A Twist for Graph Classification: Optimizing Causal Information Flow in Graph Neural Networks

Zhe Zhao[13], **Pengkun Wang**[12*], **Haibin Wen**[4], **Yudong Zhang**[1], **Zhengyang Zhou**[12], **Yang Wang** [12*]

[1]University of Science and Technology of China, Hefei 230026, China
[2]Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou 215123, China
[3]City University of Hong Kong
[4]Shaoguan University
{zz4543, zyd2020}@mail.ustc.edu.cn, {pengkun, zzy0929, angyan}@ustc.edu.cn, haibin65535@gmail.com

Abstract

Graph neural networks (GNNs) have achieved state-of-the-art results on many graph representation learning tasks by exploiting statistical correlations. However, numerous observations have shown that such correlations may not reflect the true causal mechanisms underlying the data and thus may hamper the ability of the model to generalize beyond the observed distribution. To address this problem, we propose an Information-based Causal Learning (ICL) framework that combines information theory and causality to analyze and improve graph representation learning to transform information relevance to causal dependence. Specifically, we first introduce a multi-objective mutual information optimization objective derived from information-theoretic analysis and causal learning principles to simultaneously extract invariant and interpretable causal information and reduce reliance on non-causal information in correlations. To optimize this multi-objective objective, we enable a causal disentanglement layer that effectively decouples the causal and non-causal information in the graph representations. Moreover, due to the intractability of mutual information estimation, we derive variational bounds that enable us to transform the above objective into a tractable loss function. To balance the multiple information objectives and avoid optimization conflicts, we leverage multi-objective gradient descent to achieve a stable and efficient transformation from informational correlation to causal dependency. Our approach provides important insights into modulating the information flow in GNNs to enhance their reliability and generalization. Extensive experiments demonstrate that our approach significantly improves the robustness and interpretability of GNNs across different distribution shifts. Visual analysis demonstrates how our method converts informative dependencies in representations into causal dependencies.

Introduction

The continuous advances in representational capacity (**?**), architectural flexibility (Gasteiger, Bojchevski, and Günnemann 2018; Wang et al. 2023), and computational efficiency (Chen, Zhu, and Song 2017; Chiang et al. 2019) of graph neural networks (GNNs) have propelled graph classification to unprecedented levels of success across diverse domains including bioinformatics (Chen, Zhu, and Song 2017; Zhao et al. 2021), social network analysis (Wu et al. 2020; Yang et al. 2023), and computer vision (Wang et al. 2019). Recent research (Lv et al. 2022; Kipf et al. 2018) has shown that in graph classification tasks, the salient properties that determine a graph's label often originate from specific causal substructures in the graph. Contemporary graph neural networks (GNNs) learn via end-to-end backpropagation on structure-rich graph inputs and predominantly rely on exploiting statistical correlations between graph features and outputs. As such, GNNs exhibit a tendency to utilize potentially spurious non-causal features for making predictions, as long as they are associated with the target labels. However, non-causal features that are correlated with labels but not causally related tend to vary significantly across domains. Overfitting to one domain may increase spurious correlations, thereby compromising the generalization and reliability of graph neural networks (Jaber et al. 2020; Dai and Wang 2021).

Why does the existence of non-causal features affect the generalization process of graph neural network learning? To understand this problem, we first investigate the decision-making process of GNNs for graph classification from the perspective of mutual information. Referring to the causal hypothesis (Ghorbani and Zou 2019; Schölkopf et al. 2012), non-causal features act as confounding factors (Carlucci et al. 2019; Arjovsky et al. 2019), which open backdoor paths and spuriously correlate causal features and predictions. By modeling the causal features as C and the non-causal features as S, the mutual information between the input graph G and the predicted label Y can be decomposed as:

$$I(Y;G) = I(Y;C) + I(Y;S|C). \tag{1}$$

Here, $I(Y;C)$ represents the mutual information between causal features C and predictions Y, capturing invariant explanatory mechanisms. $I(Y;S|C)$ denotes the mutual information between non-causal features S and predictions Y given causal features C, representing spurious correlations that do not generalize across distributions. Since GNNs are prone to exploit arbitrary statistical associations, they tend to maximize $I(Y;S)$ while ignoring the underlying $I(Y;C)$. However, reliance on $I(Y;S)$ impairs out-of-distribution generalization because the non-causal pattern varies across

*Corresponding author.

domains. Therefore, reducing the non-causal information extracted and predicted by the model during the learning process and enhancing the causal information will enable the model to extract more valuable relevant information, thereby reducing the obstruction of irrelevant information and enhancing the generalization performance in different domains.

To this end, we propose an Information-based Causal Learning (ICL) strategy, which decomposes the process of maximizing the mutual information between the participation graph and prediction into a non-causal information learning process and a causal information learning process according to the mutual information chain rule. This constrains the non-causal information extracted during training while enhancing the extraction of causal information. Specifically, We introduce a *composite objective function* combining the causal feature enhancement term and the non-causal regularization term. The causal term maximizes the mutual information $I(Y;C)$ between the causal information and the prediction target to extract invariant explanatory factors. Meanwhile, the non-causal term minimizes $I(Y;S)$ to reduce dependence on superficial statistical patterns. By optimizing this composite objective, we can steer the learning of GNN dynamically towards the invariant $I(Y;C)$ while avoiding fragile $I(Y;S)$. To further regulate these components, we propose an optimization framework to shape GNN knowledge accretion towards causal mechanisms to improve generalization. Through the joint learning of the two information objectives and achieving Pareto optimality, the model can extract the real relevant information and eliminate irrelevant information to the greatest extent.

