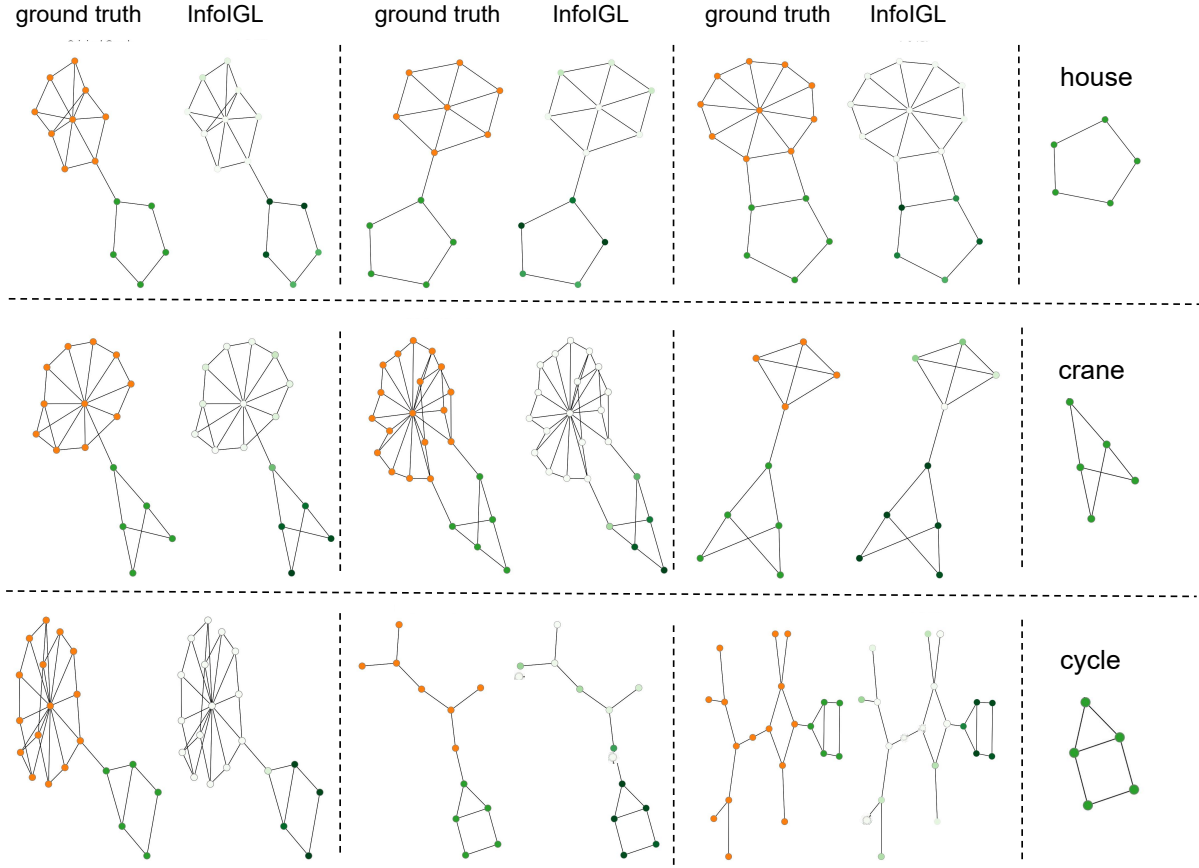
Fig. 5 The invariance obtained by InfoIGL.

display the identified invariance after training using InfoIGL. We colorize nodes and edges based on the attention weights assigned by the model, and the nodes with darker colors indicate a higher degree of invariance. Remarkably, we observe that almost all of the darker colors accurately align with the invariant areas. This further demonstrates the efficacy of InfoIGL in successfully capturing invariant features through information bottleneck theory.

5.8 Running Time Comparison

In our experiment, we conduct a comparative analysis of the running time between ERM, InfoIGL, and its variants (including InfoIGL-S and InfoIGL-I) on various datasets. The results are detailed in Table 8. We can observe that the time spent by

InfoIGL is close to that of ERM, with a slight increase attributed to contrastive learning. This observation highlights the reasonable time complexity and efficiency of InfoIGL. Moreover, thanks to InfoIGL's exceptional capability in tackling graph OOD problems, we can circumvent the need for retraining or finetuning the model under varied data distributions, which can conserve computational resources and enhance both time and space efficiency.

