






Causal Subgraphs and Information Bottlenecks: Redefining OOD Robustness in Graph Neural Networks

Weizhi An¹, Wenliang Zhong¹, Feng Jiang¹, Hehuan Ma¹, and
Junzhou Huang^{1*}

The University of Texas at Arlington, Arlington TX 76010, USA

*jzhuang@uta.edu

Abstract. Graph Neural Networks (GNNs) are increasingly popular in processing graph-structured data, yet they face significant challenges when training and testing distributions diverge, common in real-world scenarios. This divergence often leads to substantial performance drops in GNN models. To address this, we introduce a novel approach that effectively enhances GNN performance in Out-of-Distribution (OOD) scenarios, called **Causal Subgraphs and Information Bottlenecks (CSIB)**. CSIB is guided by causal modeling principles to generate causal subgraphs while concurrently considering both Fully Informative Invariant Features (FIIF) and Partially Informative Invariant Features (PIIF) situations. Our approach uniquely combines the principles of invariant risk minimization and graph information bottleneck. This integration not only guides the generation of causal subgraphs but also underscores the necessity of balancing invariant principles with information compression in the face of various distribution shifts. We validate our model through extensive experiments across diverse shift types, demonstrating its effectiveness in maintaining robust performance under OOD conditions.

Keywords: Graph Neural Network · Invariant Learning · Graph Out of Distribution Generation

1 Introduction

Graph Neural Network (GNN) has become a promising solution for various graph-based learning tasks [23, 30, 32, 39], such as social recommendation [12, 34, 35], drug discovery [15, 21, 22, 31], adversarial robustness [18, 19], and biomedical applications [5, 13, 38]. Despite their widespread success, traditional GNN approaches often rely on the assumption that training and testing sets are from the identical distribution, which may not hold true in real-world scenarios, leading to performance degradation under distribution shifts. Most GNNs presume the in-distribution (ID) assumption and may not perform well in out-of-distribution (OOD) settings. For instance, in drug discovery, models trained on limited data may face challenges when tested on a much larger and diverse

set of molecules, underlining the necessity for GNNs with robust OOD generalization abilities [15, 24]. In graph OOD scenarios, two predominant types of distribution shifts are observed [10]: covariate shift and concept shift. Covariate shift arises when the distribution of node or edge features changes, while the underlying predictive relationships remain constant. In contrast, concept shift occurs when these foundational relationships themselves alter, necessitating a profound re-evaluation of the learned patterns by the GNNs. Although there is a big success of the invariant learning on regular Euclidean data [2, 16, 29] to address feature-level distribution shifts, due to the complex nature of graphs, it is still a challenge for graph generalization. Since shifts on the graph can occur at either the structure level or the feature level, considering how to transfer the Euclidean invariant learning paradigm to graphs is worthwhile.

Another concern is existing methods in graph-based OOD generalization stem from the inherent uncertainty in real-world data regarding the nature of invariant features. In practical scenarios, it is often unclear if the invariant features within a graph are Fully Informative Invariant Features (FIIF) or Partially Informative Invariant Features (PIIF). This ambiguity presents a significant challenge to existing graph generalization methods. Traditional approaches like Empirical Risk Minimization (ERM) [8] and invariant learning methodologies such as Invariant Risk Minimization (IRM) [2], along with recent advancements like Graph Information Bottleneck (GIB) [36, 41], Domain Invariant Representation (DIR) [37] and Graph Stochastic Attention (GSAT) [25], primarily cater to scenarios assuming invariant features are fully informative. However, these methods often inadequately address situations where invariant features are only partially informative, leading to considerable difficulties in accurately identifying causal subgraph in this scenario. In such partially informative scenarios, the causal subgraph does not encompass all the necessary information for accurate prediction of labels across various environments, thereby necessitating a reliance on additional non-causal features. This reliance can lead to instability in model performance across different environments and potentially misguide the model during the learning of invariant features, impacting its generalization ability. Recent causality-based methods for graph-level tasks [4, 6, 40] show promise but still primarily focus on scenarios with FIIF, neglecting the complexities introduced by PIIF. Causality inspired invariant graph learning (CIGA) [4] recognizes the significance of causality for graph-level tasks, yet it does not fully exploit environmental information, which is crucial for identifying invariant subgraphs that are robust across different environments [11].

To address existing limitations in OOD generalization for GNNs, we introduce a novel framework titled **Causal Subgraphs and Information Bottlenecks (CSIB)**. Our proposed CSIB is grounded in the Invariant Principle via Invariant Causal Prediction (ICP), predicated on the assumption that invariant features are generated in accordance with a Structural Causal Model (SCM) [26]. We leverage environmental features within the graph to extract invariant causal subgraphs through the lens of mutual information. However, the mere implementation of the Invariant Principle is insufficient. For instance, when the diver-

sity of environmental conditions is eclipsed by the number of spurious features, traditional methods falter, unable to distinguish between causal and spurious influences. Such a predicament often leads models to incorrectly incorporate both types of features for prediction, a situation typified in the PIIF scenario. Recognizing that most OOD methodologies for graphs neglect this nuance, we integrate the concept of a Graph Information Bottleneck (GIB) within our CSIB framework. This serves as a selective filter applied to our generated subgraphs, ensuring that only the most relevant causal features are preserved for prediction tasks. Theoretical analysis paired with empirical evidence demonstrates that CSIB exhibits superior performance in generating graph OOD representations, adeptly handling a spectrum of shift variations. This attests to the robustness and practicality of our approach in enhancing the OOD generalization capabilities of GNNs. Our main contributions are as follows:

- Our proposed CSIB integrates both Invariant Principle and GIB. The incorporation of information compression in CSIB aids in eliminating potential spurious features, thereby enhancing the generalization across varying shifts. This dual approach empowers our model to discern and leverage invariant features under both fully and partially informative causal structure shifts.
- We propose an end-to-end framework that integrates environmental features into causal graph generation. By minimizing the discrepancy between the predictions of causal and environmental causal graphs, our approach effectively identifies features invariant to environmental changes.
- Empirically, we conduct extensive experiments on five datasets including both synthetic and real-world scenarios. The results demonstrate significant improvements in handling structural and feature-level shifts in the graph and highlight the model’s robustness and versatility.

