A Recipe for Causal Graph Regression: Confounding Effects Revisited

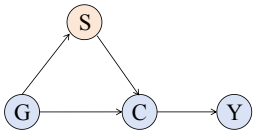
Yujia Yin [*1] **Tianyi Qu** [*2 3] **Zihao Wang** [4] **Yifan Chen** [1]

Abstract

Through recognizing causal subgraphs, causal graph learning (CGL) has risen to be a promising approach for improving the generalizability of graph neural networks under out-of-distribution (OOD) scenarios. However, the empirical successes of CGL techniques are mostly exemplified in classification settings, while regression tasks, a more challenging setting in graph learning, are overlooked. We thus devote this work to tackling causal graph regression (CGR); to this end we reshape the processing of confounding effects in existing CGL studies, which mainly deal with classification. Specifically, we reflect on the predictive power of confounders in graph-level regression, and generalize classification-specific causal intervention techniques to regression through a lens of contrastive learning. Extensive experiments on graph OOD benchmarks validate the efficacy of our proposals for CGR. The model implementation and the code are provided on https://github.com/causal-graph/CGR.

1. Introduction

Causal graph learning (CGL) (Lin et al., 2021) holds particular importance due to its relevance in fields such as drug discovery (Qiao et al., 2024) and climate modeling (Zhao et al., 2024). However, previous CGL studies focus on classification settings. Some of them cannot be directly extended to regression tasks, such as property prediction (Rollins et al., 2024), traffic flow forecasting (Li et al., 2021), and credit risk scoring (Ma et al., 2024), because the transition from finite to infinite support makes discrete labels unavailable. Graphs thus cannot be informatively grouped. A systematical understanding of how CGL techniques should be adapted to graph-level regression is still under-explored.



G : Full Graph
C : Causal Subgraph
S : Confounding Subgraph
Y : Response

Figure 1. Structural causal model (SCM) for graph regression.

The core methodology of causal learning involves the identification and differentiation of causal features from confounding ones. As shown in Figure 1, causal features C are those directly deciding responses Y, whereas confounding features S (shorthand for "spurious") solely present spurious correlations. Therefore, understanding how causal features (as well as confounding features) and responses interact plays a central role in practical designing of causal learning methods. From this perspective, causal graph regression (CGR) warrants specialized handling since the interaction between features and responses therein is significantly different from classification. Furthermore, regression is in general a more challenging task than classification, and techniques working for classification, Perceptron (Rosenblatt, 1958) for example, may not apply to regression.

Specifically in CGL, the identification of causal subgraphs is seemingly transferable since this step, explicitly or implicitly, relies on the calculation of mutual information and is compatible with both settings (c.f. Section 3.2). However, the empirical performance of this vanilla adaptation on regression tasks is dwarfed by empirical risk minimization (Vapnik, 1991, ERM) w.r.t. least squares loss (see the results in Sections 5.3 and 5.4).

To crack CGR, we revisit the processing of confounding effects, which conceptually constitutes causal graph learning along with causal subgraph identification as shown in Figure 1. Existing CGL methods, such as CAL (Sui et al., 2022) and DisC (Fan et al., 2022), are built on a strong assumption that confounding subgraphs contain strictly no predictive power. We reflect on this assumption and speculate it is hardly practical due to the contradiction with real-world observations: in molecular property prediction, for example, molecular weight is noncausal to toxicity while does exhibit strong correlations.

[*]Equal contribution [1]Hong Kong Baptist University [2]SF Tech [3]Zhejiang University [4]Hong Kong University of Science and Technology. Correspondence to: Tianyi Qu <qutianyi@sf-express.com>, Yifan Chen <yifanc@hkbu.edu.hk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

In this work, we develop an enhanced graph information bottleneck (GIB) loss function, which no longer takes the strong assumption. Moreover, some confounding effect processing techniques, such as backdoor adjustment (Sui et al., 2022; 2024) and counterfactual reasoning (Guo et al., 2025), heavily rely on discrete label information and cannot be adapted to regression at all. We follow the principle of those methods and generalize it from class separation to instance discrimination; the discrimination principle aligns with the philosophy of contrastive learning (CL) and CL techniques are therefore leveraged to tackle CGR in our proposal.

Following the intuition, we develop a new framework for causal graph regression, which spotlights the confounding effects within. In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to explicitly consider the predictive role of confounding features in graph regression tasks, a critical yet overlooked aspect in graph OOD generalization.
- We introduce a new causal intervention approach that generates random graph representations by leveraging a contrastive learning loss to enhance causal representation, outperforming label-dependent methods.
- Extensive experiments on OOD benchmarks demonstrate that our method significantly improves generalization in graph regression tasks.

2. Related Work

Out-of-distribution (OOD) challenges in graph learning has drawn significant attention, particularly in methods aiming to disentangle causal and confounding factors (Ma, 2024). Existing approaches can be broadly categorized into invariant learning (Wu et al., 2022a), causal modeling (Sui et al., 2024), and stable learning (Li et al., 2022).

Invariant learning focuses on identifying features that remain stable across different environments, filtering out spurious correlations in the process. While not explicitly grounded in causal reasoning, prior studies (Wang & Veitch, 2022; Mitrovic et al., 2020) have highlighted its inherent connection to causality. Methods in invariant learning, such as CIGA (Chen et al., 2022), GSAT (Miao et al., 2022), and GALA (Chen et al., 2024), aim to learn invariant representations by isolating causal components.

However, these approaches are typically designed for classification tasks, limiting their out-of-distribution (OOD) generalization capability in regression settings. Post-hoc methods, such as PGExplainer (Luo et al., 2020) and RegExplainer (Zhang et al., 2023), attempt to discover invariant subgraphs after training. However, these methods fail to equip the model with the ability to learn invariant representations during the training process.

Causal modeling leverages structural causal models (SCMs) to improve the performance of graph neural networks (GNNs) on out-of-distribution (OOD) data. These approaches incorporate various traditional causal inference techniques, such as backdoor adjustment (e.g., CAL (Sui et al., 2022), CAL+ (Sui et al., 2024)), frontdoor adjustment (e.g., DSE (Wu et al., 2022c)), instrumental variables (e.g., RCGRL (Gao et al., 2023)), and counterfactual reasoning (e.g., DisC (Fan et al., 2022)). By simulating causal interventions through supervised training, these methods aim to achieve OOD generalization. However, they often disregard the predictive potential of confounding features, which hinders effective disentanglement. Moreover, the supervised loss functions tailored for classification tasks are not easily adaptable to regression problems, as the inherent complexity of regression introduces additional challenges.

Stable learning aims to ensure consistent performance across environments by reweighting samples or balancing covariate distributions. For example, StableGNN (Fan et al., 2023) employs a regularizer to reduce the influence of confounding variables. However, such methods often rely on heuristic reweighting strategies, which may not fully disentangle causal from confounding factors.

In addition to graph-based approaches, traditional machine learning methods have also explored causality in regression tasks. For instance, Pleiss et al. (2019) observed that causal features tend to concentrate in a low-dimensional subspace, whereas non-causal features are more randomly distributed. Similarly, Amini et al. (2020) proposed a framework for learning continuous targets by placing an evidence prior on a Gaussian likelihood function and training a non-Bayesian neural network to infer the hyperparameters of the evidence distribution. These methods highlight the potential of leveraging causal insights for improved regression performance.

3. Preliminaries and Notations

Along this paper, we denote a graph G as $(\boldsymbol{A}, \boldsymbol{X})$. Here, $\boldsymbol{A} \in \{0,1\}^{n\times n}$ is the adjacency matrix indicating connectivity among n nodes ($\boldsymbol{A}_{ij} = 1$ if nodes i and j are connected, otherwise 0); $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is the node feature matrix, where each row $\boldsymbol{X}_i$ represents the d-dimensional feature vector of node i. The regression task in graph learning is to learn a function $f : G \mapsto y$, where $y \in \mathbb{R}$ denotes the response for the graph G.