Our contributions are summarized as follows:

- *New theoretical insight:* for the first time, we examine causal feature learning in graph classification from an information-theoretic perspective and explain the impact of these two features on generalization performance.
- *New advisable strategy:* we propose an information-based causal learning strategy (ICL) for graph classification, which can simultaneously focus on enhancing the extraction of real relevant information and suppressing the extraction of irrelevant information during the learning process, and guarantees the maximum extraction of real relevant information through the Pareto optimality of the two objectives.
- *Compelling empirical results:* extensive experiments on synthetic and real-world benchmarks demonstrate ICL yields significant improvements in out-of-distribution robustness across diverse distribution shift types. Visual analysis demonstrates how our method converts informative dependencies in representations into causal dependencies.

Related Work

Graph neural networks (GNNs) have emerged as a powerful technique for representation learning on graph-structured data (Trivedi, Yang, and Zha 2020; Jin et al. 2020; Li, Han, and Wu 2018). By propagating node features across graph topology, GNNs can learn expressive embeddings useful for node and graph-level prediction tasks. A variety of GNN architectures including GCNs (Kipf and Welling 2016), GATs (Veličković et al. 2017), and GraphSAGE (Hamilton, Ying, and Leskovec 2017) have advanced state-of-the-art across applications like molecular property prediction (Duvenaud et al. 2015; Zitnik, Agrawal, and Leskovec 2018) and social network analysis (Fan et al. 2019). However, recent work has revealed limitations in out-of-distribution generalization stemming from sensitivity to spurious correlations (Bahng et al. 2020; Ustun, Spangher, and Liu 2019; Yang et al. 2020).

Concurrently, causal learning has gained prominence for discovering explanatory structures from observational data (Lopez-Paz et al. 2017; Parascandolo et al. 2018). By modeling conditional independences and interventions, causal models can encode invariant mechanisms to improve generalization (Zhang et al. 2020; De Haan, Jayaraman, and Levine 2019). Recent works have explored integrating deep learning and causality, using causal principles for representation learning and employing neural networks for causal structure discovery. Our work contributes to this emerging area by proposing causal objectives tailored to improving GNN robustness. In the graph domain, researchers have developed causal discovery methods leveraging topological patterns and proposed causal graph convolutional networks integrating connectivity into representation learning. We build on these works by introducing information theoretic objectives regularizing causal and anti-causal factors during GNN training for robust graph classification. Our analysis reveals promising new directions at the intersection of causal modeling, graph neural networks and domain generalization. Our method provides a new perspective on causal learning in graph neural networks.

Methodology

Our method consists of three main parts. First, we provide a theoretical analysis from an information theoretic perspective for graph representation learning and propose an optimization objective based on mutual information that aims to approximate causal dependencies and eliminate non-causal dependencies. Second, we introduce how to disentangle causal and non-causal information in the graph representation and optimize the proposed objectives through variational approximation. Finally, we discuss how to combine and trade off multiple optimization objectives to achieve optimal prediction performance, robustness, and generalization.

Analysis and Objectives

In this section, we analyze the graph representation learning from an information-theoretic perspective and reveal its deficiencies in extracting causal dependencies. Then, we decompose the information extraction process in neural networks based on causal assumptions and the mutual information chain rule and propose a mutual information based optimization objective based.

In graph classification, attention and pooling-based graph neural networks (GNNs) are commonly used methods that

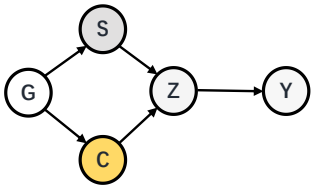

Figure 1: Structural Causal Model (SCM). It consists of graphical data G, causal features C, non-causal features S, graph representation Z, and prediction Y. The arrow $\rightarrow$ indicates causation from one variable to another, i.e., cause $\rightarrow$ effect.

can extract key features from the input graph to support prediction. From a mutual information perspective, this is equivalent to maximizing the mutual information between the representation and prediction target $I(Z;Y)$ (**objective I**), where Z is the learned representation and Y is the prediction target. However, mutual information can only characterize correlations between variables and cannot measure causality. This means it can only absorb all statistical correlations between input features and labels from the training data, without distinguishing between causal and non-causal influences of features. Due to the existence of non-causal features and their shortcut effect in prediction, optimization of predictive correlations may largely stem from non-causal features, thus affecting generalization performance.

To address this problem, we incorporate structural causal models (SCM) and information theory to disentangle the causal and non-causal parts of the mutual information objective I from the causal perspective. Figure 1 shows the structural causal model, which depicts the causal relationships between variables in graph representation learning. It is worth noting that there are the following causal relationships in graph representation learning in SCM:

- $C \leftarrow G \rightarrow S$. C are causal features that directly determine graph properties through causal mechanisms, and S are correlated but non-causal features that serve as 'shortcut' cues. The co-existence of C and S in G leads to this causality.
- $Z \rightarrow Y$. The ultimate goal of graph representation learning is to predict the properties of the input graph. The representation-to-label mappings predicted by the classifier lead to this causality.
- $C \rightarrow Z \leftarrow S$. Z is the representation of given graphical data G. GNN models learn graph representations using both causal and non-causal features simultaneously.