2 Graph OOD Causal Generation

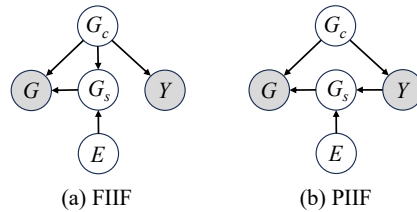


Fig. 1: Demonstrations of structural causal models (SCMs) for FIIF and PIIF. G_s in FIIF is directly controlled by G_c . G_s in PIIF is indirectly controlled by G through Y .

In this study, we address the significant challenge of improving the generalization capability of GNNs in OOD scenarios. Commonly, GNN models presuppose

that the distributions of training and testing data are identical. However, this assumption often does not hold in practical applications, leading to a marked decrease in model performance when faced with data that substantially diverges from the training distribution.

2.1 Problem and Objective in Graph OOD Generation

Our primary research goal is to address OOD generalization challenges in graph data by learning an invariant subgraph G_c that remains consistent across varied environmental conditions and distributional shifts. We consider a collection of graph datasets $\mathcal{D} = \{\mathcal{D}_e\}_{e \in \mathcal{E}_{\text{all}}}$, where each dataset \mathcal{D}_e represents a unique environment e within the full set of environments \mathcal{E}_{all} . Each dataset \mathcal{D}_e consists of graph samples (G, Y) , where G is a graph from environment e and Y is the corresponding label. The central aim is to optimize a GNN model ϕ that is capable of identifying and leveraging an invariant subgraph structure G_c within these graphs. This invariant subgraph G_c should include the essential features of the graphs that are consistent and predictive of the labels Y across all environments, irrespective of the environmental changes or distributional shifts. The challenge lies in ensuring that G_c is robust to environmental variations, thereby enabling the GNN ϕ to generalize well to unseen environments, and maintain its predictive precision.

2.2 Causal Graph Generation under Structural Causal Models

To mitigate the challenge of OOD generation in graphs, we introduce to generate causal graphs with Structural Causal Models (SCMs) as shown in Figure 1. Our primary goal is to identify variables that hold a stable causal relationship with the target variable Y across varying environmental conditions. We extend the invariance principle to our model.

Let G be a graph with a corresponding target variable Y . There exists an invariant causal subgraph $G_c \subseteq G$ encapsulating the causal mechanism $G_c \rightarrow Y$. This causal mechanism is independent of any external environmental factors E , ensuring the conditional distribution $P(Y|G_c)$ is invariant across different environments. Formally:

$$\forall e \in \mathcal{E}_{\text{all}}, \quad P(Y|G_c, E = e) = P(Y|G_c), \quad (1)$$

where \mathcal{E}_{all} represents the set of all possible environments. This indicates that G_c contains all necessary information for consistent prediction of Y , irrespective of environmental variations or shifts.

Building on this assumption, our framework introduces a causal graph generator aimed at uncovering the invariant features within graph data. We leverage environmental features as a form of auxiliary information, integrating them into the graph generation process to further emphasize the invariant characteristics of G_c . To further construct an invariant subgraph G_c that consistently represents the underlying causal mechanisms across different environmental settings,

we define Fully Informative Invariant Features (FIIF) and Partially Informative Invariant Features (PIIF) in Equations 2 and 3, respectively. Identifying variables with a stable causal relationship to the target variable Y under both FIIF and PIIF is essential. The Independent Causal Mechanisms assumption [3, 27] suggests that the process of labeling, represented as $G_c \rightarrow Y$, is not influenced by other external factors. This means that the conditional distribution $P(Y|G_c)$ should remain consistent, regardless of any interventions on the environmental latent variable E . However, extracting the invariant causal subgraph G_c from the overall graph G is challenging, especially when specific information about the environmental variable E is missing. This absence complicates the task of ensuring that the representations learned are independent of E .

- **Fully Informative Invariant Features (FIIF)**: These are features within a graph that provide complete and consistent information about the target variable Y across all environments. Formally, a feature set is considered as FIIF if it satisfies:

$$Y \perp\!\!\!\perp E | G_c, \quad (2)$$

where G_c represents the invariant subgraph that captures all the necessary information to predict Y reliably, irrespective of the environmental variable E . FIIF ensures that the relationship between G_c and Y is stable and unaffected by changes in E .

- **Partially Informative Invariant Features (PIIF)**: These features in a graph provide partial information about the target variable Y and may require additional contextual information for accurate prediction. PIIF is defined by the condition:

$$Y \not\perp\!\!\!\perp E | G_c, G_s, \quad (3)$$

where G_s denotes the spurious graph. In this case, G_c still contains relevant information about Y , but its predictive power is not complete and can be influenced by environmental changes represented by E . PIIF indicates that while G_c is informative, it may not be sufficient on its own to account for the variability in Y across different environments.