3.1. Causal Graph Learning

In causal graph learning, a graph G can be split into a **causal subgraph** C and a **confounding subgraph** S. This process is non-trivial and our proposed paradigm will hinge on the output of this process. We follow the definition in Sui et al.

(2022) and first introduce the construction of the causal subgraph C:

$$C := (\boldsymbol{M}_{\text{edge}} \odot \boldsymbol{A}, \boldsymbol{M}_{\text{node}} \cdot \mathbf{X}), \tag{1}$$

where the mask matrix $\boldsymbol{M}_{\text{edge}} \in [0,1]^{n \times n}$ and the diagonal matrix $\boldsymbol{M}_{\text{node}}$ (whose diagonal elements are in $[0,1]$) will filter out the non-causal nodes and edges. The confounding subgraph is then the "complement": $S := G - C$.

In our framework, these masks $\boldsymbol{M}_{\text{edge}}$ and $\boldsymbol{M}_{\text{node}}$ are not predefined. Instead, they are learnable soft masks, generated by MLPs conditioned on the representations of G. The parameters of these MLPs are optimized end-to-end as part of the overall model training, enabling the model to autonomously learn how to construct C and S. Further architectural details are provided in Appendix B and illustrated in Figure 2.

Notably, mutual information plays an essential role in CGL, and we introduce its calculation, exemplified by the mutual information between the hidden embedding vectors (learned by a graph neural network) of the causal subgraphs and the original graphs, as follows:

$$I(C;G) := \mathbb{E}_{C,G}\left[\log p(C \mid G)/p(C)\right], \tag{2}$$

where we follow the convention in CGL literature and abuse the notation C, G to represent a random variable following the **underlying distribution of embedding pairs** $H_{g,i}$'s and $H_{c,i}$'s. In particular, those hidden embeddings are assumed Gaussian and the joint distribution can thus be well-estimated by sample embedding pairs. We refer readers interested to Miao et al. (2022, Appendix A) for more details. Moreover, the computation/approximation of the mutual information terms is a crucial component in causal graph learning, while still under-explored for CGR; we will dissect the computation of our proposed terms in Section 4.2 through deriving the variational bounds.

3.2. Graph Information Bottleneck

The information bottleneck (Tishby et al., 2000; Tishby & Zaslavsky, 2015, IB) principle aims to balance the trade-off between preserving the information necessary for prediction and discarding irrelevant redundancy. Specifically, IB suggests to maximize $I(Z;Y)$ while minimizing $I(Z;X)$ for regular data compression, where Z is the compressed representation, X is the input, and Y is the response.

Graph information bottleneck (GIB) (Wu et al., 2020) extends the IB principle to graph-structured data, facilitating the identification of subgraphs that are most relevant for predicting graph-level responses. By minimizing the mutual information $I(C;G)$ between the extracted causal subgraph C and the original graph G, GIB reduces redundant information. However, GIB alone does not guarantee the extraction of a purely causal subgraph, as isolating causal effects requires additional interventions (Miao et al., 2022; Chen et al., 2022).

Formally, the GIB objective is expressed as:

$$-I(C;Y) + \alpha I(C;G), \tag{3}$$

where $I(C;Y)$ quantifies the predictive information retained by C (and thus needs to maximize). $I(C;G)$ serves as a regularizer to exclude irrelevant details from the original graph; the parameter α controls the trade-off between information preservation and compression.

3.3. Causal Intervention in GNNs

We borrow the structural causal model (SCM) diagram in Figure 1 to illustrate the causal intervention techniques. As shown in Figure 1, the graph G decides both the causal subgraph C and the confounding subgraph S, and the former C affects the prediction of response Y. In more detail,

- $C \leftarrow G \rightarrow S$: Graph data G encodes both C, which directly impacts Y, and S, which introduces spurious correlations.
- $S \rightarrow C \rightarrow Y$: The causal feature C has the potential to predict Y not only directly but also indirectly through its influence along this backdoor path $S \rightarrow C \rightarrow Y$.

In causal inference, confounder S incurs spurious correlations, preventing the discovery of underlying causality. To address this issue, backdoor adjustment methods focus on the interventional effect $P(Y|\text{do}(C))$, and suggest to estimate it by stratifying over S and calculating the conditional distribution $P(Y|C,S)$ (Pearl, 2014; Sui et al., 2024).

4. Revisiting Confounding Effects for CGR

In this section, we present a causal graph regression paradigm that integrates an enhanced graph information bottleneck (GIB) objective with causal discovery, reshaping the processing of confounding effects in CGL.

4.1. Overview

We first provide an overview of how graph inputs are turned into regression outputs. As shown in Figure 2, we follow the framework of Sui et al. (2024) and first encode graph embeddings $H_{g,i}$'s using a GNN-based encoder. Attention modules are then adopted to generate soft masks for extracting causal and confounding subgraphs (c.f. Equation (1)). These subgraphs are processed through two GNN modules ($\mathcal{G}_c$ and $\mathcal{G}_s$) with shared parameters to extract causal ($H_{c,i}$'s) and confounding ($H_{s,i}$'s) representations, which are passed through distinct readout layers for regression.

The optimization features an enhanced graph information bottleneck (GIB) loss L_{GIB}, comprising the causal part L_c and the confounding part L_s, to disentangle causal signals

