KDD-23 Research Track Paper

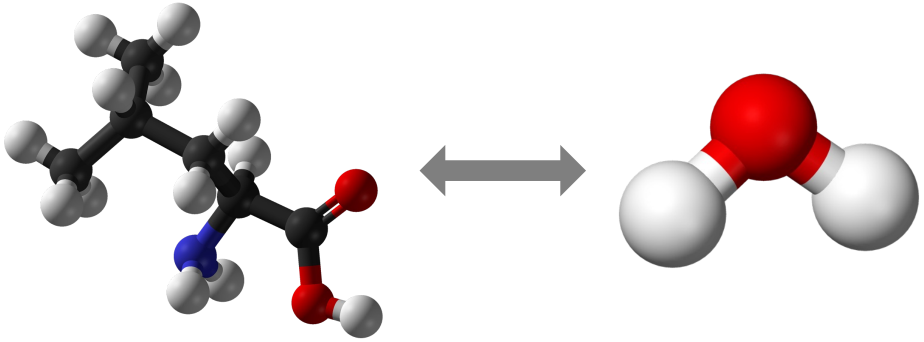
Shift-Robust Molecular Relational Learning with Causal Substructure

Namkyeong Lee, Kanghoon Yoon, Gyoung S. Na,
Sein Kim, Chanyoung Park

TABLE OF CONTENTS

- **Background**
  - Molecular Relational Learning
  - Distribution Shift in Molecules
  - Causal Inference
  - Causal Inference for Graph Structured Data
- **Methodology: Shift-Robust Molecular Relational Learning with Causal Substructure**
- **Experiments**
- **Conclusion**

BACKGROUND MOLECULAR RELATIONAL LEARNING



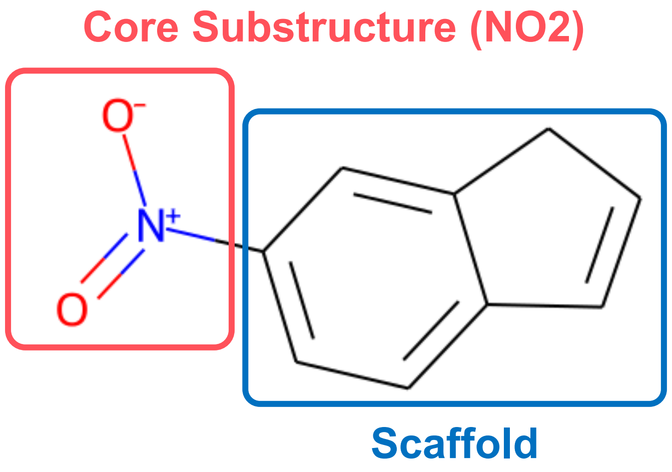
Molecular Relational Learning

Learning the interaction behavior between a pair of molecules

Examples

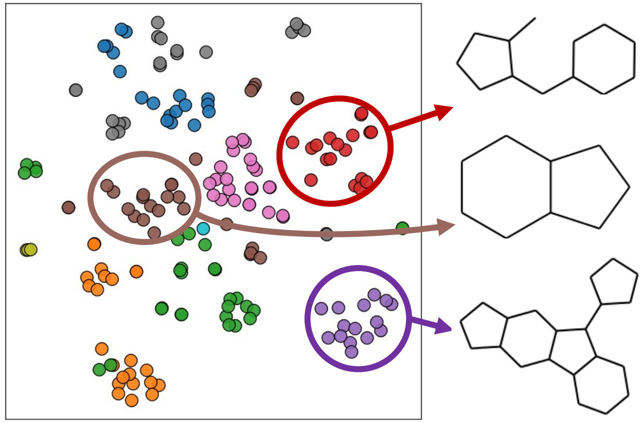
- Predicting solubility when a **drug** and **solvent** react
- Predicting side effects when taking **two types of drugs** simultaneously
- Predicting optical properties when a **Chromophore** and **Solvent** react

BACKGROUND DISTRIBUTION SHIFT IN MOLECULES


Core Substructure (NO2)