Traditional approaches such as IRM [2], and recent developments like GIB [36, 41], DIR [37], and GSAT [25] primarily focus on scenarios characterized with FIIF. These methods are under the assumption that invariant features comprehensively inform the target variable across different environments. However, they may falter in PIIF contexts where invariant features only partially convey essential information for accurate predictions. Although CIGA [4] takes PIIF into account, it does not leverage environmental features, limiting its effectiveness in OOD generalization. This highlights the importance of discerning between FIIF and PIIF, a critical aspect for advancing OOD generalization in graph-based data. Specifically, in the SCMs as shown in Figure 1 for FIIF, the spurious graph G_s is directly influenced by the causal graph G_c . In contrast, in PIIF scenarios, G_s is affected by G_c indirectly through label Y . In PIIF scenarios,

the causal subgraph G_c does not fully capture all necessary information for accurate prediction of Y across environments. Some subgraphs of G_s may hold additional information about Y beyond what G_c offers, leading to a potential incorporation of parts of G_s into G_c . This partial informativeness necessitates reliance on additional non-causal features, potentially leading to instability in model performance across various shifts. It can misguide the model during the learning of G_c , leading it to over-rely on these unstable, non-causal features, thereby affecting its generalization capability.

3 Graph Invariant Causal Generation

Our CSIB introduces a novel methodology for Graph OOD generation. We implement the invariance principle, and generate the invariant causal graph, forming the backbone of our OOD generalization strategy. To further refine this process and overcome the limitations inherent in previous approaches, we also incorporate the concept of information bottleneck. The details of these implementations are described in this section.

3.1 Invariant Causal Graph Extraction

The first phase involves the identification of the invariant causal graph G_c within the given graph G , utilizing a specially designed GNN, denoted as g_ϕ . This process is designed to discern the subgraph structure G_c that inherently contains features invariant across diverse environments. During training, edge selection is guided by stochastic sampling from Bernoulli distributions, enabling the generation of subgraphs that most contribute to the label prediction. This mechanism highlights the subgraphs relevant to stable predictions, thereby capturing the essential invariant structures within the graph.

From an information-theoretic interpretation [14], IRM [2] can be understood as the process of identifying features within data that maintain a consistent predictive relationship with the output variable, irrespective of the environment. This interpretation aligns with our objective of extracting an invariant causal graph G_c from the input graph G . By maximizing the mutual information between Y and G_c , and concurrently minimizing the conditional mutual information between Y and environmental variables given G_c , our model adheres to the invariant principle. This approach ensures the extraction of features from G that are not only informative about Y but also stable across different environments. Thus, we aim to maximize the mutual information between the graph’s label Y and the invariant causal graph G_c , while also considering the conditional mutual information between Y and environmental factors E given G_c . The objective is formalized as follows:

$$\max_{\phi, \theta} I(Y; G_c) - \beta I(Y; E | G_c). \quad (4)$$

This formulation aims to ensure that the model captures the crucial information from G_c for predicting Y and takes into account environmental influences that

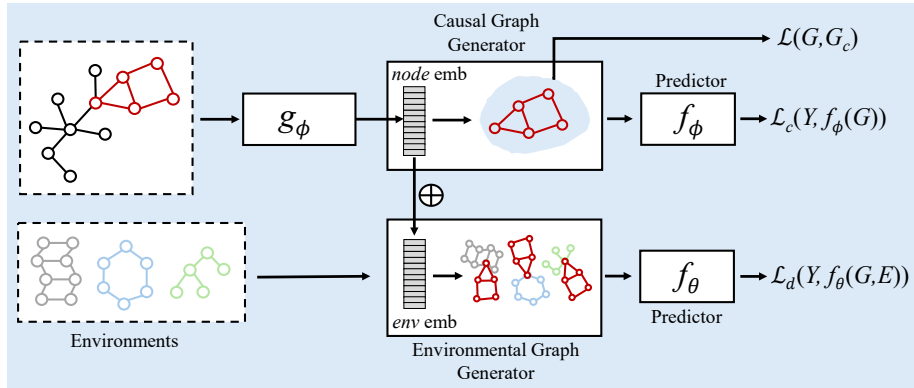


Fig. 2: The CSIB framework comprises three key components: a causal graph predictor f_ϕ , a conditional environmental predictor f_θ , and our GNN g_ϕ . Specifically, the causal graph can be generated using a Causal Graph Generator. Environmental graph features are concatenated with the output of g_ϕ and then fed into the Environmental Graph Generator to sample environmental conditioned graphs. This process aligns with our optimization objective $\max_{\phi, \theta} I(Y; G_c) - \beta I(Y; E | G_c) - \lambda I(G; G_c)$, where the three loss terms correspond to the respective components of our model.

might affect this prediction. For $I(Y; E | G_c)$, we have :

$$I(Y; E | G_c) = H(Y | G_c) - H(Y | E, G_c), \quad (5)$$

where we dissect the $I(Y; E | G_c)$ into $H(Y | G_c)$ and $H(Y | E, G_c)$. We utilize an i.i.d. sample set $(y_i, G_{c_i})_{i=1}^N$ from the joint distribution $p(y, G_c)$, approximating these terms using the empirical distribution. This decomposition $H(Y | G_c)$ and $H(Y | E, G_c)$ are represented as:

$$H(Y | G_c) = \frac{1}{N} \sum_{i=1}^N \log q(y_i | G_{c_i}), \quad (6)$$

$$H(Y | E, G_c) = \frac{1}{N} \sum_{i=1}^N \log q(y_i | G_{c_i}, e_i), \quad (7)$$

where $q(y | G_c)$ and $q(y | G_c, e)$ function as variational approximations of the actual conditional probabilities $p(y | G_c)$ and $p(y | G_c, E)$, respectively. This step is crucial for operationalizing the mutual information in a computationally feasible manner, ensuring the practicality of our approach.