6 Limitations

Informal hard negative mining. There are several techniques for generating hard negative samples in machine learning. One approach is to choose negative samples that closely resemble positive examples by sampling from a pool of negatives. These

Table 7 Results of experiments of InfoIGL on node classification tasks on Cora and Amazon-Photo datasets. The best results are in **bold**.

methods	Cora			Amazon-Photo		
	GCN	GAT	SGC	GCN	GAT	SGC
ERM	$67.28_{\pm5.37}$	$72.39_{\pm6.45}$	$64.79_{\pm4.78}$	$88.06_{\pm0.65}$	$75.85_{\pm1.23}$	$79.26_{\pm2.41}$
EERM	$71.02_{\pm4.52}$	$74.31_{\pm4.87}$	$66.21_{\pm5.68}$	$90.79_{\pm2.18}$	$81.14_{\pm0.76}$	$81.56_{\pm1.53}$
InfoIGL	$\mathbf{72.28_{\pm4.86}}\uparrow$	$\mathbf{77.40_{\pm6.36}}\uparrow$	$\mathbf{69.08_{\pm4.47}}\uparrow$	$\mathbf{91.73_{\pm0.47}}\uparrow$	$\mathbf{82.03_{\pm2.02}}\uparrow$	$\mathbf{83.28_{\pm1.05}}\uparrow$

Table 8 Comparison of running time for ERM, InfoIGL, and its variants.

Dataset	ERM	InfoIGL-S	InfoIGL-I	InfoIGL
Motif-size	00h 21m 32s	00h 20m 41s	00h 20m 33s	00h 21m 21s
Motif-base	00h 33m 10s	00h 32m 29s	00h 32m 48s	00h 33m 49s
HIV-size	00h 08m 43s	00h 10m 44s	00h 12m 07s	00h 12m 39s
HIV-scaffold	00h 14m 27s	00h 15m 08s	00h 14m 23s	00h 16m 13s
Molbbbp-size	00h 03m 44s	00h 03m 31s	00h 03m 54s	00h 04m 29s
Molbbbp-scaffold	00h 03m 45s	00h 06m 18s	00h 06m 37s	00h 06m 37s
CMNIST-color	00h 49m 56s	00h 55m 58s	00h 59m 22s	01h 06m 36s

selected samples can be relabeled as "hard negatives" and included in the training process. Another method involves the use of sophisticated algorithms like online hard example mining (OHEM), which identifies challenging negative samples based on their loss values during training. However, instead of these methods, we select hard negative samples by computing the distance between the negative samples and the semantic center that corresponds to the positive sample. While this informal hard negative mining technique may conserve computational resources, it could also introduce a certain degree of error.

Lack of testing on larger real-world datasets. Due to the scarcity of mature datasets in the field of graph OOD generalization, we initially validated the effectiveness of our approach on the dataset of "GOOD" [63], including Motif, HIV, CMNIST, and Molbbbp. These datasets are created with distribution shifts, which can be divided into two kinds: concept shift and covariate shift. We plan to further validate InfoIGL on a larger and more comprehensive dataset as the next step. Besides, we have the utmost confidence in the effectiveness of our method for performing cross-domain fusion when the dataset quality is ensured. In the future, we will explore expanding the method to tasks like analyzing material structures, predicting properties, or understanding material relationships. The applicability of InfoIGL on larger real-world datasets is what we leave for future work.

7 Conclusion

In this paper, we propose a novel framework, InfoIGL, to extract invariant representation for graph OOD generalization based on information bottle-

neck theory. To satisfy the invariance and suffi-
ciency conditions of invariant learning, we com-
press redundant information from spurious features
with redundancy filter and maximize mutual infor-
mation of graphs from the same class with multi-
level contrastive learning. Extensive experiments
demonstrate the superiority of InfoIGL, highlight-
ing its potential for real-world applications.

8 Acknowledgements

9 Notation

For a better understanding, We summarize and pro-
vide some key notations in Table 9.