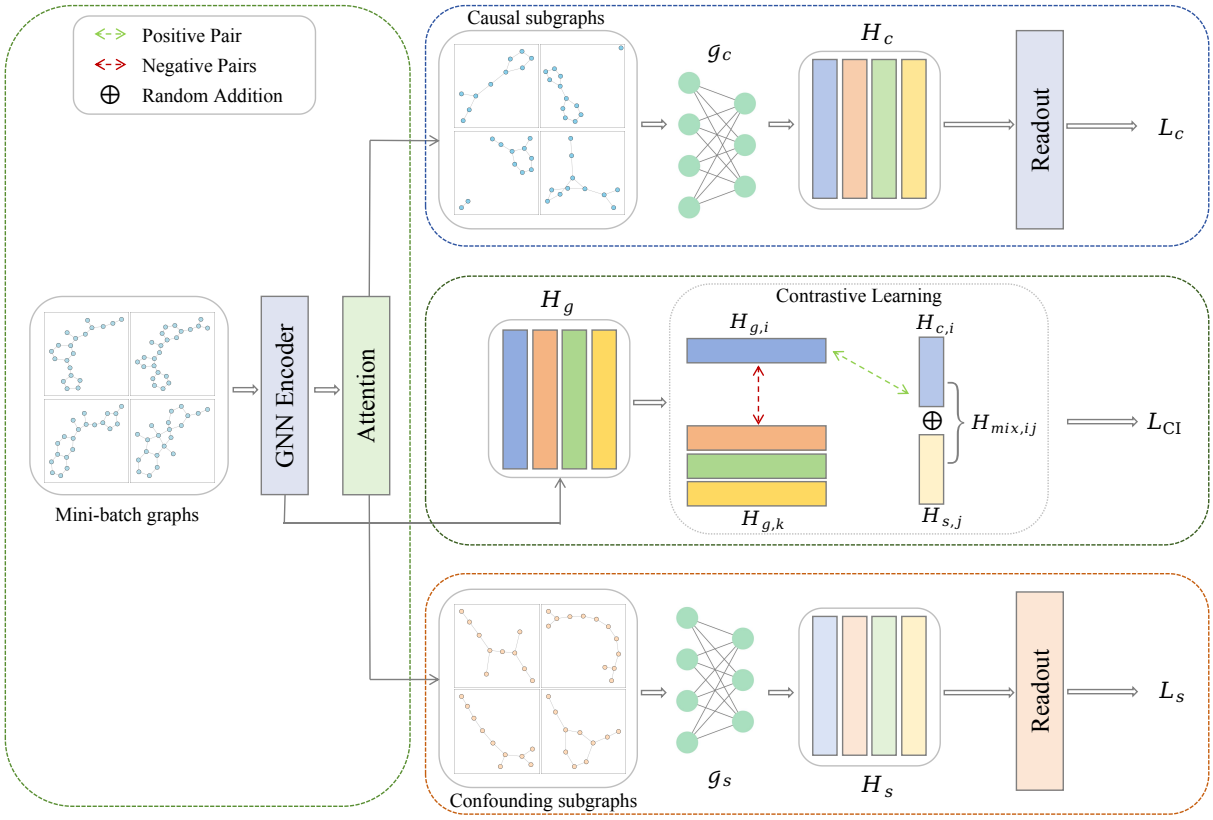


Figure 2. Given a mini-batch of graphs, (1) the GNN encoder computes the graph embeddings H_g, and an attention layer generates soft masks to extract causal and confounding subgraphs. (2) GNN $\mathcal{G}_c$ processes the causal subgraph C, generates its representation H_c, and employs readout to predict responses; it is optimized with causal subgraph loss L_c. (3) GNN $\mathcal{G}_s$, sharing parameters with $\mathcal{G}_c$, processes the confounding subgraph S, generates H_s, and applies readout for prediction; it is optimized with confounding subgraph loss L_s. (4) For causal intervention, contrastive learning guides the process. Given a graph $H_{g,i}$, the positive sample is a mixed graph $H_{\mathrm{mix},ij}$ from random addition, while any other graph $H_{g,k}$ serves as the negative sample. The causal intervention loss L_{CI} is used accordingly.

(c.f. Section 4.2). Also, counterfactual samples ($H_{\mathrm{mix},ij}$) are generated by randomly injecting confounding representations into causal ones; unsupervised learning is then performed, guided by contrastive-learning-based causal intervention loss L_{CI} (c.f. Section 4.3). More implementation details of the overall framework are deferred to Appendix B.

4.2. Enhanced GIB Objective

CGL adopts the GIB objective to extract subgraphs that retain essential predictive information while excluding redundant components (Zhang et al., 2023), which aligns with the disentanglement of causal subgraph C and confounding subgraph S in CGL. Original GIB assumes the confounding subgraph S is pure noise and cannot predict the response Y (Chen et al., 2022), while as we discussed in Section 1 S may still contain information that is predictive of the response Y. In its current form, the GIB framework overlooks this aspect, causing the model to allocate all Y-relevant information to C and to potentially lose meaningful content.

This limitation leads to incomplete causal disentanglement, which impacts the generalization of models to out-of-distribution (OOD) settings. To overcome this issue, we propose an enhanced GIB loss function that takes the predictive roles of both C and S into consideration. By introducing mutual information terms on S during optimization, we avoid overburdening C with all relevant information, and consequently enable a more precise disentanglement.

Overall, our enhanced GIB objective is defined as follows:

$$-I(C;Y) + \alpha I(C;G) - \beta I(S;Y), \tag{4}$$

which formally extends the original GIB objective by introducing a confounder-related term $I(S;Y)$ to capture the

predictive capacity of S, along with a parameter β. In particular, we intentionally exclude the $I(S;G)$ term because, in the SCM diagram of Figure 1, S primarily introduces shortcut rather than directly encoding causality; overly imposing structural regularization on S can disrupt disentanglement and lead to suboptimal separation between C and S. Notably, the conceptual objective (4) is incomputable in practice. We devote the remainder of this subsection to the practical computation of Equation (4) for CGR.

Variational bounds for approximating $I(C;G)$. The mutual information $I(C;G)$ is mathematically defined based on the marginal distributin $p(C) = \sum_G p(C|G)p(G)$. Since $p(C)$ is intractable, a variational distribution $q(C)$ is introduced and induces an upper bound:

$$I(C;G) \leq \mathbb{E}_{p(G)}\big[\mathrm{KL}\big(p(C \mid G) \| q(C)\big)\big]. \tag{5}$$

To efficiently compute the KL divergence in Equation (5), we follow the literature (Chechik et al., 2003; Kingma et al., 2013) and assume that $p(C \mid G)$ and $q(C)$ are multivariate Gaussian distributions:

$$p(C \mid G) = \mathcal{N}(\mu_\phi(G), \Sigma_\phi(G)), \quad q(C) = \mathcal{N}(0, I), \tag{6}$$

where $\mu_\phi(G)$ and $\Sigma_\phi(G)$ are the mean vector and covariance matrix estimated by GNNs. To simplify computation and stabilize training, we further assume $\Sigma_\phi(G)$ is an identity matrix, removing the need to learn covariance parameters. This simplification is not only practical but also theoretically justified, as any full-rank covariance can be whitened without loss of generality (Chechik et al., 2003, Appendix A). $\mathrm{KL}\big(p(C \mid G) \| q(C)\big)$ then reduces to:

$$\begin{aligned} &\frac{1}{2}\Big[\mathrm{tr}(\Sigma_\phi(G)) + \|\mu_\phi(G)\|^2 - d - \log\det\Sigma_\phi(G)\Big] \\ =&\frac{1}{2}\|\mu_\phi(G)\|^2. \end{aligned} \tag{7}$$

where d is the dimensionality of C. Further substituting Equation (7) into Equation (5), we obtain an upper bound for $I(C;G)$:

$$I(C;G) \leq \frac{1}{2}\mathbb{E}_{p(G)}\big[\|\mu_\phi(G)\|^2\big], \tag{8}$$

which serves as an easy-to-compute proxy for $I(C;G)$.

Variational bounds for approximating $I(C;Y), I(S;Y)$. We first recall $I(C;Y)$ mathematically reads:

$$I(C;Y) = H(Y) - H(Y \mid C), \tag{9}$$

where $H(Y)$ denotes the entropy of Y, representing the overall uncertainty in the target variable. Since $H(Y)$ remains constant, maximizing $I(C;Y)$ reduces to minimizing the conditional entropy $H(Y \mid C)$, given by:

$$H(Y \mid C) = -\mathbb{E}_{C,Y}[\log p(Y \mid C)]. \tag{10}$$

The computation of $H(Y \mid C)$ is supposed to hinge on the hidden embeddings $H_{c,i}$'s produced by a GNN $\mathcal{G}_c$ (see Section 4.1); we model the conditional distribution $p(Y \mid H_c)$ as a Gaussian distribution:

$$p(Y \mid H_c) = \mathcal{N}(Y; \mu_{(c)}, \sigma^2_{(c)}), \tag{11}$$

where $\mu_{(c)}$ and $\sigma^2_{(c)}$ represent the scalar conditional mean and variance of Y (estimated by networks) given a causal subgraph representation H_c. The probability density function for this Gaussian is:

$$p(Y \mid H_c) = \frac{1}{\sqrt{2\pi\sigma^2_{(c)}}} \exp\left(-\frac{(Y-\mu_{(c)})^2}{2\sigma^2_{(c)}}\right). \tag{12}$$

Substituting Equation (12) into Equation (10), we can further approximate $H(Y \mid C)$ through empirical data:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{(Y_i-\mu_{(c),i})^2}{2\sigma^2_{(c),i}} + \frac{1}{2}\log(2\pi\sigma^2_{(c),i})\right], \tag{13}$$

where N represents sample size, Y_i is the target response for the i-th sample. and $\mu_{(c),i}$ and $\sigma^2_{(c),i}$ are the corresponding mean and variance of Y given $H_{c,i}$.