Given the above causal relationships, optimizing objective I is equivalent to maximizing the information flowing of $Z \rightarrow Y$. However, the path $C \rightarrow Z \leftarrow S$ causes $I(Z;Y)$ to contain information from both causal features C and non-causal features S. That is, *maximizing $I(Z;Y)$ will use the correlation from both S and C, making it impossible to distinguish whether the correlation is causal or not.* To address this issue, we decompose the mutual information objective by using the mutual information chain rule and the above causal relationships:

$$I(Z;Y) = I(C,S;Y) = I(C;Y) + I(S;Y|C) \quad (2)$$

where $I(C;Y)$ is the mutual information between causal features C and prediction Y, representing true information and causal dependence. $I(S;Y|C)$ is the mutual information between non-causal features S and prediction Y given C, representing the remaining noise and non-causal dependencies unexplained by C. Due to the entanglement of C and S and the indiscriminate encoding, only optimizing $I(Z;Y)$ will increase both $I(C;Y)$ and $I(S;Y|C)$, resulting in non-causal dependencies and noise that negatively affect model generalization. To learn more causal dependencies, we optimize the causal objective and non-causal objective in Eq. (2) differently, thereby converting mutual information relevance into causal dependence, i.e., we replace the optimization objective I with objective II:

$$\max I(Z;Y) \;\&\; \max I(C;Y) \;\&\; \min I(S;Y|C) \quad (3)$$

This means that while improving prediction accuracy, we can exploit stable causal dependencies to improve robustness and generalization and reduce fitting to noise and non-causal dependencies. Figure 2 illustrates the difference between our proposed objective II and the traditional graph representation learning (objective I). However, the co-optimization of objective II faces many challenges, such as the effective decoupling of the two features and conflicts between information objectives. Therefore, we next make a reasonable combination and trade-off of multiple optimization objectives.

Disentanglement and Optimization

To optimize this objective, we first disentangle causal and non-causal information in the representation. To achieve this, we propose two attention layers that disentangle causal and non-causal information at the edge and node levels, respectively. Specifically, given a GNN encoder $f(\cdot)$ and graph $G = \{A, X\}$, we obtain the encoded representation:

$$H = f(A, X); \quad (4)$$

where H contains both causal and non-causal information from G. To separate them, we propose the causal information extractor Att_C and non-causal information extraction layer Att_S, which learn attention over nodes and edges that are causally and non-causally relevant from the representation, respectively:

$$\begin{cases} \alpha_x, \alpha_a = \sigma(\text{Att}_\text{C}(H, A)) \\ O_x, O_a = \sigma(\text{Att}_\text{S}(H, A)) \end{cases} \quad (5)$$

where α_x, α_a represent the node- and graph-level causal attentions, indicating the importance of nodes and edges in causal dependencies. O_x, O_a represent the corresponding non-causal attentions, indicating importance in non-causal dependencies. The original graph G is disentangled into causal graph Z_c and non-causal graph Z_s based on the two attentions:

$$\begin{cases} Z_c = \text{GConv}_c(A \odot \alpha_a, X \odot \alpha_x) \\ Z_s = \text{GConv}_t(A \odot O_a, X \odot O_x) \end{cases} \quad (6)$$

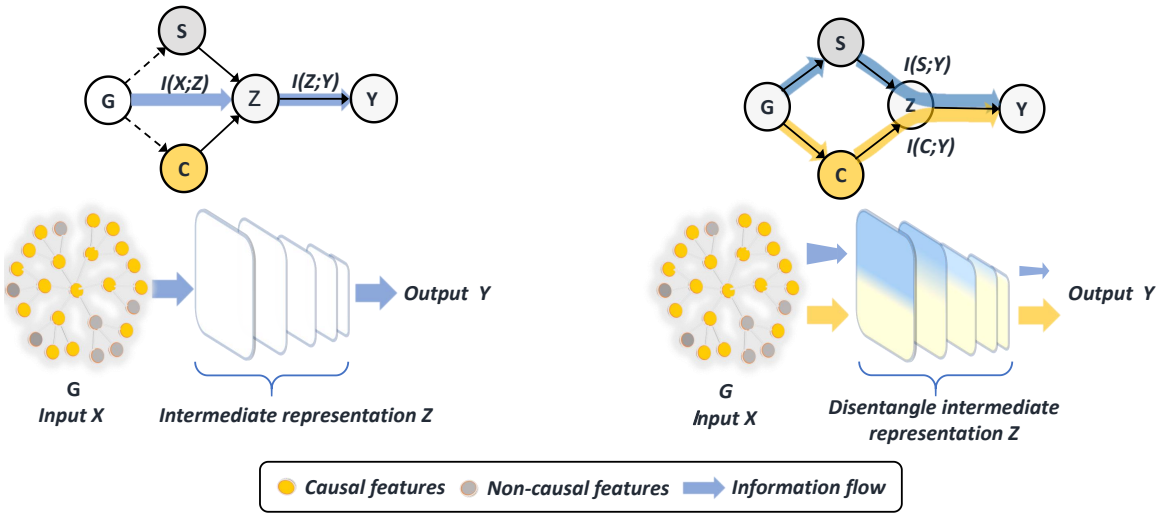

Figure 2: Difference between our objective II and the traditional objective I. The left part of the figure illustrates the learning process of Objective I, that is, causal and non-causal information in the data cannot be distinguished and entangled in prediction-related information to be extracted to support the prediction. The subgraph on the right is our proposed ICL strategy. Our method realizes the transformation from predicted correlation to causal dependence by optimizing objective II, which gradually maximizes the causal information flow and minimizes the non-causal information flow in representation to simultaneously improve prediction performance, robustness, and generalization.

With the disentangled causal representation Z_c and non-causal Z_s, we can optimize objective II. According to the analysis in Section 3.1, to increase learned true causal dependencies, we attempt to maximize $I(C;Y)$. However, due to the high complexity of mutual information optimization, we derive a variational lower bound and minimize it:

$$I(C;Y) \geq E_p(c,y)[\log q(y|c)] - H(Y) \tag{7}$$

where $q(y|c)$ is a conditional probability distribution that can be modeled by a classifier $f_c(\cdot)$, $H(Y)$ is the entropy of Y which is a constant, and $E_p(c,y)[\log q(y|c)]$ is the expectation of classification results overall causal features. The lower bound of $I(C;Y)$ can be optimized by maximizing $E_p(c,y)[\log q(y|c)]$, as shown in **Appendix C**. Thus, we optimize the causal classification loss L_C to enhance the learning of causal dependencies:

$$L_C = -\sum_{c\in C}\sum_{y\in Y} p(c,y)\log q(y|c) \tag{8}$$

where L_C is the cross-entropy loss for the causal representation. Similarly, to maximize $I(Z;Y)$ and improve overall predictive performance, we optimize the cross-entropy loss of the global representation with both causal and non-causal information:

$$L_Y = -\sum_{z\in Z}\sum_{y\in Y} p(z,y)\log q(y|z) \tag{9}$$

Then, to minimize $I(S;Y|C)$ and reduce the influence of non-causal features on prediction, we make S independent of Y by minimizing the KL-divergence between Z_s and the uniform distribution:

$$L_S = \mathrm{KL}(S||u(S)) = -\sum_{s\in S} p(s)\log\frac{p(s)}{u(s)} \tag{10}$$

where $p(s)$ is the distribution of non-causal representation s, and $u(S)$ is the uniform distribution over S. Since the uniform distribution has maximum entropy, minimizing L_S encourages S not to contain information about Y, thus minimizing dependence between non-causal representation S and prediction Y. By optimizing the above loss functions, we can improve overall prediction accuracy and causal dependence, and reduce non-causal dependence. However, despite achieving the conversion from mutual information relevance to causal dependence, there are still interactions and even conflicts between these objectives (e.g. reducing non-causal dependence v.s. improving overall prediction). We next discuss how to combine these objectives to achieve efficient information disentanglement and stable causal learning trade-offs.

Combinations and Tradeoffs

Due to the significant correlation between the optimization objectives and the complexity of causal dependence, we treat the optimization scheme as a multi-objective optimization problem. Mathematically, our problem can be formulated as:

$$\min_{\theta\in\Theta} L(\theta) = \min_{\theta\in\Theta}(L_Y(\theta), L_C(\theta), L_S(\theta)) \tag{11}$$

Multi-objective optimization problems are usually solved with the goal of achieving overall optimality (Pareto optimality). Pareto optimality represents the optimal weights between multiple objectives. In our proposed problem, Pareto optimality can be defined as follows:

Definition 1 (Pareto optimality)**.** *For the multi-objective optimization problem* $\min_{\theta\in\Theta}(L_Y(\theta), L_C(\theta), L_S(\theta))$*, a solution* $\theta^* \in \Theta$ *is* ***Pareto optimal*** *if there does not exist another*

Method	MUTAG	NCI1	PROTEINS	COLLAB	IMDB-B	IMDB-M	AVG
DiffPool	85.61 ± 6.22	75.06 ± 3.66	76.25 ± 4.21	79.24 ± 1.66	74.47 ± 3.84	49.20 ± 3.10	76.38
SortPool	86.17 ± 7.53	79.00 ± 1.68	75.48 ± 1.62	77.84 ± 1.22	73.00 ± 3.50	49.53 ± 2.29	76.49
AGNN	79.77 ± 8.54	79.96 ± 2.37	75.66 ± 3.94	81.10 ± 2.39	73.10 ± 4.07	49.73 ± 3.72	76.15
GCN	88.20 ± 7.33	82.97 ± 2.34	75.65 ± 3.24	81.72 ± 1.64	73.89 ± 5.74	51.53 ± 3.28	78.00
GCN + CAL	88.89 ± 7.16	83.16 ± 1.73	73.32 ± 2.20	82.24 ± 1.62	74.00 ± 5.68	51.7 ± 3.37	79.05
GCN + ICL	88.33 ± 6.89	83.21 ± 2.17	74.93 ± 2.88	**82.68 ± 1.90**	**74.70 ± 5.21**	51.27 ± 3.02	<u>79.22</u>
GIN	<u>89.42 ± 7.40</u>	82.71 ± 1.52	**76.21 ± 3.83**	82.08 ± 1.51	73.40 ± 3.78	<u>51.53 ± 2.97</u>	78.35
GIN + CAL	87.81 ± 10.51	82.73 ± 2.24	73.22 ± 3.46	82.66 ± 1.93	73.60 ± 5.70	51.47 ± 2.77	78.31
GIN + ICL	88.39 ± 8.80	83.36 ± 2.22	75.02 ± 3.51	<u>82.68 ± 1.06</u>	<u>74.50 ± 4.09</u>	**52.00 ± 4.18**	**79.35**
GAT	88.58 ± 7.54	82.11 ± 1.43	<u>75.96 ± 3.26</u>	81.42 ± 1.41	72.70 ± 4.37	50.60 ± 3.75	77.88
GAT + CAL	88.83 ± 6.82	<u>83.36 ± 0.85</u>	74.40 ± 4.14	81.86 ± 1.42	71.90 ± 5.20	50.07 ± 2.84	77.79
GAT + ICL	**91.02 ± 7.02**	**83.38 ± 1.7**	75.12 ± 3.31	81.82 ± 1.20	72.60 ± 2.46	50.67 ± 3.60	78.98

Table 1: Test Accuracy (%) of classification on TUDataset. **Bold** indicates the best performance while <u>underline</u> indicates the second best. We report the average results of ten random trials.

solution $\theta \in \Theta$ ***dominates*** *it, i.e.,*

$$\begin{aligned} \theta \in \Theta \textit{ such that } L_i(\theta) \leq L_i(\theta^*) \textit{ for all } i = Y, C, S \\ \textit{and } L_j(\theta) < L_j(\theta^*) \textit{ for some } j = Y, C, S \end{aligned} \tag{12}$$

where Θ is the feasible solution set. This means that a Pareto optimal solution cannot improve on any one objective without at least worsening another objective. The set of all Pareto optimal solutions is called the **Pareto Frontier**. Although Pareto optimality is considered to have many good properties and is the ultimate goal of many multi-objective problems, it is not suitable for our problem with causal assumptions. Because if objective II reaches Pareto optimality, then when $\max I(Z;Y)$ could be further improved, the optimization of $\max I(Z;Y)$ would have to stop to avoid damaging $\max I(C;Y)$ and $\min I(S;Y|C)$. That is, improving causal dependence and reducing non-causal dependence would hinder improving overall prediction, which clearly contradicts our motivation and causal hypothesis. Therefore, we consider converting the goal to the following:

$$\theta \in \Theta \max I(Z;Y) \;\text{ s.t. } \theta \in P(I(C;Y), -I(S;Y|C)) \tag{13}$$

