# A Recipe for Causal Graph Regression: Confounding Effects Revisited

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

[1]Hong Kong Baptist University [2]SF Tech [3]Zhejiang University [4]HKUST

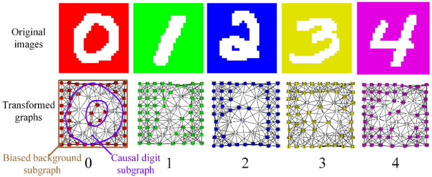
ICML 2025

# Outline

# The Challenge of Causal Graph Regression

## Motivation

- Causal Graph Learning (CGL) is crucial for out-of-distribution (OOD) generalization in fields like drug discovery and climate modeling.
- Problem: Existing CGL methods focus almost exclusively on **classification** tasks.
- Regression is a more challenging setting, and classification-specific techniques often don't apply.



(a) Examples of graphs in CMNIST-75sp.

[a]

[a]Fan, Shaohua, *et al.*, "Debiasing graph neural networks via learning disentangled causal substructure." NeurIPS 2022.

# The Challenge of Causal Graph Regression

## Our Core Idea

- We must revisit how we handle confounding effects for regression tasks.
- Existing methods assume confounders are pure noise, but in reality, they can have predictive power (e.g., molecular weight vs. toxicity).
- We need label-agnostic methods for causal intervention.

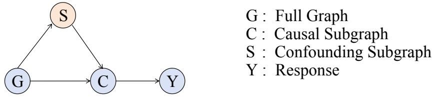



Figure: Structural Causal Model (SCM).

# Enhanced Graph Information Bottleneck (GIB)

## Limitation of Standard GIB

The standard GIB objective aims to find a small, predictive causal subgraph $C$:

$$\mathscr{L}_{\text{GIB}} = -I(C;Y) + \alpha I(C;G)$$

This implicitly assumes the confounding subgraph $S$ is non-predictive noise. This is often not true in real-world regression.

# Enhanced Graph Information Bottleneck (GIB)

## Our Proposal: An Enhanced GIB Objective

We explicitly model the predictive power of the confounding subgraph $S$ to achieve better disentanglement.

$$L_{\text{GIB}} = \underbrace{-I(C;Y) + \alpha I(C;G)}_{\text{Predictive Causal Subgraph}} \quad \underbrace{-\beta I(S;Y)}_{\text{Acknowledge Confounder's Predictive Power}}$$

- By penalizing the mutual information between $S$ and the label $Y$, we discourage the model from relying on spurious correlations from $S$.
- This encourages a cleaner separation of causal ($C$) and confounding ($S$) factors.
- These terms are made practical using variational bounds, resulting in simple regression and regularization losses.

# Causal Intervention via Contrastive Learning

## The Problem with Traditional Intervention

Methods like backdoor adjustment often rely on stratifying by **class labels** to block confounding paths: $P(Y|\mathrm{do}(C)) = \sum_s P(Y|C,s)P(s)$.

- This is infeasible for regression tasks with continuous labels $Y$.

# Causal Intervention via Contrastive Learning

## The Problem with Traditional Intervention

Methods like backdoor adjustment often rely on stratifying by **class labels** to block confounding paths: $P(Y|\mathrm{do}(C)) = \sum_s P(Y|C,s)P(s)$.

- This is infeasible for regression tasks with continuous labels $Y$.

## Our Proposal: A Label-Agnostic Intervention

We generalize intervention from "class separation" to "instance discrimination" using contrastive learning.

- **Create Counterfactuals:** Mix the causal part of a graph $i$ with a *random* confounding part from another graph $j$: $H_{\mathrm{mix},ij} = H_{c,i} + H_{s,j}$.

# Causal Intervention via Contrastive Learning

## The Problem with Traditional Intervention

Methods like backdoor adjustment often rely on stratifying by **class labels** to block confounding paths: $P(Y|\mathrm{do}(C)) = \sum_s P(Y|C,s)P(s)$.

- This is infeasible for regression tasks with continuous labels $Y$.

## Our Proposal: A Label-Agnostic Intervention

We generalize intervention from "class separation" to "instance discrimination" using contrastive learning.

- **Create Counterfactuals:** Mix the causal part of a graph $i$ with a *random* confounding part from another graph $j$: $H_{\mathrm{mix},ij} = H_{c,i} + H_{s,j}$.
- **Contrastive Objective:** A robust causal representation should be invariant to the confounding part it's mixed with.

# Causal Intervention via Contrastive Learning

## Our Proposal: The Contrastive Objective

The core idea is implemented with the InfoNCE loss:

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

This pulls the original graph representation ($H_{g,i}$) and its counterfactual version ($H_{\mathrm{mix},ij}$) together, learning confounder-invariant causal features.