If a constant conditional variance (i.e., $\sigma^2_{(c)} = 1$) is assumed, a choice adopted for stability and aligning with approaches in (Nix & Weigend, 1994; Yu et al., 2024), then $I(C;Y)$ (or, equivalently, $-H(Y \mid C)$) reduces to the least squares loss:

$$\begin{aligned} &-\frac{1}{N}\sum_{i=1}^{N}\left[\frac{(Y_i-\mu_{(c),i})^2}{2\sigma^2_{(c),i}} + \frac{1}{2}\log(2\pi\sigma^2_{(c),i})\right] \\ \propto &-\frac{1}{N}\sum_{i=1}^{N}(Y_i-\mu_{(c),i})^2, \end{aligned} \tag{14}$$

which turns to the **causal subgraph objective** L_{CP}.

Similarly, the mutual information $I(S;Y)$ can induce the **confounding subgraph objective**

$$L_{\text{SP}} \propto -\frac{1}{N}\sum_{i=1}^{N}(Y_i-\mu_{(s),i})^2. \tag{15}$$

Empirically, we employ two independent readout layers to compute the causal and confounding subgraph mean $\mu_{(c),i}$'s and $\mu_{(s),i}$'s.

In summary, our enhanced GIB objective can be decomposed into two distinct loss components: the causal subgraph loss $L_c(G,C,Y) = -I(C;Y) + \alpha I(C;G)$ and the confounding subgraph loss $L_s(S,Y) = -I(S;Y)$. The complete enhanced GIB objective we propose is:

$$\begin{aligned} L_{\text{GIB}} &= L_c + \beta L_s \\ &= -I(C;Y) + \alpha I(C;G) - \beta I(S;Y), \end{aligned}$$

and in practice we use $-L_{\text{CP}} + \alpha\mathbb{E}_{p(G)}\big[\|\mu_\phi(G)\|^2\big] - \beta L_{\text{SP}}$.

4.3. Causal Intervention

To further strengthen causal learning in CGR, we introduce a causal intervention loss and reshape the processing of confounding effects therein. In general, our approach injects randomness at the graph level by randomly pairing confounding subgraphs with target causal subgraphs from the entire dataset. By generating counterfactual graph representations through the random combination of these subgraphs, we effectively implement causal intervention.

This strategy can be understood as an implicit realization of backdoor adjustment (Pearl, 2014) in the representation space. In existing research on graph classification tasks (Fan et al., 2022; Sui et al., 2024), causal intervention is typically modeled by predicting $P(Y|C,S)$ through intervened graphs, adjusting for causal effects by comparing predictive distributions under different confounding conditions. However, in regression tasks, Y is a continuous variable, and directly modeling $P(Y|C,S)$ becomes significantly more challenging. To overcome this, we follow the spirit of contrastive learning to get rid of the reliance on explicit labels.

In more detail, following Sui et al. (2022), we use a random addition method to pair the confounding subgraph with the target causal subgraph , which gives H_{mix}:

$$H_{\text{mix},ij} = H_{c,i} + H_{s,j}. \tag{16}$$

Comparing the predictions of H_{mix} with the original graph's labels, as shown in Sui et al. (2022), can inadvertently force the mixed graph to discard all confounding effects, thereby nullifying the intended causal disentanglement.

To mitigate this issue, we suggest learning causal representations through contrastive learning. Specifically, the causal subgraph, when combined with different confounding subgraphs, consistently produces mixed graph representations that are aligned with the original graph representation. This formulation enables the model to learn causal subgraphs that are invariant across varying confounders, and to avoid the causal subgraphs boiled down to non-informative ones.

To achieve this, we propose a causal intervention loss guided by contrastive learning. Specifically, the method aligns the representation of the original graph with that of its corresponding random mixture graph, while simultaneously ensuring that representations of unrelated graphs remain distinct. In implementation, draw inspiration from the InfoNCE loss (Oord et al., 2018), we treat H_g and H_{mix} from the same causal subgraph as positive pairs, and H_g with representations of other graphs within the batch as negative pairs. Formally, the mixed graph contrastive loss is defined as:

$$L_{\text{CI}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(H_{g,i}, H_{\text{mix},ij})}{\sum_{k=1,k\neq i}^{B}\exp(\text{sim}(H_{g,i}, H_{g,k})}, \tag{17}$$

where B is the batch size, $H_{\text{mix},ij}$ is the representation of the mixed graph combining the i-th causal subgraph and the j-th confounding subgraph, and $H_{g,i}$ is the representation of the original graph.

Remark 4.1. The ultimate loss used in our paradigm is a simple combination of the GIB objective and the causal intervention loss: $L = L_{\text{GIB}} + \lambda L_{\text{CI}}$.

5. Experiments

In this section, we evaluate the prediction performance and OOD generalization ability of our method. We comprehensively compare our method with existing models to demonstrate the superior generalization ability of our method on regression tasks. We briefly introduce the dataset, baselines, and experimental settings here.

5.1. Datsets

GOOD-ZINC. GOOD-ZINC is a regression task in the GOOD benchmark (Gui et al., 2022), which aims to test the out-of-distribution performance of real-world molecular property regression datasets from the ZINC database (Gómez-Bombarelli et al., 2018). The input is a molecular graph containing up to 38 heavy atoms, and the task is to predict the restricted solubility of the molecule (Jin et al., 2018; Kusner et al., 2017). GOOD-ZINC includes four specific OOD types: Scaffold-Covariate, Scaffold-Concept, Size-Covariate, and Size-Concept. Scaffold OOD involves changes in molecular structures, while Size OOD varies graph size. Each can manifest as Covariate Shift ($P(X)$ changes, $P(Y|X)$ remains stable) or Concept Shift (spurious correlations in training break in testing).

ReactionOOD-SOOD. In addition to the GOOD benchmark, we also used three S-OOD datasets in the ReactionOOD benchmark (Wang et al., 2023), namely Cycloaddition (Stuyver et al., 2023), E2&S_N2 (von Rudorff et al., 2020), and RDB7 (Spiekermann et al., 2022), which are designed to extract information outside the structural distribution during molecular reactions. Cycloaddition and RDB7 have two domains: Total Atom Number (where the total number of atoms in a reaction exceeds the training range) and First Reactant Scaffold (where the first reactant has a new molecular scaffold unseen in training), while E2&S_N2 dataset contains reactions with molecules whose scaffold cannot be properly defined, which prevents the scaffold from being an applicable domain index for this dataset. The definitions of two shifts Covariate and Concept in ReactionOOD are consistent with those in GOOD.

5.2. Baselines and Setup

As our framework is general and aims to address distribution shifts, we compare it against several baseline methods.

Table 1. OOD generalization performance on GOOD-ZINC dataset, with **boldface** being the best and underline being the runner-up.

GOOD-ZINC	SCAFFOLD				SIZE			
	COVARIATE		CONCEPT		COVARIATE		CONCEPT	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD
ERM	0.1188±0.0030	0.1660±0.0093	0.1174±0.0013	0.1248±0.0018	0.1222±0.0061	0.2331±0.0169	0.1304±0.0010	0.1406±0.0002
IRM	0.1258±0.0033	0.2313±0.0243	0.1176±0.0052	0.1245±0.0062	0.1217±0.0014	0.5840±0.0039	0.1331±0.0045	0.1338±0.0011
VREX	0.0978±0.0016	0.1561±0.0021	0.1928±0.0021	0.1271±0.0020	0.1841±0.0009	0.2276±0.0005	0.1206±0.0008	0.1289±0.0039
MIXUP	0.1348±0.0025	0.2157±0.0098	0.1192±0.0026	0.1296±0.0049	0.1431±0.0070	0.2573±0.0042	0.1625±0.0121	0.1660±0.0063
DANN	0.1152±0.0021	0.1734±0.0005	0.1284±0.0031	0.1289±0.0020	0.1053±0.0081	0.2254±0.0140	0.1227±0.0008	0.1271±0.0039
CORAL	0.1252±0.0043	0.1734±0.0034	0.1173±0.0029	0.1260±0.0024	0.1164±0.0004	0.2243±0.0147	0.1246±0.0062	0.1270±0.0020
CIGA	0.1568±0.0034	0.2986±0.0041	0.1926±0.0120	0.2415±0.0115	0.1500±0.0001	0.6102±0.0148	0.3560±0.0160	0.3240±0.0451
DIR	0.2483±0.0056	0.3650±0.0032	0.2510±0.0001	0.2619±0.0076	0.2515±0.0529	0.4224±0.0679	0.4831±0.0823	0.3630±0.0872
GSAT	0.0890±0.0031	0.1419±0.0043	0.0928±0.0029	0.0999±0.0029	0.0876±0.0032	0.2112±0.0033	0.1002±0.0013	0.1043±0.0001
OURS	**0.0514±0.0061**	**0.1046±0.0007**	**0.0659±0.0041**	**0.0518±0.0007**	**0.0466±0.0034**	**0.1484±0.0033**	**0.0577±0.0008**	**0.0580±0.0004**