It means that we maximize the predictive relevance of the representation under the premise of achieving the optimal causal dependence of predictive relevant information. To reach the Pareto optimality of maximizing causal dependence and minimizing non-causal dependence, we use a multi-objective gradient descent algorithm MGDA to optimize $L_C(\theta)$ and $L_S(\theta)$, which is a gradient-based multi-objective optimization algorithm that finds a step balancing the gradients of multiple objective functions at each iteration. The key idea of MGDA is to find a direction d at each iteration, such that taking a small step η along d leads to improvement in all objective functions. In our problem, since there are only two sub-objectives $L_C(\theta)$ and $L_S(\theta)$, we can simplify the MGDA solution. We first compute the gradients of the two sub-objectives $\nabla L_C(\theta)$ and $\nabla L_S(\theta)$ and then compute the angle α between the two gradients:

$$\alpha = \arccos \frac{\nabla L_C(\theta)^T \nabla L_S(\theta)}{\|\nabla L_C(\theta)\| \|\nabla L_S(\theta)\|} \tag{14}$$

By calculating the angle, we determine a direction in which both targets can descend together. We will give more details about MGDA in **Appendix B** and prove that it can find the Pareto optimality of the proposed objective. With the Pareto optimal solution θ^*, we can use it as a constraint to optimize the main objective $I(Z;Y)$. This corresponds to finding a solution on the Pareto front that maximizes $I(Z;Y)$ to obtain the final representation Z^*, for use in prediction tasks. Through this multi-objective optimization process, we achieve a stable trade-off of predictive correlation and causal dependence.

Experiments

Experimental Settings

Datasets. To comprehensively evaluate the efficacy of ICL, we conduct extensive experiments on a diverse set of benchmark datasets. Following Dir, we test on synthetic and real-world graphs exhibiting greater degrees of bias, including Spurious-Motif (Wu et al. 2022) with varying bias levels b, along with MNIST-75sp (Knyazev, Taylor, and Amer 2019), Graph-SST2 (Yuan et al. 2022), and Molhiv (Hu et al. 2020). Furthermore, we assess the capability of ICL to distill causal and non-causal patterns on real-world data across different domains. To this end, we employ six distinct datasets from the TUDataset (Morris et al. 2020) encompassing three biomedical (MUTAG, NCI1, and PROTEINS) and three social graphs (COLLAB, IMDB-B, and IMDB-M). Please refer to **Appendix A** for dataset statistics and details.

Evaluation Metrics. We employ precision for Spurious-Motif to evaluate interpretability and ROC-AUC for Molhiv. For MNIST-75SP, Graph-SST2, and six real-world datasets from TUDataset, we use classification accuracy (Acc) as the metric.

Method	Spurious-Motif			MNIST-75SP	Graph-SST2	Molhiv	AVG
	b = 0.5	b = 0.7	b = 0.9				
Attention	39.42 ± 1.50	37.41 ± 0.86	33.46 ± 0.43	15.19 ± 2.62	81.57 ± 0.71	75.84 ± 1.33	53.87
Top-k Pool	41.21 ± 7.05	40.27 ± 7.12	33.60 ± 0.91	14.91 ± 3.25	79.78 ± 1.35	73.01 ± 1.65	53.94
SAG Pool	43.82 ± 6.32	40.45 ± 7.50	33.60 ± 1.18	14.31 ± 2.44	80.24 ± 1.72	73.26 ± 0.84	54.36
DIR	43.88 ± 4.27	41.87 ± 1.81	39.12 ± 3.51	19.47 ± 1.69	81.89 ± 0.73	68.04 ± 6.24	54.95
GCN	46.20 ± 2.34	38.12 ± 4.56	34.55 ± 1.23	11.35 ± 2.01	82.09 ± 3.45	95.79 ± 1.56	55.55
GCN + CAL	75.31 ± 11.35	69.38 ± 10.79	58.57 ± 5.94	15.76 ± 2.31	84.39 ± 0.28	96.59 ± 0.05	71.26
GCN + ICL	77.45 ± 12.45	75.09 ± 8.40	63.69 ± 8.73	17.04 ± 3.59	**84.67 ± 0.37**	96.89 ± 0.07	72.53
GIN	81.07 ± 3.12	69.30 ± 4.56	59.93 ± 2.34	11.80 ± 1.33	84.37 ± 2.56	96.84 ± 3.21	70.59
GIN + CAL	82.89 ± 8.53	86.86 ± 9.55	86.56 ± 8.91	18.74 ± 2.02	84.59 ± 0.33	97.19 ± 0.07	81.20
GIN + ICL	**82.95 ± 8.53**	**89.00 ± 6.65**	**86.62 ± 5.40**	19.07 ± 1.57	84.46 ± 0.46	**97.19 ± 0.05**	**81.55**
GAT	33.45 ± 2.12	33.60 ± 1.62	33.77 ± 3.45	9.80 ± 1.23	82.10 ± 4.56	97.01 ± 2.34	53.94
GAT + CAL	38.02 ± 6.87	39.42 ± 6.01	35.67 ± 4.35	20.64 ± 5.3	84.30 ± 0.4	97.24 ± 0.06	59.15
GAT + ICL	42.30 ± 10.29	42.01 ± 10.36	40.20 ± 7.78	**21.29 ± 8.40**	84.31 ± 0.4	97.27 ± 0.05	60.82

Table 2: Performance on the Synthetic Dataset and Real Datasets. b is the indicator of the confounding effect in Spurious-Motif dataset. **Bold** indicates the best performance while underline indicates the second best. We report the average results of ten random trials.