3.2 Causal Graph Generator and Environmental Graph Generator on GNNs

In our approach, the causal subgraph G_c from the input graph G is facilitated through a GNN g_ϕ . It processes the graph G to generate node representations

$\{h_v | v \in V\}$, where V represents the set of nodes in the graph. Following the Gilbert random graph theory [9], we model each edge in G with a binary random variable r_{ij} , indicating the presence $r_{ij} = 1$ or absence of an edge between nodes v_i and v_j . The probability p_{uv} for each edge is computed using a Multi-layer Perception (MLP) layer that applies a sigmoid function to the concatenated node representations (h_u, h_v) to a probability score $p_{uv} \in [0, 1]$.

To facilitate gradient-based optimization and enable stochasticity in edge selection, we employ a categorical reparameterization for r_{ij} :

$$r_{ij} = \text{Sigmoid} \left(\log \varepsilon - \log(1 - \varepsilon) + \frac{\alpha_{ij}}{\tau} \right), \quad (8)$$

where $\varepsilon \sim \text{Uniform}(0, 1)$, α_{ij} is the logit of the edge existence probability, and τ is a temperature parameter controlling the approximation accuracy. The causal subgraph G_c is extracted based on the edge probabilities p_{uv} , with its adjacency matrix A_c derived from the original matrix A and the edge scores. This process effectively identifies the relevant connections and nodes that contribute to the prediction task. Finally, the prediction model f_ϕ utilizes the extracted subgraph G_c to make predictions about the label Y .

It is known that OOD generation is impracticable without environmental information [1, 11, 26]. In the absence of environmental information, distinguishing between the causal subgraph G_c and the spurious subgraph G_s becomes challenging, leading to potential confusion in their identification. In our framework, the causal graph generation on GNNs extends beyond structural features of G by incorporating environmental factors. We augment the graph features with extracted environmental features enhancing the model’s capacity to account for external influences. This concatenation enriches our graph representation, aligning it with the specific context of each environment. The enriched graph representation, now comprising both graph and environmental features, is used to sample environmentally conditioned subgraphs. These subgraphs, embodying comprehensive environmental and structural data, are then fed into our prediction model f_θ , which is adept at leveraging these enriched subgraphs for making predictions about Y . The introduction of environmental features into Equation 4 is pivotal to our invariant learning strategy. It allows us to minimize the influence of environmental variables on the causal graph, thereby learning a representation of G_c that is invariant across different environments.

3.3 Introduce Graph Information Bottleneck into Graph OOD Generation Failures

Incorporating the Graph Information Bottleneck (GIB) principle into the Graph OOD generation process addresses inherent limitations, particularly under PIIF scenarios. As illustrated in Figure 1, certain subgraphs within the spurious graph G_s may inadvertently influence the target variable Y , complicating the OOD generalization task. In the context of PIIF, the target label Y is affected not only by the invariant causal subgraph G_c but also by a subset of the spurious graph. While the optimization of Equation 4 aids in isolating the invariant causal

subgraph G_c , it may neglect the influence of the spurious graph, which also contributes to predicting Y . This oversight can lead to a model that, despite being trained to focus on invariant features, still inadvertently captures non-causal or spurious features that can degrade its performance in unseen environments.

To relieve the failure model, we propose to select subgraphs with the least capacity, i.e., those that minimize the mutual information $I(G, G_c)$. This selection criterion leads us to an objective formulation that integrates both the Invariant principle and the GIB principle [41]:

$$\max_{\phi, \theta} I(Y; G_c) - \beta I(Y; E|G_c) - \lambda I(G; G_c), \quad (9)$$

To effectively operationalize this objective, we introduce a variational distribution $q(G_S)$ to approximate the mutual information term $I(G; G_c)$, leading to an upper bound that can be tractably optimized. This is expressed as:

$$I(G; G_c) \leq \mathbb{E}_{p(G)} [\text{KL}(p_\alpha(G_S | G) || q(G_S))] \quad (10)$$

$$= \mathbb{E}_{p(G)} \left[\sum_{i,j=1}^N \text{KL}(p_\alpha(e_{ij} | G) || q(e_{ij})) \right] \quad (11)$$

$$= \mathbb{E}_{p(G)} \left[\sum_{i,j=1}^N \log \frac{p_\alpha(e_{ij} | G)}{q(e_{ij})} \right], \quad (12)$$

where $p_\alpha(G_S | G)$ represents the causal graph generator, and $q(G_S)$ is typically set as $q(G_S) = C \cdot \prod_{i,j=1}^N p_\pi(e_{ij})$, $e_{ij} \sim \text{Bern}(\pi)$, with C being a constant determined by the hyper-parameter π .

The regularization parameters β and λ play a crucial role in balancing the trade-offs between maintaining adherence to the invariance principle and imposing the information bottleneck constraint. By carefully adjusting these parameters, our model aims to capture the invariant structure within G that is predictive of Y while minimizing the influence of environmental factors and extraneous information, thereby enhancing the model’s generalization capability across diverse and unseen environments.