Empirical Risk Minimization (ERM) (Vapnik, 1991) serves as a non-OOD baseline for comparison with OOD methods. We consider both Euclidean and graph-based state-of-the-art OOD approaches: (1) Euclidean OOD methods include IRM (Arjovsky et al., 2019), VREx (Krueger et al., 2021), GroupDRO (Sagawa et al., 2019), DANN (Ganin et al., 2016), Coral (Sun & Saenko, 2016), and Mixup (Zhang, 2017); (2) Graph OOD methods include CIGA (Chen et al., 2022), GSAT (Miao et al., 2022), and DIR (Wu et al., 2022b).

For a fair comparison, all methods are implemented with consistent architectures and hyperparameters, ensuring that performance differences arise solely from the method itself. To provide reliable results, each experiment is repeated three times with different random seeds, and we report the mean and standard error of the results. Detailed settings and hyperparameter configurations are described in Appendix A.4.

5.3. Results of GOOD

As shown in Table 1, our proposed method achieves SOTA performance on GOOD-ZINC, consistently outperforming all baseline methods across both domains (Scaffold and Size) and under different distribution shifts (Covariate and Concept). Specifically, in terms of Mean Absolute Error (MAE), our method demonstrates significant improvements in both in-distribution (ID) and out-of-distribution (OOD) settings.

For instance, in the Scaffold domain under the Covariate shift, our method achieves an MAE of 0.0514±0.0061 (ID) and 0.1046±0.0007 (OOD), outperforming GSAT, the next-best method, by 42.2% in ID and 26.3% in OOD performance. Similarly, under the Concept shift, our method achieves 0.0659±0.0041 (ID) and 0.0518±0.0007 (OOD), representing improvements of 29.0% and 48.1%, respectively, over GSAT.

In the Size domain, our method also achieves remarkable results. Under the Covariate shift, it achieves an MAE of 0.0466±0.0034 (ID) and 0.1484±0.0033 (OOD), which translate to 46.8% lower ID error and 29.7% lower OOD error compared to GSAT. Similarly, under the Concept shift, our approach yields an MAE of 0.0577±0.0008 (ID) and 0.0580±0.0004 (OOD), improving upon GSAT by 42.4% and 44.4%, respectively.

In addition to achieving lower MAE values, our method exhibits significantly reduced variances compared to other approaches, highlighting its stability under diverse conditions. These findings confirm the strong generalization capability of our method across different domains and types of distributional shifts.

5.4. Results of ReactionOOD

Table 2 and Table 3 highlight the robust generalization ability of our method across multiple datasets and evaluation settings, as measured by RMSE. Our method achieves the best OOD performance in 6 out of 10 cases and ranks second in 2 cases. Notably, in cases where another method outperforms ours, the performance gap is within a small margin.

For instance, in the Cycloaddition dataset, under the total atom number domain with a concept shift, Our method achieves an OOD RMSE of 5.53 ± 0.12, outperforming all baseline methods. While some non-causal baselines (e.g., Coral in this specific setting, achieving an ID RMSE of 4.10 ± 0.05 versus our 4.41 ± 0.22) might get better ID performance by exploiting spurious but predictive features, such approaches can become less reliable under OOD conditions (e.g., Coral's OOD RMSE degrades to 5.74 ± 0.04). In contrast, our method's focus on identifying and removing these spurious features contributes to its stable and superior OOD performance. Even in other Cycloaddition cases where ours ranks second, such as the same domain with a covariate shift, the OOD RMSE (4.42 ± 0.24) is only 0.06 away from the best-performing method (4.36 ± 0.15).

In RDB7, a smaller dataset within the ReactionOOD where causal inference can be more difficult, our method achieves the lowest OOD RMSE (15.73 ± 0.37) under the concept shift. Our method's principled focus on true causal features, which leads to better OOD generalization ability and stabil-

Table 2. OOD generalization performance on Cycloaddition and RDB7 dataset.

Dataset	Methods	First Reactant Scaffold				Total Atom Number			
		Covariate		Concept		Covariate		Concept	
		ID	OOD	ID	OOD	ID	OOD	ID	OOD
Cycloaddition	ERM	4.38±0.04	4.80±0.38	4.79±0.03	5.60±0.02	**3.77±0.01**	**4.36±0.15**	4.22±0.04	5.69±0.03
	IRM	15.30±0.05	21.16±0.01	17.55±0.03	18.64±0.25	17.53±0.17	17.44±0.14	23.14±0.02	22.56±0.01
	VREx	5.54±0.02	6.69±0.48	5.02±0.05	6.14±0.09	4.79±0.03	5.22±0.06	4.92±0.14	6.39±0.04
	Mixup	4.51±0.04	5.24±0.83	4.90±0.01	5.90±0.05	3.90±0.13	4.53±0.03	4.11±0.09	5.93±0.13
	DANN	4.42±0.03	4.68±0.12	4.81±0.01	5.75±0.06	3.87±0.05	4.65±0.10	4.18±0.02	5.68±0.10
	Coral	**4.36±0.07**	4.95±0.30	4.82±0.03	5.72±0.16	4.39±0.59	5.05±0.48	**4.10±0.05**	5.74±0.04
	CIGA	5.26±0.04	5.67±0.04	5.30±0.29	5.64±0.03	4.93±0.05	6.62±1.09	5.03±0.09	6.21±0.06
	DIR	4.94±0.02	5.31±0.79	5.85±0.20	6.30±0.38	5.52±0.03	6.86±0.05	5.21±0.12	7.09±0.03
	GSAT	4.42±0.05	4.63±0.05	4.87±0.01	5.69±0.01	3.81±0.01	4.56±0.01	4.12±0.04	5.64±0.11
	Ours	4.57±0.13	**4.22±0.09**	**4.53±0.04**	**5.37±0.05**	4.06±0.01	4.42±0.24	4.41±0.22	**5.53±0.12**
RDB7	ERM	10.28±0.05	22.95±0.90	11.38±0.08	**14.81±0.05**	10.86±0.01	7.66±0.55	**11.28±0.15**	15.79±0.24
	IRM	59.87±0.02	76.51±0.46	65.72±0.13	63.03±0.13	63.55±0.02	69.06±0.37	81.14±0.02	46.84±0.42
	VREx	16.62±0.18	**21.89±0.02**	14.62±0.04	18.28±0.09	14.60±0.01	13.84±0.07	34.66±1.56	32.59±3.28
	Mixup	10.76±0.07	23.49±0.09	11.89±0.05	15.64±0.10	11.13±0.02	10.78±0.17	11.66±0.04	17.21±0.28
	DANN	10.28±0.05	23.54±0.07	11.28±0.01	14.93±0.05	10.77±0.22	8.29±0.10	11.34±0.05	16.28±0.15
	Coral	10.30±0.12	22.19±0.63	**11.12±0.03**	14.81±0.06	10.61±0.01	8.04±0.14	11.33±0.08	16.13±0.08
	CIGA	14.97±0.75	30.08±0.84	18.68±1.94	21.35±1.34	16.48±0.69	19.12±1.85	20.58±1.54	18.53±1.30
	DIR	14.34±0.68	26.99±0.49	17.13±1.76	20.18±1.86	14.03±2.06	15.01±0.98	13.52±0.51	16.60±1.09
	GSAT	10.52±0.04	23.45±0.11	11.26±0.25	14.85±0.12	10.80±0.01	8.66±0.10	11.58±0.03	16.08±0.41
	Ours	**10.12±0.08**	23.11±0.46	11.26±0.02	14.94±0.25	**10.51±0.08**	**6.84±0.32**	11.46±0.06	**15.73±0.37**

Table 3. OOD generalization performance on E2&S_N2 dataset.