Table 9 Notations used in our paper.

Notation	Explanation
$\mathbb{G}$	The graph space
$\mathbb{Y}$	The label space
$\mathbf{G}$	Random variable for the graph sample
$\mathbf{X}$	The node feature matrix for graph $\mathbf{G}$
$\mathbf{A}$	The adjacent matrix for graph $\mathbf{G}$
$\mathbf{Y}$	Random variable of the label for graph $\mathbf{G}$
$\mathcal{D}$	The dataset
$\mathcal{E}$	The environment set
$\mathbf{e}$	The environment factor
$\mathbf{\Phi}$	The graph encoder
θ	The classifier
$\mathbf{h}$	The graph representation after GNN
$\mathbf{z}$	The latent representation after projection head
$\mathbf{w}$	The semantic representation
$\mathbf{h}_G,$	The GNN's embedding of the graph G
$\mathbf{z}_G,$	The latent representations of the graph G after projection head
$G_+, \mathbf{z}_{G_+}$	The positive samples of the graphs and the corresponding representations
$G_-, \mathbf{z}_{G_-}$	The negative samples of the graphs and the corresponding representations
$\mathbf{z}_{inv}$	The invariant features
$\mathbf{z}_{sup}$	The spurious features
$\mathbf{h}_u^{(k)}, \mathbf{h}_{uv}^{(k)}$	The node and edge embedding obtained from the kth layer of GNN
$\mathbf{w}_c$	The semantic representation over examples belonging to class c
$\lambda_c, \lambda_i, \lambda_s$	The non-negative weight for instance constraint, instance- and semantic-level contrastive loss

References

1. Wu J, He X, Wang X, Wang Q, Chen W, Lian J, Xie X. Graph convolution machine for context-aware recommender system. Frontiers Comput. Sci., 2022, 16(6): 166614

2. Mao W, Wu J, Chen W, Gao C, Wang X, He X. Reinforced prompt personalization for recommendation with large language models. CoRR, 2024, abs/2407.17115

3. Mao W, Liu S, Liu H, Liu H, Li X, Hu L. Distinguished quantized guidance for diffusion-based sequence recommendation. In: The Web Conference. 2025

4. Wu J, Wang X, Gao X, Chen J, Fu H, Qiu T. On the effectiveness of sampled softmax loss for item recommendation. ACM Transactions on Information Systems, 2024, 42(4): 98:1–98:26

5. Wu J, Yang Y, Qian Y, Sui Y, Wang X, He X. GIF: A general graph unlearning strategy via influence function. In: The Web Conference. 2023, 651–661

6. Xu K, Hu W, Leskovec J, Jegelka S. How powerful are graph neural networks? In: International Conference on Learning Representations. 2019

7. Sui Y, Mao W, Wang S, Wang X, Wu J, He X, Chua T. Enhancing out-of-distribution generalization on graphs via causal attention learning. ACM Transactions on Knowledge Discovery from Data, 2024, 18(5): 127:1–127:24

8. Wang Z, Veitch V. A unified causal view of domain invariant representation learning. In: ICML 2022: Workshop on Spurious Correlations, Invariance and Stability. 2022

9. Wu Y, Wang X, Zhang A, He X, Chua T S. Discovering invariant rationales for graph neural networks. In: International Conference on Learning Representations. 2022

10. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019

11. Wu Q, Zhang H, Yan J, Wipf D. Handling distribution shifts on graphs: An invariance perspective. In: International Conference on Learning Representations. 2022

12. Liu G, Zhao T, Xu J, Luo T, Jiang M. Graph rationalization with environment-based augmentations. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022, 1069–1078

13. Li H, Zhang Z, Wang X, Zhu W. Learning invariant graph representations for out-of-distribution generalization. In: Advances in Neural Information Processing Systems. 2022, 11828–11841

14. Wang Y, Wang W, Liang Y, Cai Y, Hooi B. Mixup for node and graph classification. In: The Web Conference. 2021, 3663–3674

15. Fan S, Wang X, Shi C, Cui P, Wang B. Generalizing graph neural networks on out-of-distribution graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(1): 322–337