3.4 Overall Optimization Objective

The overarching goal of our CSIB is to optimize the graph neural network model to ensure robust OOD generalization. We achieve this by focusing on the identification of an invariant causal subgraph G_c and mitigating the influence of spurious features under both FIIF and PIIF. Our optimization objective combines both the principles of invariant and GIB constraint, leading to a tractable and effective learning framework for graph data.

Our optimization objective can be formulated as:

$$\begin{aligned} \min_{g_\phi, f_\phi} \max_{f_\theta} & \mathcal{L}_c(Y, f_\phi(G)) + \beta \mathcal{L}(G, g_\phi(G)) \\ & + \lambda (\mathcal{L}_c(Y, f_\phi(G)) - \mathcal{L}_d(Y, f_\theta(G, E))), \end{aligned} \quad (13)$$

where causal loss is $\mathcal{L}_c(Y, f_\phi(G)) = \frac{1}{N} \sum_{i=1}^N \log q(y_i | G_{c_i})$. By minimizing the causal loss, we ensure that the model focuses on those subgraph features that have a consistent and causal impact on Y . Environmental loss $\mathcal{L}_d(Y, f_\theta(G, E)) = \frac{1}{N} \sum_{i=1}^N \log q(y_i | G_{c_i}, e_i)$ accounts for the potential influence of environmental variables E on the prediction. By incorporating environmental loss, we ensure that the model can leverage environmental context when beneficial while maintaining its focus on invariant causal relationships. The information constraint $\mathcal{L}(G, g_\phi(G)) = \mathbb{E}_{p(G)} [\text{KL}(p_\alpha(G_S | G) || q(G_S))]$ aims to retain only the most crucial information within G_c necessary for predicting Y , and it is achieved by minimizing the mutual information between the input graph G and the causal subgraph G_c .

4 Experiments

In our experimental evaluation, we validate the effectiveness of our CSIB in graph OOD generalization. The experiments are structured to answer three questions:

- 1) How does CSIB perform compared to general OOD generation methods?
- 2) How does it stand against recent graph-specific OOD generation methods?
- 3) What is the impact of incorporating invariance principles and GIB?

4.1 Datasets

We evaluated our model on five datasets from the GOOD benchmark [10]: CMNIST-color, Motif-size, Motif-base, HIV-scaffold, and HIV-size. GOOD-HIV is a real-world molecular dataset characterized by scaffold and size domain shifts. The scaffold domain (HIV-scaffold) is based on the Bemis-Murcko scaffold, which represents the two-dimensional structural core of a molecule, while the size domain (HIV-size) is determined by the number of nodes in a molecular graph. GOOD-Motif, a synthetic dataset, is crafted for structural shifts, with graphs generated by combining a base graph and a motif, where the motif solely determines the label. The shift domains include the base graph type (Motif-base) and graph size (Motif-size). GOOD-CMNIST, a semi-synthetic dataset, is designed for node feature shifts and comprises image-derived graphs with manually applied color features (CMNIST-color), making the color shift domain irrelevant to the structure.

4.2 Implementation Details

In our experiments, we use GIN as the backbone model across all tests to maintain consistency. The optimal checkpoints were determined during the OOD validation phase and subsequently applied for OOD testing. Our experiments were conducted over three runs, each with a different random seed. We set the learning rate at 0.001 and limited the training to 200 epochs. For batch sizes, we used 32 for GOOD-Motif and GOOD-HIV datasets, and 128 for the GOOD-CMNIST dataset.

Table 1: Performance on synthetic and real-world datasets. Numbers in **bold** indicate the best performance, while underlined numbers indicate the second best.

Method	Synthetic		Semi-synthetic	Real-world	
	Motif-base	Motif-size	CMNIST-color	HIV-scaffold	HIV-size
ERM [8]	68.66 ± 4.25	51.74 ± 2.88	28.60 ± 1.87	69.58 ± 2.51	59.94 ± 2.37
IRM [2]	70.65 ± 4.17	51.41 ± 3.78	27.83 ± 2.13	67.97 ± 1.84	59.00 ± 2.92
GroupDRO [2]	68.24 ± 8.92	51.95 ± 5.86	29.07 ± 3.14	70.64 ± 2.57	58.98 ± 2.16
VREx [16]	71.47 ± 6.69	52.67 ± 5.54	28.48 ± 2.87	70.77 ± 2.84	58.53 ± 2.88
DIR [37]	62.07 ± 8.75	52.27 ± 4.56	33.20 ± 6.17	68.07 ± 2.29	58.08 ± 2.31
CAL [33]	65.63 ± 4.29	51.18 ± 5.60	27.99 ± 3.24	67.37 ± 3.61	57.95 ± 2.24
GSAT [25]	62.80 ± 11.41	53.20 ± 8.35	28.17 ± 1.26	68.66 ± 1.35	58.06 ± 1.98
OOD-GNN [17]	61.10 ± 7.87	52.61 ± 4.67	26.49 ± 2.94	70.46 ± 1.97	<u>60.60 ± 3.77</u>
StableGNN [7]	57.07 ± 14.10	46.93 ± 8.85	28.38 ± 3.49	68.44 ± 1.33	56.71 ± 2.79
CIGA [4]	66.43 ± 11.31	49.14 ± 8.34	<u>32.22 ± 2.67</u>	69.40 ± 2.39	59.55 ± 2.56
DisC [6]	51.08 ± 3.08	50.39 ± 1.15	24.99 ± 1.78	68.07 ± 1.75	58.76 ± 0.91
DropEdge [28]	45.08 ± 4.46	45.63 ± 4.61	22.65 ± 2.90	<u>70.78 ± 1.38</u>	58.53 ± 1.26
GREa [20]	56.74 ± 9.23	<u>54.13 ± 10.02</u>	29.02 ± 3.26	67.79 ± 2.56	60.71 ± 2.20
CSIB (Ours)	<u>70.18 ± 7.15</u>	60.99 ± 5.57	37.40 ± 2.24	72.53 ± 2.01	61.33 ± 3.77