Methods	Covariate		Concept	
	ID	OOD	ID	OOD
ERM	4.45±0.04	5.47±0.27	4.87±0.02	5.04±0.02
IRM	11.61±0.18	21.54±1.07	20.95±0.02	17.57±0.03
VREx	4.58±0.02	5.48±0.13	10.75±1.54	8.77±2.31
Mixup	4.55±0.09	5.55±0.01	4.69±0.08	5.11±0.01
DANN	4.51±0.06	5.38±0.04	**4.48±0.10**	5.04±0.02
Coral	4.44±0.11	5.68±0.20	4.54±0.02	**4.97±0.07**
CIGA	5.05±0.35	6.57±0.52	4.65±0.26	5.39±0.47
DIR	5.61±0.26	6.59±0.31	6.56±0.34	6.29±0.11
GSAT	4.55±0.01	5.69±0.05	4.55±0.09	5.04±0.03
Ours	**4.40±0.03**	**4.83±0.10**	4.53±0.12	5.03±0.09

ity. Even though causal methods generally face challenges in smaller datasets (Guo et al., 2020), our approach consistently outperforms other listed causal intervention baselines such as CIGA in all RDB7 settings. In the E2&S_N2 dataset, our method delivers the best OOD RMSE (4.83 ± 0.10) under the covariate shift and achieves highly competitive results under the concept shift (5.03 ± 0.09).

As noted in OOD-GNN (Tajwar et al., 2021), no method consistently performs best on every dataset due to varying distribution shifts and inductive biases. Our approach, designed under more general and weaker assumptions which do not assume that spurious features are non-predictive, aims to tackle a wider range of real-world distribution shifts.

5.5. Effectiveness of OURS in Classification Task

To validate the generality and effectiveness of our proposed losses, L_{GIB} and L_{CI}, we conduct ablation studies on the GOOD-Motif dataset under the size domain setting. The results, evaluated in terms of accuracy, are reported on the OOD dataset, as shown in Figure 3. The ablation study on L_{GIB} aims to examine our hypothesis that confounders possess certain predictive power; thus, this experiment excludes the causal intervention loss L_{CI} Conversely, the ablation study on L_{CI} evaluates whether the contrastive learning-driven causal intervention loss can independently achieve strong OOD performance. Therefore, in this experiment, we do not incorporate the predictive power of confounding factors.

Predictive power of confounding subgraphs. The left panel compares minimizing confounding subgraph prediction alone versus introducing constraints to model their predictive ability. The results show that ignore the predictive role of confounding subgraphs leads to incomplete disentanglement and weaker OOD generalization, demonstrating that accounting for their influence is crucial.

Effectiveness of contrastive learning. The right panel compares using predictions from randomly generated counterfactual graphs as causal intervention loss versus our proposed contrastive learning loss. The results show that our contrastive learning approach, initially validated in regression tasks, is equally effective in classification tasks, highlighting its general applicability.

These studies confirm the importance of explicitly modeling confounding subgraphs and the robustness of our contrastive learning loss for OOD generalization. More experimental results are provided in the Appendix A.5.

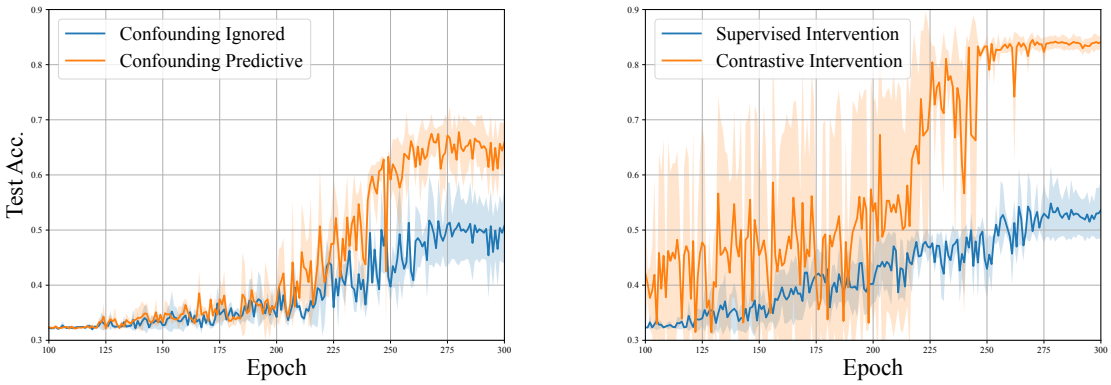


Figure 3. Ablation study on confounder predictive power (left) and causal intervention methods (right) for OOD generalization on GOOD-Motif.

6. Conclusion

In this work, we propose a recipe for causal graph regression through reshaping the processing of confounding effects in existing CGL classification-specific techniques. In particular, we develop an enhanced graph information bottleneck (GIB) loss function which highlights the impact of confounding effects and consequently benefits the recognition of causal subgraphs. Moreover, we revisit the causal intervention technique, which randomly combines causal subgraphs and confounder from the same class (label) to eliminate confounding effects. Adapting this technique to regression requires removal of label information; to this end, we analyze the principle of causal intervention and propose to connect it with unsupervised contrastive learning loss. Experimental results on graph OOD benchmarks demonstrate the effectiveness of our proposed techniques in improving the generalizability of graph regression models.

Acknowledgements

We sincerely thank the Area Chair and the anonymous reviewers for their valuable feedback and constructive suggestions, which helped improve this work. The authors acknowledge funding from Research Grants Council (RGC) under grant `22303424` and GuangDong Basic and Applied Basic Research Foundation under grant `2025A1515010259`.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

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

Chechik, G., Globerson, A., Tishby, N., and Weiss, Y. Information bottleneck for gaussian variables. *Advances in Neural Information Processing Systems*, 16, 2003.

Chen, Y., Zhang, Y., Bian, Y., Yang, H., Kaili, M., Xie, B., Liu, T., Han, B., and Cheng, J. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35:22131–22148, 2022.

Chen, Y., Bian, Y., Zhou, K., Xie, B., Han, B., and Cheng, J. Does invariant graph learning via environment augmentation learn invariance? *Advances in Neural Information Processing Systems*, 36, 2024.

Fan, S., Wang, X., Mo, Y., Shi, C., and Tang, J. Debiasing graph neural networks via learning disentangled causal substructure. *Advances in Neural Information Processing Systems*, 35:24934–24946, 2022.

Fan, S., Wang, X., Shi, C., Cui, P., and Wang, B. Generalizing graph neural networks on out-of-distribution graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Gao, H., Li, J., Qiang, W., Si, L., Xu, B., Zheng, C., and Sun, F. Robust causal graph representation learning against confounding effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7624–7632, 2023.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Gui, S., Li, X., Wang, L., and Ji, S. Good: A graph out-of-distribution benchmark. *Advances in Neural Information Processing Systems*, 35:2059–2073, 2022.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

Guo, Z., Wu, Z., Xiao, T., Aggarwal, C., Liu, H., and Wang, S. Counterfactual learning on graphs: A survey. *Machine Intelligence Research*, 22(1):17–59, 2025.

Jin, W., Barzilay, R., and Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pp. 2323–2332. PMLR, 2018.