Scaffold

Molecule: 6-nitro-1H-indene



Molecular fingerprints with various scaffolds

Molecules with different scaffolds exhibit distinct distributions

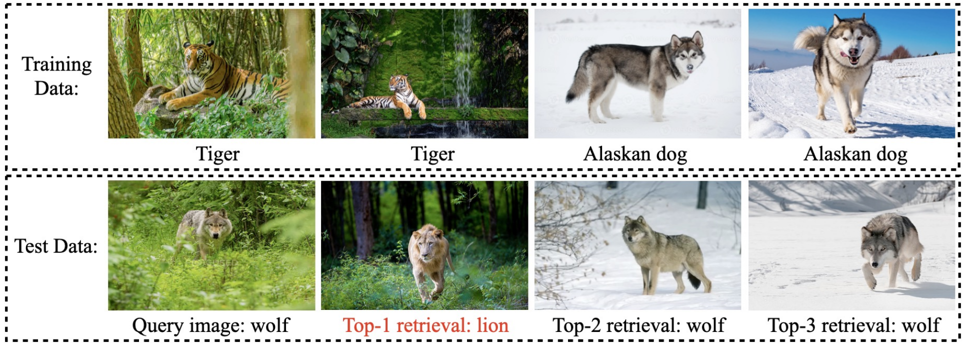
→ Learning from core substructure is crucial for the robustness of machine learning (ML) models to distribution shifts

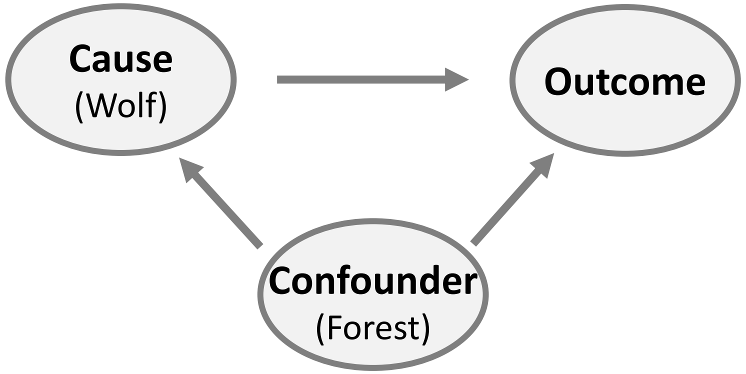
→ Enabling ML models to learn more generalized knowledge in molecules!

* Molecules with nitrogen dioxide (NO2) functional group commonly exhibit the mutagenic property

* Scaffold: The common structure characterizing a group of molecules

BACKGROUND CAUSAL INFERENCE





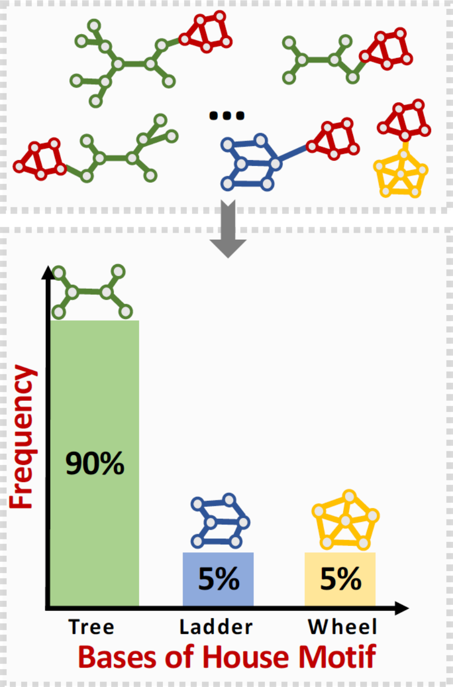
Due to the empirical process of data collection,
the data for machine learning is heavily biased

Context of the given data becomes a confounder
that misleads the machine learning model to learn spurious correlations
between pixels and labels

Ex) Spurious correlation between forest and lion in Figure

Causal Inference aims to improve model performance by removing spurious correlations

BACKGROUND CAUSAL INFERENCE FOR GRAPH STRUCTURED DATA



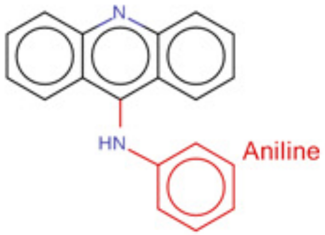
Determining House Motifs

Spurious correlation between the Tree motifs with House motifs

When facing with out-of-distribution (OOD) data,
statistical shortcuts will severely deteriorates the model performance