4.3 Baselines

In our experiments, we compare our method with general OOD generation methods, graph generation strategies, and graph augmentation methodologies. Additionally, we include ERM [8] as a foundational baseline. General OOD generation algorithms include IRM [2], GroupDRO [29], VREx [16]. Graph generation algorithms include DIR [37], CAL [33], GSAT [25], OOD-GNN [17], StableGNN [7], CIGA [4] and DisC [6]. For graph augmentation, we evaluate against DropEdge [28] and GREa [20].

4.4 Experimental Analysis under Various OOD Scenarios

Our experimental assessment demonstrates the CSIB method’s capability in addressing OOD challenges across both synthetic and real-world graph datasets, thereby substantiating the effectiveness of our approach.

Feature Level Shifts In the context of feature shifts, the CMNIST dataset acts as a critical benchmark. Our model not only eclipses current methodologies but also outperforms the state-of-the-art (SOTA) by a significant 12.7% margin, as shown in Table 1. This remarkable performance gain is especially pronounced when compared to approaches like CIGA, which also address PIIF scenarios but perhaps do not fully exploit environmental cues. The success achieved on the CMNIST dataset is a direct consequence of our model’s proficient utilization of environmental features. This accomplishment aligns seamlessly with our

method’s core focus on fortifying invariance against fluctuations in feature-level distributions, thereby substantiating our initial hypothesis concerning the pivotal importance of addressing both FIIF and PIIF for OOD generalization.

Structural Level Shifts The GOOD-Motif dataset, designed to test structural shifts, presents a different set of challenges. In base domain scenarios, which feature three distinct structural environment attributes, our model demonstrated its ability to produce results consistent with established benchmarks. This achievement emphasizes our approach’s adept handling of structural variations inherent to graph data, underscoring its adaptability and effectiveness in accommodating a variety of structural contexts within graph-based datasets. The model’s robustness was further evidenced in the size domain shift, where it surpassed prevailing methods by a significant margin of 12.6%. This success underscores the model’s resilience to variations in structural scale, attributed to the stable nature of the invariant causal graph G_c amidst environmental changes. Despite the varying size scales across different environments, the invariant causal graph G_c remains consistent, anchoring the model’s focus on crucial causal features that transcend environmental variations. This detailed evaluation illustrates our model’s ability to navigate and adapt to structural challenges, affirming its suitability for complex graph data scenarios.

Real-world Data Performance In the challenging environment of the real-world HIV dataset, our model showcased its robustness and dependability, outperforming prior state-of-the-art methods in both the scaffold and size domains. This achievement attests to the model’s capability in navigating the intricacies of complex real-world data that often exhibit a blend of feature and structural shifts. This success in a real-world context further emphasizes the robustness and efficacy of our approach, underlining its strong alignment with the requirements for OOD generalization in GNNs. Our extensive experimental evaluation across a range of datasets reinforces the central premise of our investigation. By combining invariant principles with the GIB technique, we effectively address the nuanced challenges of FIIF and PIIF scenarios. This thorough validation not only demonstrates our model’s ability to maintain high predictive accuracy and robustness across diverse OOD conditions but also represents a significant advancement in graph-based learning methodologies.

4.5 Ablation Study

To further interrogate the efficacy of our model’s components, specifically addressing "why invariance" and "why information constraint", we conducted experiments on both the HIV datasets, exploring scaffold shift and size shift scenarios, and the CMNIST dataset, examining color concept and covariate shifts as shown in Fig. 3. Our experiments aimed to validate the effectiveness of our model’s components in mitigating these challenges.

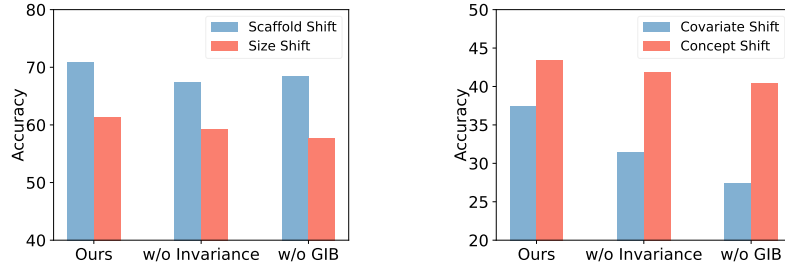


Fig. 3: *Left:* Ablation study on HIV Dataset under Scaffold and Size Shifts. *Right:* Ablation study on CMNIST-color Dataset under Covariate and Concept Shifts.

Our findings revealed that in the absence of the Invariance Principle guidance, particularly when we did not apply environmental label-guided environmental graph generation, our model experienced a significant performance drop of 16.1% on the CMNIST dataset. This decrease was notably pronounced when the CMNIST dataset exhibited color shift scenarios, suggesting that the introduction of environmental information could alleviate this issue. Moreover, further removal of the information constraint led to a substantial decrease in model performance. As discussed earlier, the information compression constraint effectively filters out irrelevant graph features, particularly in PIIF scenarios. Our results underscore the importance of these principles, as evidenced by the observed performance decrements under concept shifts. We also conducted an ablation study on the real-world drug dataset HIV. The performance on the HIV datasets drop without either the invariant principle or GIB, aligning well with our propositions.