Kingma, D. P., Welling, M., et al. Auto-encoding variational bayes, 2013.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. Grammar variational autoencoder. In *International conference on machine learning*, pp. 1945–1954. PMLR, 2017.

Li, G., Knoop, V. L., and Van Lint, H. Multistep traffic forecasting by dynamic graph convolution: Interpretations of real-time spatial correlations. *Transportation Research Part C: Emerging Technologies*, 128:103185, 2021.

Li, H., Wang, X., Zhang, Z., and Zhu, W. Ood-gnn: Out-of-distribution generalized graph neural network. *IEEE Transactions on Knowledge and Data Engineering*, 35 (7):7328–7340, 2022.

Lin, W., Lan, H., and Li, B. Generative causal explanations for graph neural networks. In *International Conference on Machine Learning*, pp. 6666–6679. PMLR, 2021.

Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., and Zhang, X. Parameterized explainer for graph neural network. *Advances in neural information processing systems*, 33:19620–19631, 2020.

Ma, F., Li, H., and Ilyas, M. Utilizing reinforcement learning and causal graph networks to address the intricate dynamics in financial risk prediction. *International Journal of Information Technologies and Systems Approach (IJITSA)*, 17(1):1–19, 2024.

Ma, J. A survey of out-of-distribution generalization for graph machine learning from a causal view. *arXiv preprint arXiv:2409.09858*, 2024.

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

Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., and Blundell, C. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

Nix, D. and Weigend, A. Learning local error bars for nonlinear regression. *Advances in neural information processing systems*, 7, 1994.

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

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

Pleiss, G., Souza, A., Kim, J., Li, B., and Weinberger, K. Q. Neural network out-of-distribution detection for regression tasks. 2019.

Qiao, G., Wang, G., and Li, Y. Causal enhanced drug–target interaction prediction based on graph generation and multi-source information fusion. *Bioinformatics*, 40 (10):btae570, 2024.

Rollins, Z. A., Cheng, A. C., and Metwally, E. Molprop: Molecular property prediction with multimodal language and graph fusion. *Journal of Cheminformatics*, 16(1):56, 2024.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Spiekermann, K., Pattanaik, L., and Green, W. H. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. *Scientific Data*, 9(1):417, 2022.

Stuyver, T., Jorner, K., and Coley, C. W. Reaction profiles for quantum chemistry-computed [3+ 2] cycloaddition reactions. *Scientific Data*, 10(1):66, 2023.

Sui, Y., Wang, X., Wu, J., Lin, M., He, X., and 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.

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

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016.

Tajwar, F., Kumar, A., Xie, S. M., and Liang, P. No true state-of-the-art? ood detection methods are inconsistent across datasets. *arXiv preprint arXiv:2109.05554*, 2021.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

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

Vapnik, V. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

von Rudorff, G. F., Heinen, S. N., Bragato, M., and von Lilienfeld, O. A. Thousands of reactants and transition states for competing e2 and s2 reactions. *Machine Learning: Science and Technology*, 1(4):045026, 2020.

Wang, Z. and Veitch, V. A unified causal view of domain invariant representation learning. 2022.

Wang, Z., Chen, Y., Duan, Y., Li, W., Han, B., Cheng, J., and Tong, H. Towards out-of-distribution generalizable predictions of chemical kinetics properties. *arXiv preprint arXiv:2310.03152*, 2023.

Wu, Q., Zhang, H., Yan, J., and Wipf, D. Handling distribution shifts on graphs: An invariance perspective. *arXiv preprint arXiv:2202.02466*, 2022a.

Wu, T., Ren, H., Li, P., and Leskovec, J. Graph information bottleneck. *Advances in Neural Information Processing Systems*, 33:20437–20448, 2020.

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

Wu, Y.-X., Wang, X., Zhang, A., Hu, X., Feng, F., He, X., and Chua, T.-S. Deconfounding to explanation evaluation in graph neural networks. *arXiv preprint arXiv:2201.08802*, 2022c.

Yu, S., Yu, X., Løkse, S., Jenssen, R., and Príncipe, J. C. Cauchy-schwarz divergence information bottleneck for regression. In *ICLR*, 2024.

Zhang, H. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhang, J., Chen, Z., Mei, H., Luo, D., and Wei, H. Regexplainer: Generating explanations for graph neural networks in regression task. *arXiv preprint arXiv:2307.07840*, 2023.

Zhao, S., Prapas, I., Karasante, I., Xiong, Z., Papoutsis, I., Camps-Valls, G., and Zhu, X. X. Causal graph neural networks for wildfire danger prediction. *arXiv preprint arXiv:2403.08414*, 2024.

A. Supplementary Experiments

A.1. GOOD Benchmark

The Graph Out-Of-Distribution (GOOD) benchmark is the most comprehensive and authoritative benchmark for assessing the OOD generalization of graph learning models. It includes 11 datasets, covering six graph-level and five node-level tasks, with 51 dataset splits across covariate shift, concept shift, and no shift scenarios. Among them, nine datasets focus on classification (binary and multi-class), one (GOOD-ZINC) on regression, and one (GOOD-PCBA) on multi-objective binary classification. GOOD is the first benchmark to incorporate both covariate and concept shifts within the same domain, enabling controlled comparisons. It evaluates 10 state-of-the-art OOD methods, including four tailored for graphs, resulting in 510 dataset-model combinations. As a result, GOOD provides a systematic and rigorous framework for benchmarking OOD generalization in graph learning

A.2. ReactionOOD Benchmark

The ReactionOOD benchmark is a specialized out-of-distribution (OOD) evaluation framework designed to systematically assess the generalization capabilities of machine learning models in predicting the kinetic properties of chemical reactions. It introduces three distinct levels of OOD shifts—structural, conditional, and mechanistic—and comprises six datasets, all formulated as regression tasks. Structural OOD (S-OOD) examines variations in reactant structures, including shifts based on total atomic count (E2 & SN2) and reactant scaffolds (RDB7, Cycloaddition). Conditional OOD (C-OOD) investigates the effect of environmental conditions on kinetic properties, considering shifts in temperature (RMG Lib. T) and combined temperature-pressure settings (RMG Lib. TP). Mechanistic OOD (M-OOD) explores the impact of different reaction mechanisms (RMG Family) on kinetic property predictions.

A.3. GOOD-ZINC Dataset Details

Table 4 presents the number of graphs/nodes in different dataset splits for the GOOD-ZINC dataset. The dataset is analyzed under three types of distribution shifts: covariate, concept, and no shift. Each row represents the number of graphs/nodes in training, in-distribution (ID) validation, ID test, out-of-distribution (OOD) validation, and OOD test sets. The no-shift scenario serves as a baseline with no distributional difference between training and test sets.

Table 4. Details of GOOD-ZINC dataset.

Dataset	Shift	Train	ID validation	ID test	OOD validation	OOD test
GOOD-ZINC	covariate	149674	24945	24945	24945	24946
	concept	101867	21828	21828	43539	60393
	no shift	149673	49891	49891	-	-

A.4. Experimental Settings

We use the GOOD-ZINC dataset from the GOOD benchmark and the S-OOD tasks from ReactionOOD, excluding other OOD tasks from ReactionOOD as they are still under maintenance. Our baseline results on ReactionOOD have been acknowledged by the original authors. We use a three-layer GIN as the backbone model, with 300 hidden dimensions, which is consistently applied in both OURS and baseline models. The model is trained for 300 epochs, with the learning rate adjusted using the cosine annealing strategy. The initial learning rate is set to 0.001, with a minimum value of 1e-8 . For the OURS model, all tunable hyperparameters in the loss function L are set to 0.5.