16. Li H, Wang X, Zhang Z, Zhu W. OOD-GNN: out-of-distribution generalized graph neural network. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(7): 7328–7340

17. Sui Y, Wang X, Wu J, Lin M, He X, Chua T. Causal attention for interpretable and generalizable graph classification. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022, 1696–1705

18. Pearl J. Interpretation and identification of causal mediation. Psychological methods, 2014, 19(4): 459

19. Saxe A M, Bansal Y, Dapello J, Advani M, Kolchinsky A, Tracey B D, Cox D D. On the information bottleneck theory of deep learning. In: International Conference on Learning Representations (Poster). 2018

20. Yue X, Zheng Z, Zhang S, Gao Y, Darrell T, Keutzer K, Sangiovanni-Vincentelli A L. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In: Conference on Computer Vision and Pattern Recognition. 2021, 13834–13844

21. Wang R, Wu Z, Weng Z, Chen J, Qi G, Jiang Y. Cross-domain contrastive learning for unsupervised domain adaptation. IEEE Trans. Multim., 2023, 25: 1665–1673

22. Yao X, Bai Y, Zhang X, Zhang Y, Sun Q, Chen R, Li R, Yu B. PCL: proxy-based contrastive learning for domain generalization. In: Conference on Computer Vision and Pattern Recognition. 2022, 7087–7097

23. Zhao H, Combes d R T, Zhang K, Gordon G J. On learning invariant representations for domain adaptation. In: International Conference on Machine Learning. 2019, 7523–7532

24. Rosenfeld E, Ravikumar P K, Risteski A. The risks of invariant risk minimization. In: International Conference on Learning Representations. 2020

25. Yang S, Fu K, Yang X, Lin Y, Zhang J, Cheng P. Learning domain-invariant discriminative features for heterogeneous face recognition. IEEE Access, 2020, 8: 209790–209801

26. Miao S, Liu M, Li P. Interpretable and generalizable graph learning via stochastic attention mechanism. In: International Conference on Machine Learning. 2022, 15524–15543

27. Rong Y, Huang W, Xu T, Huang J. Dropedge: Towards deep graph convolutional networks on node classification. In: International Conference on Learning Representations. 2020

28. Krueger D, Caballero E, Jacobsen J H, Zhang A, Binas J, Zhang D, Priol R L, Courville A. Out-of-distribution generalization via risk extrapolation (rex). In: Proceedings of the 38th International Conference on Machine Learning. 18–24 Jul 2021, 5815–5826

29. Sagawa S, Koh P W, Hashimoto T B, Liang P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In: International Conference on Learning Representations. 2020

30. Kong K, Li G, Ding M, Wu Z, Zhu C, Ghanem B, Taylor G, Goldstein T. Robust optimization as data augmentation for large-scale graphs. In: Conference on Computer Vision and Pattern Recognition. 2022, 60–69

31. Federici M, Tomioka R, Forré P. An information-theoretic approach to distribution shifts. In: Advances in Neural Information Processing Systems. 2021, 17628–17641

32. Fan S, Wang X, Mo Y, Shi C, Tang J. Debiasing graph neural networks via learning disentangled causal substructure. In: Advances in Neural Information Processing Systems. 2022, 24934–24946

33. Sui Y, Wu Q, Wu J, Cui Q, Li L, Zhou J, Wang X, He X. Unleashing the power of graph data augmentation on covariate distribution shift. In: Advances in Neural Information Processing Systems. 2023, 18109–18131

34. Yang N, Zeng K, Wu Q, Jia X, Yan J. Learning substructure invariance for out-of-distribution molecular representations. In: Advances in Neural Information Processing Systems. 2022, 12964–12978

35. Gui S, Liu M, Li X, Luo Y, Ji S. Joint learning of label and environment causal independence for graph out-of-distribution generalization. In: Advances in Neural Information Processing Systems. 2023, 3945–3978

36. Zhuang X, Zhang Q, Ding K, Bian Y, Wang X, Lv J, Chen H, Chen H. Learning invariant molecular representation in latent discrete space. In: Advances in Neural Information Processing Systems. 2023, 78435–78452