Comparison Baselines. As a general framework, ICL can be combined with various GNN architectures, so we conducted experiments on three popular GNNs: GCN (Kipf and Welling 2016), GIN (Xu et al. 2018), and GAT (Veličković et al. 2017), and took them as the most basic baseline. In addition, we compare with other strategies including graph pooling-based methods (DiffPool (Ying et al. 2018), SortPool (Zhang et al. 2018), Top-k Pool (Gao and Ji 2019), and SAGPool (Lee, Lee, and Kang 2019)) and causal heuristic-based methods (DIR (Wu et al. 2022) and CAL (Sui et al. 2022)). See **Appendix A** for more descriptions and implementation details about baselines.

Implementation. We use Pytorch to implement all neural networks and train the model on 8 NVIDIA Tesla V100 GPUs. For datasets from Tudateset, we follow the general experimental setup in CAL, using three-layer GNNs as the encoder, and the hidden layer unit is set to 128. We conduct experiments on GCN, GIN, and GAT. Our network is trained for 100 epochs by the Adma optimizer, and the learning rate is adjusted according to the cosine annealing strategy. For the synthetic data set Spurious-Motif and the real datasets MNIST-75SP, Graph-SST2, and Molhiv, we set the number of hidden layer units to 32, which is the same as Dir, and the rest of the settings remain unchanged from Tudateset. See **Appendix B** for the rest of the details. The source code is available at https://github.com/haibin65535/ICL.

Benchmark Results

Real-world Datasets. Table 1 displays the overall classification accuracies on TUDataset. Overall, the proposed ICL module consistently improves the performance over the base GNN models (GCN, GIN, and GAT) across all datasets. The ICL achieves the best results on 5 out of the 6 datasets: MUTAG, NCI1, IMDB-B, COLLAB, and IMDB-M. This indicates that integrating ICL with GAT is most effective. Comparing ICL to CAL, ICL provides better gains over the base GNNs on more datasets and exhibited higher average performance . The improvements from ICL are quite consistent across different base models like GCN, GIN, and GAT, which highlight its versatility. In summary, the proposed ICL technique consistently boosts performance over standalone GNNs and also outperforms the attention-based CAL in most cases. The combination of GIN + ICL exhibits the highest performance on these datasets, up to 79.35, integrating ICL consistently improves GNN performance, with especially large lifts on MUTAG (2.44%) and IMDB-M (2.07%). The information-theoretic learning of ICL appears highly effective for graph representation learning.

Other Datasets. As shown in Table 2, similar to the real datasets, adding ICL consistently improves the performance of GCN and GIN over their standalone versions across the biased datasets. On the Spurious Motif dataset, ICL outperforms CAL, especially when the bias level is high ($b = 0.7$ and 0.9). This aligns with my previous observation that ICL is more robust to biases than CAL. On MNIST-75SP, ICL substantially beats CAL in accuracy by large margins of 2.5-3%. This further demonstrates the effectiveness of ICL on biased graphs. For MolHIV, ICL slightly surpasses CAL in accuracy. For low bias ($b = 0.5$), CAL is slightly better than ICL on Spurious Motif. But overall, ICL dominates in high-bias settings. In summary, the consistent and substantial gains from ICL over both base GNNs and CAL on these biased datasets highlight its strength at suppressing spurious signals and retaining causal dependencies. The information-theoretic approach appears significantly more robust to biases compared to the attention-based CAL method. Integrating ICL with GNNs is an effective way to improve performance on graphs with strong confounding signals.