# Overall Framework

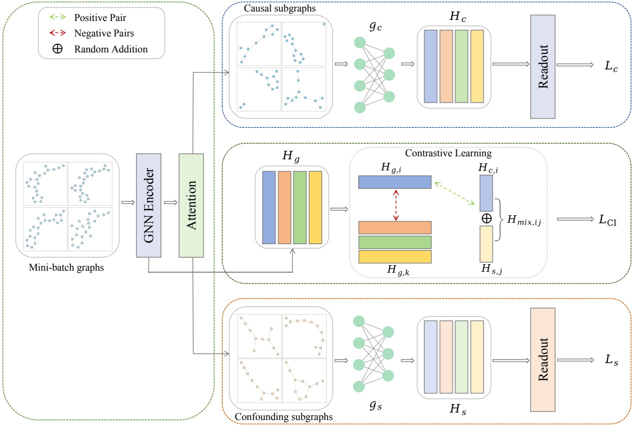
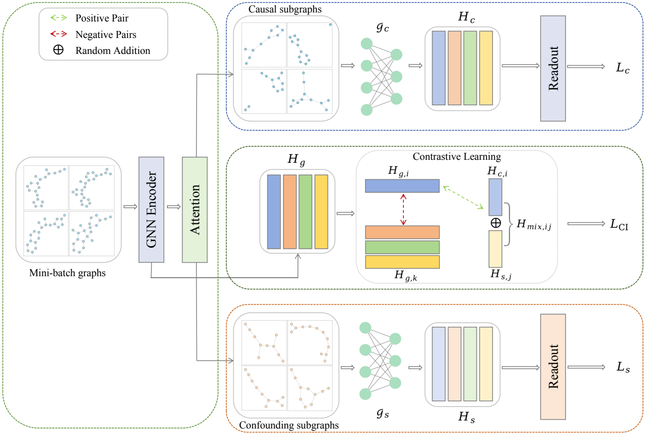



Figure: The proposed CGR framework.

# Overall Framework: The Objective




## Final Objective

The final loss combines our two proposals:

$$L = \underbrace{L_{\text{GIB}}}_{\text{GIB}} + \lambda \underbrace{L_{\text{CI}}}_{\text{CI}}$$

# Main Results on GOOD-ZINC Benchmark

- **Benchmark:** GOOD-ZINC (Molecular property regression)
- **Challenge:** OOD Generalization (Scaffold  Size shifts)
- **Metric:** Mean Absolute Error (MAE), lower is better.

*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** |

## Key Takeaway

Our CGR framework achieves SOTA performance on GOOD regression benchmark.

# Main Results on ReactionOOD Benchmark

- **Benchmark:** ReactionOOD (Chemical reaction kinetics).
- **Metric:** Root Mean Square Error (RMSE), lower is better.

*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** |

# Main Results on ReactionOOD Benchmark

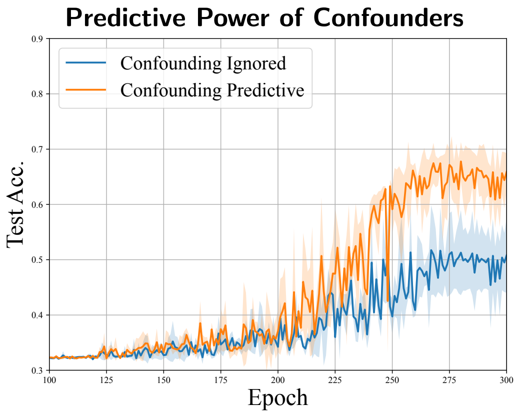
*Table 3.* OOD generalization performance on E2&$S_N2$ dataset.

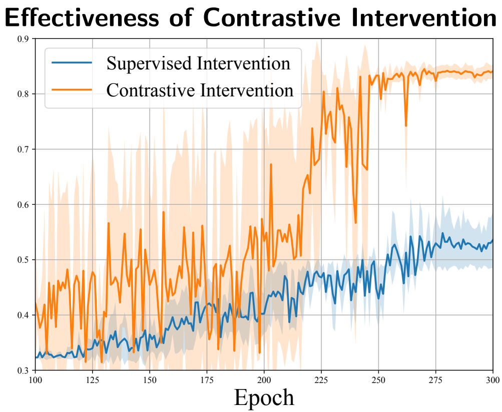
| Methods | Covariate | | Concept | |
|---|---|---|---|---|
| | ID | OOD | ID | OOD |
| ERM | 4.45±0.04 | 5.47±0.27 | 4.87±0.02 | 5.04±0.02 |
| IRM | 11.61±0.18 | 21.54±1.07 | 20.95±0.02 | 17.57±0.03 |
| VREx | 4.58±0.02 | 5.48±0.13 | 10.75±1.54 | 8.77±2.31 |
| Mixup | 4.55±0.09 | 5.55±0.01 | 4.69±0.08 | 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 |

## Key Takeaway

Our CGR framework significantly outperforms baselines on ReactionOOD benchmarks.

# Ablation Studies and Generality




- Modeling confounder's predictive power (**Confounding Predictive**) leads to better OOD accuracy than ignoring it.

- Our contrastive loss (**Contrastive Intervention**) is more effective for intervention than baseline methods.

# Conclusion

## Our Contributions

- We provide the first systematic recipe for **Causal Graph Regression (CGR)**, addressing a critical gap in CGL research.

### Rethinking Confounders

We propose an **enhanced GIB objective** that acknowledges and models the predictive power of confounding features, leading to better causal disentanglement.

### Rethinking Intervention

We introduce a **causal intervention loss based on contrastive learning**, which is powerful, effective, and crucially, does not depend on labels.

## Code Availability

Our code will be open-source: `https://github.com/causal-graph/CGR`