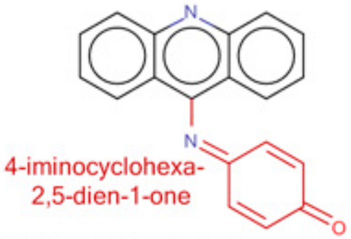
BACKGROUND CAUSAL INFERENCE FOR GRAPH STRUCTURED DATA

Mutagenic



(3) N-phenylacridin-9-amine
0.94 (17/18)

Non-Mutagenic



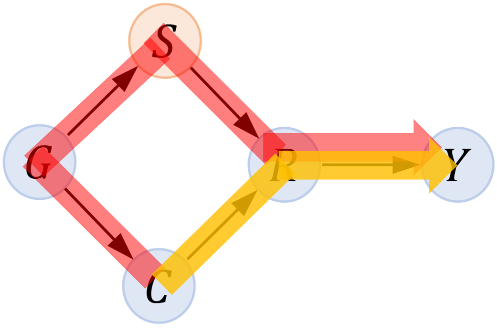
(24) 4-acridine-9-yliminocyclohexa-
2,5-dien-1-one
0 (0/1)

Instead of probing into the causal effect of the functional groups, Model focuses on “carbon rings” as the cues of the mutagenic class

In fact, “Carbon ring” has no relationship with mutagenicity

Spurious correlation becomes even severe in molecules!

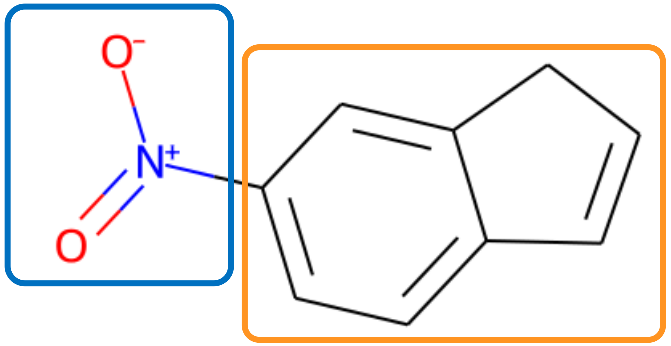
BACKGROUND STRUCTURAL CAUSAL MODEL



G : graph data
C : causal feature
S : shortcut feature
R : representation
Y : prediction

Structure Causal Model (SCM) for
molecular property prediction

Causal feature



Shortcut feature

Causal-Effect relationship in molecular property prediction

$\underline{C \leftarrow G \rightarrow S}$: $\underline{C}$ and $\underline{S}$ naturally coexist in molecule $\underline{G}$.
$\underline{C \rightarrow R \leftarrow S}$: The variable $\underline{R}$ is the representation of the given molecule $\underline{G}$.

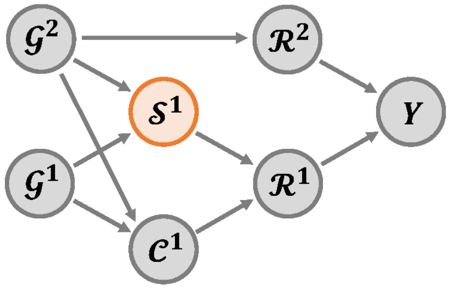
➡ $\underline{C \rightarrow R \rightarrow Y}$: Causality we are interested in

➡ $\underline{C \leftarrow G \rightarrow S \rightarrow R \rightarrow Y}$: Backdoor path