37. Han X, Jiang Z, Liu N, Hu X. G-mixup: Graph data augmentation for graph classification. In: International Conference on Machine Learning. 2022, 8230–8248

38. Chen Y, Bian Y, Zhou K, Xie B, Han B, Cheng J. Does invariant graph learning via environment augmentation learn invariance? In: Advances in Neural Information Processing Systems. 2023, 71486–71519

39. Tishby N, Pereira F C, Bialek W. The information bottleneck method. arXiv preprint physics/0004057, 2000

40. Fang J, Zhang G, Wang K, Du W, Duan Y, Wu Y, Zimmermann R, Chu X, Liang Y. On regularization for explaining graph neural networks: An information theory perspective. IEEE Transactions on Knowledge and Data Engineering, 2024, 1–14

41. Ye N, Li K, Bai H, Yu R, Hong L, Zhou F, Li Z, Zhu J. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In: Conference on Computer Vision and Pattern Recognition. 2022, 7937–7948

42. Du Y, Xu J, Xiong H, Qiu Q, Zhen X, Snoek C G M, Shao L. Learning to learn with variational information bottleneck for domain generalization. In: European Conference on Computer Vision. 2020, 200–216

43. Ahuja K, Caballero E, Zhang D, Gagnon-Audet J, Bengio Y, Mitliagkas I, Rish I. Invariance principle meets information bottleneck for out-of-distribution generalization. In: Advances in Neural Information Processing Systems. 2021, 3438–3450

44. Li B, Shen Y, Wang Y, Zhu W, Reed C, Li D, Keutzer K, Zhao H. Invariant information bottleneck for domain generalization. In: AAAI. 2022, 7399–7407

45. Poole B, Ozair S, Oord v. d A, Alemi A A, Tucker G. On variational bounds of mutual information. In: International Conference on Machine Learning. 2019, 5171–5180

46. He K, Fan H, Wu Y, Xie S, Girshick R B. Momentum contrast for unsupervised visual representation learning. In: Conference on Computer Vision and Pattern Recognition. 2020, 9726–9735

47. Chen T, Kornblith S, Norouzi M, Hinton G E. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. 2020, 1597–1607

48. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, Krishnan D. Supervised contrastive learning. In: Advances in Neural Information Processing Systems. 2020, 18661–18673

49. Tian Y, Sun C, Poole B, Krishnan D, Schmid C, Isola P. What makes for good views for contrastive learning? In: Advances in Neural Information Processing Systems. 2020, 6827–6839

50. Hassani K, Ahmadi A H K. Contrastive multi-view representation learning on graphs. In: International Conference on Machine Learning. 2020, 4116–4126

51. Huang W, Yi M, Zhao X, Jiang Z. Towards the generalization of contrastive self-supervised learning. In: International Conference on Learning Representations. 2023

52. Zhang M, Sohoni N S, Zhang H R, Finn C, Re C. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In: International Conference on Machine Learning. 2022, 26484–26516

53. Chen Y, Zhang Y, Bian Y, Yang H, Kaili M, Xie B, Liu T, Han B, Cheng J. Learning causally invariant representations for out-of-distribution generalization on graphs. In: Advances in Neural Information Processing Systems. 2022, 22131–22148

54. Boonlia H, Dam T, Ferdaus M M, Anavatti S G, Mullick A. Improving self-supervised learning for out-of-distribution task via auxiliary classifier. 2022 IEEE International Conference on Image Processing (ICIP), 2022, 3036–3040

55. Yang M, Fang Z, Zhang Y, Du Y, Liu F, Ton J F, Wang J, Wang J. Invariant learning via probability of sufficient and necessary causes. In: Advances in Neural Information Processing Systems. 2023, 79832–79857

56. Xu M, Wang H, Ni B, Guo H, Tang J. Self-supervised graph-level representation learning with local and global structure. In: International Conference on Machine Learning. 2021, 11548–11558

57. Zhang T, Qiu C, Ke W, Süsstrunk S, Salzmann M. Leverage your local and global representations: A new self-supervised learning strategy. In: Conference on Computer Vision and Pattern Recognition. 2022,