A.5. Ablation Studies

Effectiveness Analysis To evaluate the effectiveness of the proposed loss functions L_{GIB} and L_{CI} in improving the model's OOD generalization ability, we conducted a series of ablation studies across four ood datasets: ZINC, Cycloaddition, E2SN2, and RDB7. Ours w/o BO serves as the baseline model, where both loss functions are removed, and only the causal subgraph readout layer's l_1 loss is used for optimization. Ours w/o GIB ablates L_{GIB}, eliminating the constraint on confounding subgraphs to assess the impact of removing confounder control on generalization. Conversely, Ours w/o CI removes L_{CI} while keeping L_{GIB}, allowing us to examine the contribution of the causal intervention loss to OOD generalization. Ours represents the complete model, incorporating both loss functions for optimization. Notably, ZINC is

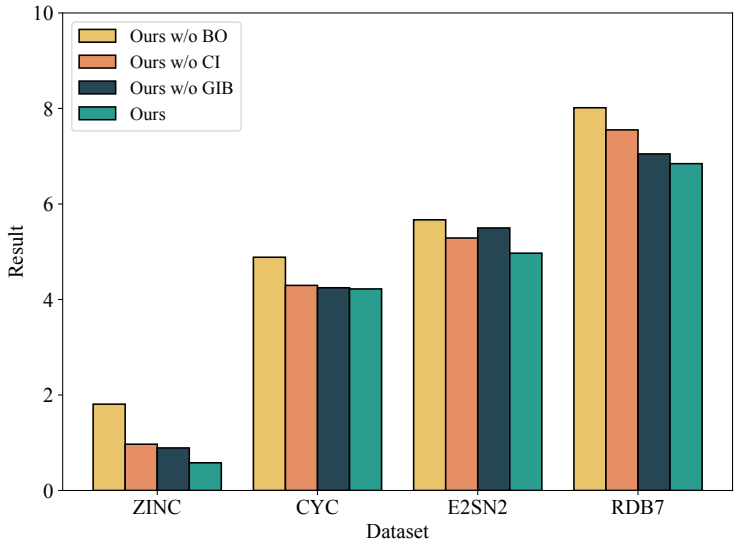


Figure 4. The comparison of different components.

evaluated using MAE, while the other datasets adopt RMSE as the evaluation metric. Given that the ZINC results are small (approximately 0.0x), we scale them by a factor of 10 in the Figure 4 for better visualization and comparison.

The results reveal several key insights. The full model (green) consistently achieves the lowest RMSE across all datasets, demonstrating the effectiveness of jointly applying both the enhanced GIB loss and the CI loss. Removing both components (yellow) leads to the worst performance, confirming that both components are essential. Between the two losses, removing CI (orange) generally causes a larger degradation than removing GIB (blue), suggesting that CI plays a more dominant role. On E2SN2, however, GIB contributes more significantly. These results indicate that GIB and CI provide complementary benefits, and that using both yields the best OOD generalization.

Parameter Sensitivity Analysis In this experiment, we analyzed the sensitivity of loss function hyperparameters under different settings in the Cycloaddition dataset, focusing on two key components of our proposed loss function: the hyperparameter λ for the causal intervention term and α, β for the confounding constraint term. The results in Figure 5

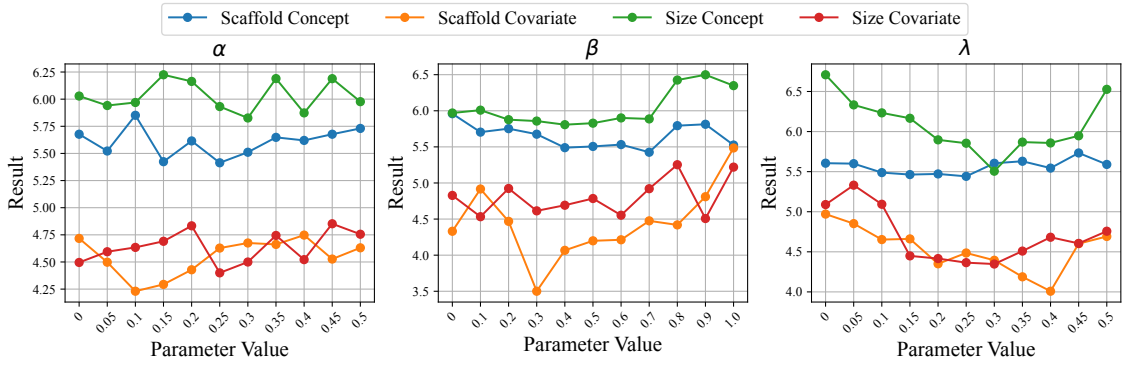


Figure 5. Parameter sensitivity.

indicate that, there is no clear trend toward getting better or worse for α. For β, which balances the GIB loss, there is a gradual increase in RMSE when it is too large, especially in scaffold-covariate settings, suggesting an optimal range around 0.3–0.6. For λ, which controls the causal intervention loss, has the strongest impact. A suitable parameter interval (0.2–0.4) consistently leads to lower RMSE, while overly large or small λ causes performance degradation, especially in the size-concept setting. This demonstrates the importance of carefully tuning λ to achieve effective OOD generalization.

B. Framework Details

Given a GNN-based encoder $f(\cdot)$ and a graph $G_i = (A_i, X_i)$, the graph representation is computed as:

$$H_{g,i} = f(A_i, X_i), \tag{18}$$

Then, to estimate attention scores, inspired by (Sui et al., 2022), we utilize separate MLPs for nodes and edges. The node attention scores, which can be seen as the node-level soft mask can be computed as:

$$\mathrm{M}_{\mathrm{node}}, \bar{\mathrm{M}}_{\mathrm{node}} = \sigma(\mathrm{MLP}_{\mathrm{node}}(H_{g,i})), \tag{19}$$

where σ denotes the softmax operation applied across attention dimensions. Similarly, edge-level soft masks are determined by concatenating node embeddings from connected edges, followed by an edge-specific MLP:

$$\mathrm{M}_{\mathrm{edge}}, \bar{\mathrm{M}}_{\mathrm{edge}} = \sigma(\mathrm{MLP}_{\mathrm{edge}}([H_{g,i}[\mathrm{row}], H_{g,i}[\mathrm{col}]])), \tag{20}$$

These soft masks serve as weighting mechanisms, allowing the model to focus on the most relevant nodes and edges while maintaining differentiability.

Next, we decompose the initial graph to causal and confounding attened-subgraph:

$$\mathrm{C}_i = \left\{A_i \odot M_{edge}, X_i \odot M_{node}\right\}, \tag{21}$$

$$\mathrm{S}_i = \left\{A_i \odot \bar{M}_{edge}, X_i \odot \bar{M}_{node}\right\}. \tag{22}$$

To encode these subgraphs, C_i and S_i are processed through a pair of GNNs with shared parameters, extracting causal and confounding representations H_c and H_s, respextively. Finally, the representations of the two subgraphs are respectively used to obtain the predictions of the regression task through the corresponding readout layers.

C. Variational Bounds for the GIB Objective

The mutual information $I(C;G)$ quantifies the dependency between C and G and is defined as:

$$I(C;G) = \mathbb{E}_{p(C,G)}\left[\log \frac{p(C \mid G)}{p(C)}\right]. \tag{23}$$

However, computing the marginal distribution $p(C) = \sum_G p(C \mid G)p(G)$ is intractable, to overcome this challenge, we approximate $p(C)$ with a variational distribution $q(C)$. Substituting $q(C)$ into Eq. (23), we reformulate $I(C;G)$ as:

$$I(C;G) = \mathbb{E}_{p(C,G)}\left[\log \frac{p(C \mid G)}{q(C)}\right] - \mathrm{KL}\big(p(C) \| q(C)\big). \tag{24}$$

The KL divergence term $\mathrm{KL}\big(p(C) \| q(C)\big)$ is non-negative, providing an upper bound for $I(C;G)$:

$$I(C;G) \leq \mathbb{E}_{p(G)}\big[\mathrm{KL}\big(p(C \mid G) \| q(C)\big)\big]. \tag{25}$$