KDD-23 Research Track Paper

Shift-Robust Molecular Relational Learning with Causal Substructure

METHODOLOGY CAUSALITY IN MOLECULAR RELATIONAL LEARNING



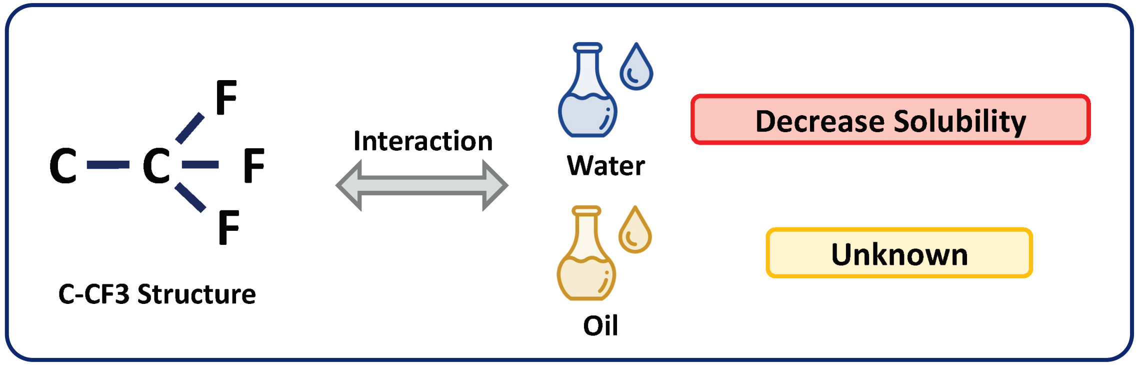
$\mathcal{G}^1$: **Molecule 1**
$\mathcal{G}^2$: **Molecule 2**
$\mathcal{C}^1$: **Causal Substructure in Molecule 1**
$\mathcal{S}^1$: **Shortcut Substructure in Molecule 1**
$\mathcal{R}^1$: **Molecule 1 Representation**
$\mathcal{R}^2$: **Molecule 2 Representation**
Y : **Target Value**

Structure Causal Model (SCM) for
Molecular Relational Learning

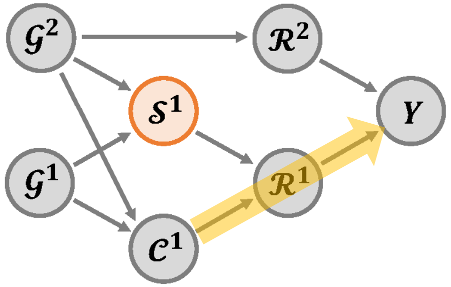
Key causal-effect relationship in molecular relational learning

$$\mathcal{G}^1 \longrightarrow \mathcal{C}^1 \longleftarrow \mathcal{G}^2$$

Causal substructure $\mathcal{C}^1$ of molecule $\mathcal{G}^1$
→ Determined by not only $\mathcal{G}^1$but also $\mathcal{G}^2$



METHODOLOGY CAUSALITY IN MOLECULAR RELATIONAL LEARNING



$\mathcal{G}^1$: Molecule 1
$\mathcal{G}^2$: Molecule 2
$\mathcal{C}^1$: Causal Substructure in Molecule 1
$\mathcal{S}^1$: Shortcut Substructure in Molecule 1
$\mathcal{R}^1$: Molecule 1 Representation
$\mathcal{R}^2$: Molecule 2 Representation
Y : Target Value

Structure Causal Model (SCM) for
Molecular Relational Learning

➡ Causality we are interested in ($\mathcal{C}^1 \rightarrow Y$)

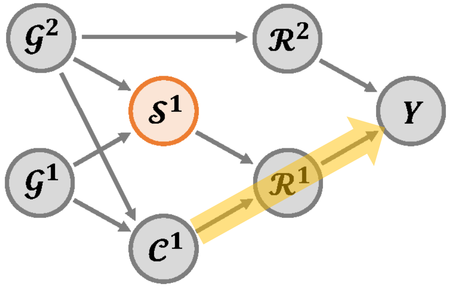
4 Backdoor paths that confound the model

$$\mathcal{C}^1 \leftarrow \mathcal{G}^1 \rightarrow \mathcal{S}^1 \leftarrow \mathcal{G}^2 \rightarrow \mathcal{R}^2 \rightarrow Y$$
$$\mathcal{C}^1 \leftarrow \mathcal{G}^2 \rightarrow \mathcal{R}^2 \rightarrow Y$$
$$\mathcal{C}^1 \leftarrow \mathcal{G}^2 \rightarrow \mathcal{S}^1 \rightarrow \mathcal{R}^1 \rightarrow Y$$
$$\mathcal{C}^1 \leftarrow \mathcal{G}^1 \rightarrow \mathcal{S}^1 \rightarrow \mathcal{R}^1 \rightarrow Y$$

In molecular relational learning,
$\mathcal{G}^2$ is given and utilized during model prediction

$\mathcal{C}^1 \leftarrow \mathcal{G}^1 \rightarrow \mathcal{S}^1 \rightarrow \mathcal{R}^1 \rightarrow Y$ Only remaining backdoor path!