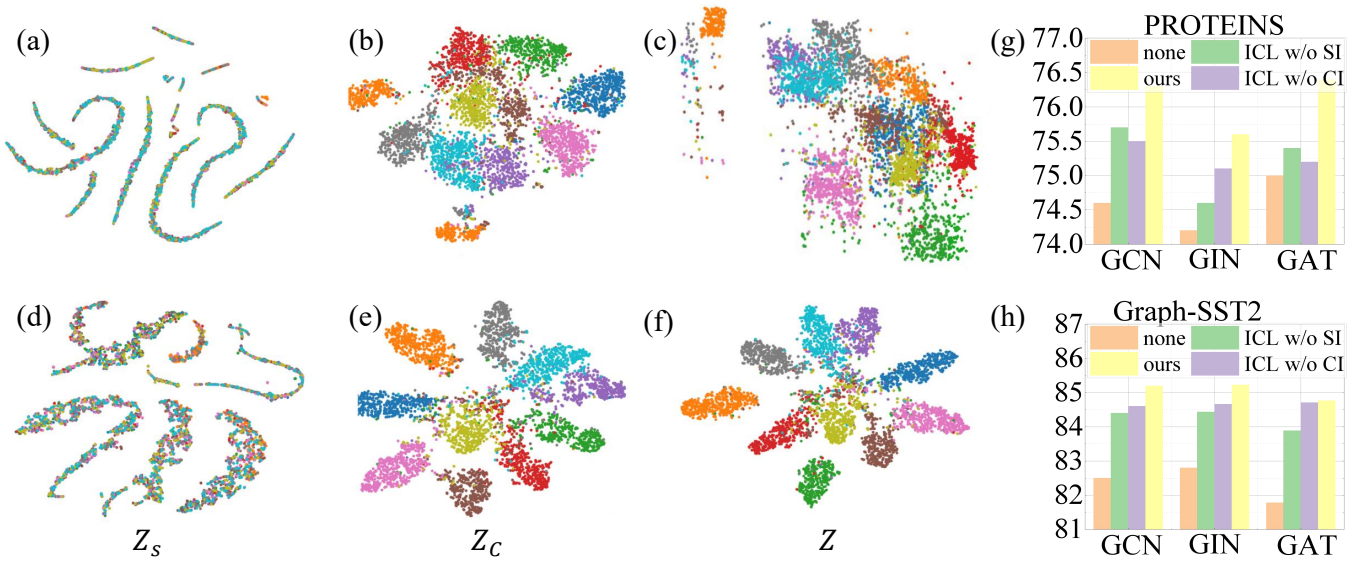

Figure 3: Subfigures (a) to (f) depict the T-SNE analysis of various representations and the progression of information evolution. Dimensionality reduction and clustering techniques are applied to diverse information types within the disentangled representations. Subfigures (g) and (h) illustrate the ablation studies conducted on the proposed method and its variations.

Further Analysis

In this section, we perform T-SNE visualization analysis on the representation to demonstrate the effectiveness of ICL. We observe the effects of different information control objectives on the distribution of representations to examine the process of converting statistical correlations to causal dependencies in the representations.For more analyses of the representations to demonstrate the effectiveness of our method, please refer to the **Appendix E**.

Subfigures (a) (b) (c) show representations Z_s containing only non-causal information, representations Z_c containing only causal information, and representations Z containing both causal and non-causal information, respectively.

- Z_c has clearly separated cluster boundaries between different classes that aligns well with the label distribution, indicating strong intra-class correlation and prediction correlation.
- Z_s is clearly different from Z_c. After removing the deterministic causal part, the distribution of Z_s barely overlaps with the label distribution, indicating almost no prediction correlation.
- Z containing both causal and non-causal information shows a partial intra-class correlation and prediction correlation, but not as significant as Z_c. This shows that the prediction correlation in Z mainly comes from causal information.

Subfigures (d) (e) (f) show the distribution of representations Z_c, Z_s and Z after optimizing with their respective information objectives. We analyze the effects of the proposed information objectives.

- By maximizing the mutual information $I(C;Y)$, the intra-class correlation and prediction correlation of Z_c further increase after the classifier.
- Meanwhile, the clustering effect of S decreases, indicating the success of minimizing $I(S;Y)$ in reducing correlations between non-causal features and predictions.
- Under these objectives, the correlation between the joint representation Z and predictions significantly increases. This validates the effectiveness of objective $I(Z;Y)$.

In addition, the changes from (a) to (d) and (b) to (e) show that the increase of $I(Z;Y)$ is accompanied by the decrease of $I(S;Y)$ and increase of $I(C;Y)$. That is, the improvement of causal dependencies leads to better prediction correlations. Our proposed optimization objectives in Eq. 3 are satisfied and facilitate the conversion from prediction correlations to causal dependencies.

Ablation experiments

Subfigures (g) and (h) in Figure 3 shows the results of our ablation study on the proposed different information control objectives. *None* refers to the baseline without information control that uses only cross-entropy loss and aims to maximize prediction correlation of the representation. *Ours* refers to our proposed ICL method with objectives in Eq. (3). *ICL w/o CI* and *ICL w/o SI* are variants of our method that removes the $max\ I(C;Y)$ and $min\ I(S;Y|C)$ objectives in Eq. (3), respectively. We conducted experiments on MUTAG, PROTEINS and Graph-SST2 datasets. The results on MUTAG please refre to Appendex D. In most cases, the complete ICL achieves the highest performance, significantly outperforming the baseline and the incomplete ICL variants. This demonstrates the effectiveness of our proposed information control objectives. By maximizing the mutual information between causal features and predictions, the model learns more correlation consistent with causal dependencies. Similarly, by minimizing the mutual information between non-causal features and predictions, the model

avoids using correlations inconsistent with causality. When optimizing the complete objectives in Eq. 3, the model's extraction of prediction correlations will strictly conform to causal dependencies. This results in consistently higher overall performance across datasets and model architectures.

Conclusion

This work proposes a novel framework to improve graph neural network generalization by optimizing causal dependencies. Through an information-theoretic analysis, we identify limitations in exploiting spurious correlations and introduce objectives maximizing causal mutual information while minimizing non-causal terms. We achieve this via a causal disentanglement module and multi-objective optimization. Extensive experiments demonstrate significant gains in out-of-distribution robustness across diverse shifts. Our approach provides important insights into regulating information flow in GNNs by transforming statistical associations into robust causal dependencies.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grants No. 62072427, No. 12227901), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No. YSBR-005), and the Academic Leaders Cultivation Program, USTC. Additionally, we acknowledge funding from the Research Grants Council of the Hong Kong Special Administrative Region, China [GRF Project No. Cityu11215723], and the Key Basic Research Foundation of Shenzhen, China (JCYJ20220818100005011).

References

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

Bahng, H.; Chun, S.; Yun, S.; Choo, J.; and Oh, S. J. 2020. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, 528–539. PMLR.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2229–2238.

Chen, J.; Zhu, J.; and Song, L. 2017. Stochastic training of graph convolutional networks with variance reduction. *arXiv preprint arXiv:1710.10568*.

Chiang, W.-L.; Liu, X.; Si, S.; Li, Y.; Bengio, S.; and Hsieh, C.-J. 2019. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 257–266.

Clancey, W. J. 1979. *Transfer of Rule-Based Expertise through a Tutorial Dialogue*. Ph.D. diss., Dept. of Computer Science, Stanford Univ., Stanford, Calif.

Clancey, W. J. 1983. Communication, Simulation, and Intelligent Agents: Implications of Personal Intelligent Machines for Medical Education. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*, 556–560. Menlo Park, Calif: IJCAI Organization.

Clancey, W. J. 1984. Classification Problem Solving. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, 45–54. Menlo Park, Calif.: AAAI Press.