Hyperparameter Sensitivity Study In further investigation, we delve into the impact of hyperparameters β and λ on the efficacy of our CSIB method, particularly within the contexts of the GOOD-Motif and CMNIST datasets. These hyperparameters play pivotal roles: β modulates the integration of environmental information into the invariant learning process, while λ governs the strength of the information bottleneck constraint. Adjusting these parameters offers insights into the balance between capturing invariant features and mitigating the influence of spurious correlations. In our hyperparameter sensitivity analysis, we focused on the parameters β and λ , identified through preliminary experiments as having optimal values of 0.1 and 0.01, respectively. To rigorously assess the impact of these hyperparameters on our model performance, we systematically varied them across a set range of values: 0, 0.001, 0.01, 0.1, and 1. This comprehensive exploration allowed us to observe the model’s behavior under diverse settings. In Fig. 4, we demonstrate the robustness of the CSIB method to variations in hyperparameters β and λ . Our analysis reveals that while the CSIB method maintains stability across a range of hyperparameter settings, it exhibits sensitivity to the extremities of λ ’s value spectrum. Specifically, perfor-

mance degrades when λ is set too high or too low, highlighting the importance of balanced regularization to mitigate the model’s susceptibility to spurious correlations inherent in distribution shifts. Conversely, the impact of the information constraint parameter β becomes more pronounced at higher values (0.1 and 1), suggesting its efficacy in filtering out irrelevant information and reinforcing the model’s focus on invariant features.

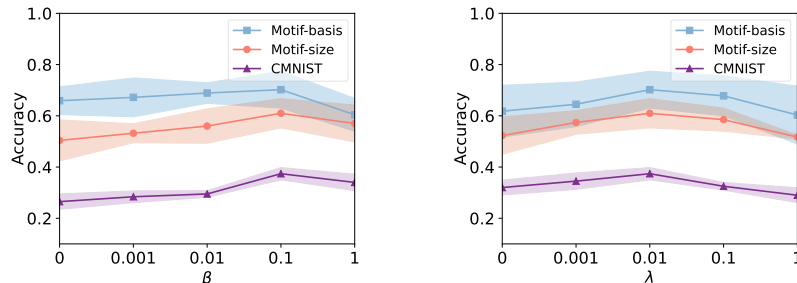


Fig. 4: Sensitivity analysis of β (left) and λ (right). For each sensitivity study, we fix other hyperparameters with the values selected from the previous experiments.

5 Conclusion

We introduce **CSIB (Causal Subgraphs and Information Bottlenecks)**, a novel framework designed to enhance Out-of-Distribution (OOD) generalization in Graph Neural Networks. Our approach is grounded in the integration of invariant causal graph generation and the information bottleneck principle, addressing the critical challenge of identifying invariant features in graph data that are reliable predictors across diverse environments. Key contributions of our work include the development of an end-to-end framework that effectively leverages environmental features into the causal graph generation process. This framework employs mutual information theory to optimize the model, focusing on extracting invariant causal graphs that capture the essential, environment-independent structures within the graph data. Furthermore, our incorporation of the information bottleneck principle allows for the compression of the graph representation, effectively filtering out spurious features that could otherwise lead to model instability and reduced predictive accuracy in OOD scenarios. Empirically, our extensive experiments across various datasets, including both synthetic and real-world scenarios, demonstrate the efficacy of CSIB in handling different types of distribution shifts. Our framework shows significant improvements in managing structural and feature-level shifts, underscoring its robust generalization capability.

Acknowledgements

This work was partially supported by US National Science Foundation IIS-2412195, CCF-2400785 and the Cancer Prevention and Research Institute of Texas (CPRIT) award (RP230363).

References

1. Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.C., Bengio, Y., Mitliagkas, I., Rish, I.: Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems* **34**, 3438–3450 (2021)
2. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019)
3. Chang, S., Zhang, Y., Yu, M., Jaakkola, T.: Invariant rationalization. In: *International Conference on Machine Learning*. pp. 1448–1458. PMLR (2020)
4. 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. *Advances in Neural Information Processing Systems* **35**, 22131–22148 (2022)
5. Ding, K., Zhou, M., Wang, Z., Liu, Q., Arnold, C.W., Zhang, S., Metaxas, D.N.: Graph convolutional networks for multi-modality medical imaging: Methods, architectures, and clinical applications. *arXiv preprint arXiv:2202.08916* (2022)
6. Fan, S., Wang, X., Mo, Y., Shi, C., Tang, J.: Debiasing graph neural networks via learning disentangled causal substructure. *Advances in Neural Information Processing Systems* **35**, 24934–24946 (2022)
7. 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* (2023)
8. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**(11), 665–673 (2020)
9. Gilbert, E.N.: Random graphs. *The Annals of Mathematical Statistics* **30**(4), 1141–1144 (1959)
10. Gui, S., Li, X., Wang, L., Ji, S.: Good: A graph out-of-distribution benchmark. *Advances in Neural Information Processing Systems* **35**, 2059–2073 (2022)
11. Gui, S., Liu, M., Li, X., Luo, Y., Ji, S.: Joint learning of label and environment causal independence for graph out-of-distribution generalization. *Advances in Neural Information Processing Systems* **36** (2024)
12. He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., Wang, M.: Lightgcn: Simplifying and powering graph convolution network for recommendation. In: *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. pp. 639–648 (2020)
13. Huang, J., Li, R.: Adaptive graph convolutional neural network and its biomedical applications. In: *State of the Art in Neural Networks and Their Applications*, pp. 105–132. Elsevier (2023)
14. Huszar, F.: Invariant risk minimization: An information theoretic view (2019)