16559–16568

58. Brody S, Alon U, Yahav E. How attentive are graph attention networks? In: International Conference on Learning Representations. 2022

59. Belghazi M I, Baratin A, Rajeswar S, Ozair S, Bengio Y, Hjelm R D, Courville A C. Mutual information neural estimation. In: International Conference on Machine Learning. 2018, 530–539

60. Jing L, Vincent P, LeCun Y, Tian Y. Understanding dimensional collapse in contrastive self-supervised learning. In: International Conference on Learning Representations. 2022

61. Xuan H, Stylianou A, Liu X, Pless R. Hard negative examples are hard, but useful. In: European Conference on Computer Vision. 2020, 126–142

62. Robinson J D, Chuang C, Sra S, Jegelka S. Contrastive learning with hard negative samples. In: International Conference on Learning Representations. 2021

63. Gui S, Li X, Wang L, Ji S. Good: A graph out-of-distribution benchmark. In: Advances in Neural Information Processing Systems. 2022, 2059–2073

64. Yang Z, Cohen W, Salakhudinov R. Revisiting semi-supervised learning with graph embeddings. In: International Conference on Machine Learning. 2016, 40–48

65. Rozemberczki B, Allen C, Sarkar R. Multi-scale attributed node embedding. J. Complex Networks, 2021, 9(2)

66. Ying Z, Bourgeois D, You J, Zitnik M, Leskovec J. Gnnexplainer: Generating explanations for graph neural networks. In: Advances in Neural Information Processing Systems. 2019, 9240–9251

67. Hu W, Fey M, Zitnik M, Dong Y, Ren H, Liu B, Catasta M, Leskovec J. Open graph benchmark: Datasets for machine learning on graphs. In: Advances in Neural Information Processing Systems. 2020, 22118–22133

68. Wu Z, Ramsundar B, Feinberg E N, Gomes J, Geniesse C, Pappu A S, Leswing K, Pande V. Moleculenet: a benchmark for molecular machine learning. Chemical science, 2018, 9(2): 513–530

69. Peng Z, Huang W, Luo M, Zheng Q, Rong Y, Xu T, Huang J. Graph representation learning via graphical mutual information maximization. In: The Web Conference. 2020, 259–270

70. Sun F, Hoffmann J, Verma V, Tang J. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In: International Conference on Learning Representations. 2020

71. Hafidi H, Ghogho M, Ciblat P, Swami A. Graphcl: Contrastive self-supervised learning of graph represen-

tations. CoRR, 2020, abs/2007.08025

Wenyu Mao is currently a graduate student at the University of Science and Technology of China (USTC). She earned her Bachelor's degree from Beijing University of Posts and Telecommunications in 2023. Her research interests encompass recommender systems, graph learning, and large language models.

Jiancan Wu is an associate research professor at the University of Science and Technology of China (USTC), where he completed his B.E. and Ph.D in 2017 and 2022, respectively. His research interests focus on information retrieval, data mining, and large language models, particularly in recommender systems and graph learning. He has published several conference papers in top-tier venues such as NeruIPS, ICLR, SIGIR, WWW, and KDD. Moreover, he has served as the program committee member for several top conferences including ICML, NeurIPS, ICLR, SIGIR, WWW, KDD, etc., and the regular reviewer for journals including TKDE, TOIS, etc.

Haoyang Liu earned his Bachelor of Science degree from the School of Mathematics at the University of Science and Technology of China in Hefei, China, in 2023. He is currently a graduate student in the Department of Electronic Engineering and Information Science at the same university.

Yongduo Sui has received his Ph.D. degree in computer science from the University of Science and Technology of China (USTC), Hefei, in 2024. He is now a senior researcher at Tencent, Shenzhen. His research interests include Large Language Models, Agents, Graph Learning, and Recommendation.

Xiang Wang is now a professor at the University of Science and Technology of China (USTC). He received his Ph.D. degree from National University of Singapore in 2019. His research interests include recommender systems, graph learning, AI explainability, and AI security. He has published some academic papers on international conferences such as NeurIPS, ICLR, KDD, WWW, SIGIR, and AAAI. He serves as a program committee member for several top conferences such as SIGIR and WWW.