Clancey, W. J. 2021. The Engineering of Qualitative Models. Forthcoming.

Dai, E.; and Wang, S. 2021. Towards self-explainable graph neural network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 302–311.

De Haan, P.; Jayaraman, D.; and Levine, S. 2019. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32.

Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28.

Engelmore, R.; and Morgan, A., eds. 1986. *Blackboard Systems*. Reading, Mass.: Addison-Wesley.

Fan, W.; Ma, Y.; Li, Q.; He, Y.; Zhao, E.; Tang, J.; and Yin, D. 2019. Graph neural networks for social recommendation. In *The world wide web conference*, 417–426.

Gao, H.; and Ji, S. 2019. Graph u-nets. In *international conference on machine learning*, 2083–2092. PMLR.

Gasteiger, J.; Bojchevski, A.; and Günnemann, S. 2018. Predict then propagate: Graph neural networks meet personalized pagerank. *arXiv preprint arXiv:1810.05997*.

Ghorbani, A.; and Zou, J. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, 2242–2251. PMLR.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hasling, D. W.; Clancey, W. J.; and Rennels, G. 1984. Strategic explanations for a diagnostic consultation system. *International Journal of Man-Machine Studies*, 20(1): 3–19.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33: 22118–22133.

Jaber, A.; Kocaoglu, M.; Shanmugam, K.; and Bareinboim, E. 2020. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, 33: 9551–9561.

Jin, W.; Derr, T.; Liu, H.; Wang, Y.; Wang, S.; Liu, Z.; and Tang, J. 2020. Self-supervised learning on graphs: Deep insights and new direction. *arXiv preprint arXiv:2006.10141*.

Kipf, T.; Fetaya, E.; Wang, K.-C.; Welling, M.; and Zemel, R. 2018. Neural relational inference for interacting systems. In *International conference on machine learning*, 2688–2697. PMLR.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Knyazev, B.; Taylor, G. W.; and Amer, M. 2019. Understanding attention and generalization in graph neural networks. *Advances in neural information processing systems*, 32.

Lee, J.; Lee, I.; and Kang, J. 2019. Self-attention graph pooling. In *International conference on machine learning*, 3734–3743. PMLR.

Li, Q.; Han, Z.; and Wu, X.-M. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Lopez-Paz, D.; Nishihara, R.; Chintala, S.; Scholkopf, B.; and Bottou, L. 2017. Discovering causal signals in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6979–6987.

Lv, F.; Liang, J.; Li, S.; Zang, B.; Liu, C. H.; Wang, Z.; and Liu, D. 2022. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8046–8056.

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. *arXiv preprint arXiv:2007.08663*.

NASA. 2015. Pluto: The 'Other' Red Planet. https://www.nasa.gov/nh/pluto-the-other-red-planet. Accessed: 2018-12-06.

Parascandolo, G.; Kilbertus, N.; Rojas-Carulla, M.; and Schölkopf, B. 2018. Learning independent causal mechanisms. In *International Conference on Machine Learning*, 4036–4044. PMLR.

Rice, J. 1986. Poligon: A System for Parallel Problem Solving. Technical Report KSL-86-19, Dept. of Computer Science, Stanford Univ.

Robinson, A. L. 1980. New Ways to Make Microcircuits Smaller. *Science*, 208(4447): 1019–1022.

Schölkopf, B.; Janzing, D.; Peters, J.; Sgouritsa, E.; Zhang, K.; and Mooij, J. 2012. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*.

Sui, Y.; Wang, X.; Wu, J.; Lin, M.; He, X.; and Chua, T.-S. 2022. Causal attention for interpretable and generalizable graph classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1696–1705.

Trivedi, R.; Yang, J.; and Zha, H. 2020. Graphopt: Learning optimization models of graph formation. In *International Conference on Machine Learning*, 9603–9613. PMLR.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, 10–19.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, B.; Zhang, Y.; Shi, J.; Wang, P.; Wang, X.; Bai, L.; and Wang, Y. 2023. Knowledge Expansion and Consolidation for Continual Traffic Prediction With Expanding Graphs. *IEEE Transactions on Intelligent Transportation Systems*.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.

Wu, Y.-X.; Wang, X.; Zhang, A.; He, X.; and Chua, T.-S. 2022. Discovering invariant rationales for graph neural networks. *arXiv preprint arXiv:2201.12872*.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1): 4–24.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Yang, K.; Zhou, Z.; Sun, W.; Wang, P.; Wang, X.; and Wang, Y. 2023. Extract and refine: Finding a support subgraph set for graph representation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2953–2964.

Yang, S.; Wang, Y.; Van De Weijer, J.; Herranz, L.; and Jui, S. 2020. Unsupervised domain adaptation without source data by casting a bait. *arXiv preprint arXiv:2010.12427*, 1(2): 5.

Ying, Z.; You, J.; Morris, C.; Ren, X.; Hamilton, W.; and Leskovec, J. 2018. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31.

Yuan, H.; Yu, H.; Gui, S.; and Ji, S. 2022. Explainability in graph neural networks: A taxonomic survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(5): 5782–5799.

Zhang, A.; Lyle, C.; Sodhani, S.; Filos, A.; Kwiatkowska, M.; Pineau, J.; Gal, Y.; and Precup, D. 2020. Invariant causal prediction for block mdps. In *International Conference on Machine Learning*, 11214–11224. PMLR.

Zhang, M.; Cui, Z.; Neumann, M.; and Chen, Y. 2018. An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhao, J.; Dong, Y.; Ding, M.; Kharlamov, E.; and Tang, J. 2021. Adaptive diffusion in graph neural networks. *Advances in neural information processing systems*, 34: 23321–23333.

Zitnik, M.; Agrawal, M.; and Leskovec, J. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13): i457–i466.