15. Ji, Y., Zhang, L., Wu, J., Wu, B., Huang, L.K., Xu, T., Rong, Y., Li, L., Ren, J., Xue, D., et al.: Drugood: Out-of-distribution (ood) dataset curator and benchmark for ai-aided drug discovery—a focus on affinity prediction problems with noise annotations. arXiv preprint arXiv:2201.09637 (2022)
16. Krueger, D., Caballero, E., Jacobsen, J.H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., Courville, A.: Out-of-distribution generalization via risk extrapolation (rex). In: International Conference on Machine Learning. pp. 5815–5826. PMLR (2021)
17. Li, H., Wang, X., Zhang, Z., Zhu, W.: Ood-gnn: Out-of-distribution generalized graph neural network. IEEE Transactions on Knowledge and Data Engineering (2022)
18. Li, K., Liu, Y., Ao, X., Chi, J., Feng, J., Yang, H., He, Q.: Reliable representations make a stronger defender: Unsupervised structure refinement for robust gnn. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 925–935 (2022)
19. Li, K., Liu, Y., Ao, X., He, Q.: Revisiting graph adversarial attack and defense from a data distribution perspective. In: The Eleventh International Conference on Learning Representations (2022)
20. Liu, G., Zhao, T., Xu, J., Luo, T., Jiang, M.: Graph rationalization with environment-based augmentations. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1069–1078 (2022)
21. Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., Tang, J.: Pre-training molecular graph representation with 3d geometry. arXiv preprint arXiv:2110.07728 (2021)
22. Ma, H., An, W., Wang, Y., Sun, H., Huang, R., Huang, J.: Deep graph learning with property augmentation for predicting drug-induced liver injury. Chemical research in toxicology **34**(2), 495–506 (2020)
23. Ma, H., Bian, Y., Rong, Y., Huang, W., Xu, T., Xie, W., Ye, G., Huang, J.: Cross-dependent graph neural networks for molecular property prediction. Bioinformatics **38**(7), 2003–2009 (2022)
24. Ma, H., Jiang, F., Rong, Y., Guo, Y., Huang, J.: Robust self-training strategy for various molecular biology prediction tasks. In: Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. pp. 1–5 (2022)
25. Miao, S., Liu, M., Li, P.: Interpretable and generalizable graph learning via stochastic attention mechanism. In: International Conference on Machine Learning. pp. 15524–15543. PMLR (2022)
26. Pearl, J.: Causality. Cambridge university press (2009)
27. Peters, J., Bühlmann, P., Meinshausen, N.: Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society Series B: Statistical Methodology **78**(5), 947–1012 (2016)
28. Rong, Y., Huang, W., Xu, T., Huang, J.: Dropedge: Towards deep graph convolutional networks on node classification. arXiv preprint arXiv:1907.10903 (2019)
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. arXiv preprint arXiv:1911.08731 (2019)
30. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE transactions on neural networks **20**(1), 61–80 (2008)
31. Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., Tang, J.: Graphaf: a flow-based autoregressive model for molecular graph generation. arXiv preprint arXiv:2001.09382 (2020)

32. Shi, Y., Ma, H., Zhong, W., Tan, Q., Mai, G., Li, X., Liu, T., Huang, J.: Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. In: 2023 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 515–520. IEEE (2023)
33. Sui, Y., Wang, X., Wu, J., Lin, M., He, X., Chua, T.S.: Causal attention for interpretable and generalizable graph classification. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1696–1705 (2022)
34. Wang, X., He, X., Wang, M., Feng, F., Chua, T.S.: Neural graph collaborative filtering. In: Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval. pp. 165–174 (2019)
35. Wu, S., Sun, F., Zhang, W., Xie, X., Cui, B.: Graph neural networks in recommender systems: a survey. *ACM Computing Surveys* **55**(5), 1–37 (2022)
36. Wu, T., Ren, H., Li, P., Leskovec, J.: Graph information bottleneck. *Advances in Neural Information Processing Systems* **33**, 20437–20448 (2020)
37. Wu, Y.X., Wang, X., Zhang, A., He, X., Chua, T.S.: Discovering invariant rationales for graph neural networks. arXiv preprint arXiv:2201.12872 (2022)
38. Yan, Y., He, S., Yu, Z., Yuan, J., Liu, Z., Chen, Y.: Investigation of customized medical decision algorithms utilizing graph neural networks. arXiv preprint arXiv:2405.17460 (2024)
39. Yang, J., Zhao, P., Rong, Y., Yan, C., Li, C., Ma, H., Huang, J.: Hierarchical graph capsule network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 10603–10611 (2021)
40. Yang, N., Zeng, K., Wu, Q., Jia, X., Yan, J.: Learning substructure invariance for out-of-distribution molecular representations. *Advances in Neural Information Processing Systems* **35**, 12964–12978 (2022)
41. Yu, J., Xu, T., Rong, Y., Bian, Y., Huang, J., He, R.: Graph information bottleneck for subgraph recognition. arXiv preprint arXiv:2010.05563 (2020)