METHODOLOGY BACKDOOR ADJUSTMENT



$\mathcal{G}^1$: Molecule 1
$\mathcal{G}^2$: Molecule 2
$\mathcal{C}^1$: Causal Substructure in Molecule 1
$\mathcal{S}^1$: Shortcut Substructure in Molecule 1
$\mathcal{R}^1$: Molecule 1 Representation
$\mathcal{R}^2$: Molecule 2 Representation
Y : Target Value

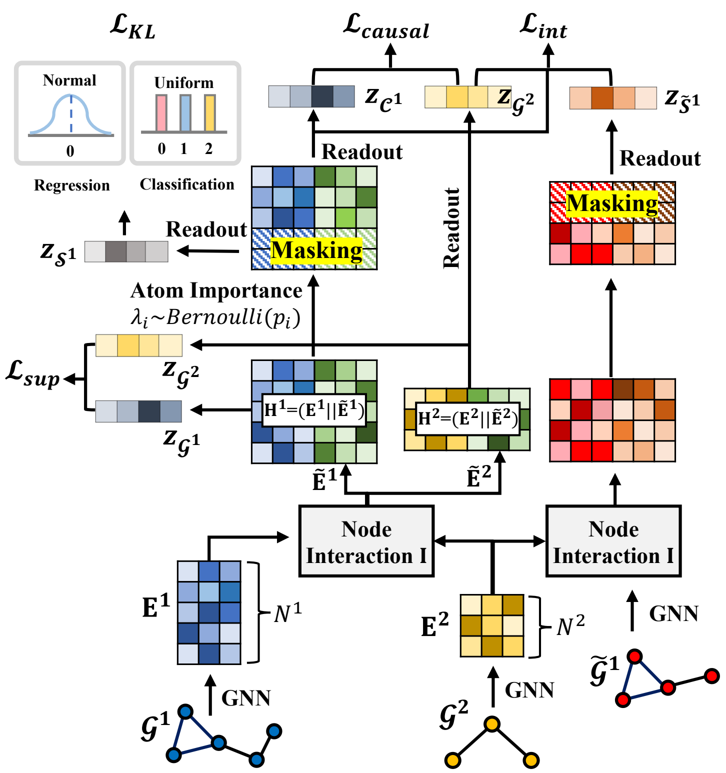
Structure Causal Model (SCM) for Molecular Relational Learning

$$
\begin{aligned}
P(\mathrm{Y}|do(C^1),\mathcal{G}^2) &= \tilde{P}(\mathrm{Y}|C^1,\mathcal{G}^2) \\
&= \sum_s \tilde{P}(\mathrm{Y}|C^1,\mathcal{G}^2,s)\cdot\tilde{P}(s|C^1,\mathcal{G}^2) \text{ (Bayes' Rule)} \\
&= \sum_s \tilde{P}(\mathrm{Y}|C^1,\mathcal{G}^2,s)\cdot\tilde{P}(s|\mathcal{G}^2) \text{ (Independence)} \\
&= \sum_s P(\mathrm{Y}|C^1,\mathcal{G}^2,s)\cdot P(s|\mathcal{G}^2),
\end{aligned}
$$

Backdoor Adjustment

Alleviate confounding effect via Backdoor adjustment!

METHODOLOGY CAUSAL MOLECULAR RELATIONAL LEARNER



Disentangling with Atom Representation Masks

Separate the causal substructure $\mathcal{C}^1$and shortcut substructure $\mathcal{S}^1$ from $\mathcal{G}^1$

→ Not trivial to explicitly manipulate molecular structure

→ Let's separate in representation space by masking atom representation!

$$p_i = \mathrm{MLP}(\mathrm{H}_i^1)$$ Importance of atom i

$$\mathrm{C}_i^1 = \lambda_i \mathrm{H}_i^1 + (1-\lambda_i)\epsilon$$ Causal substructure

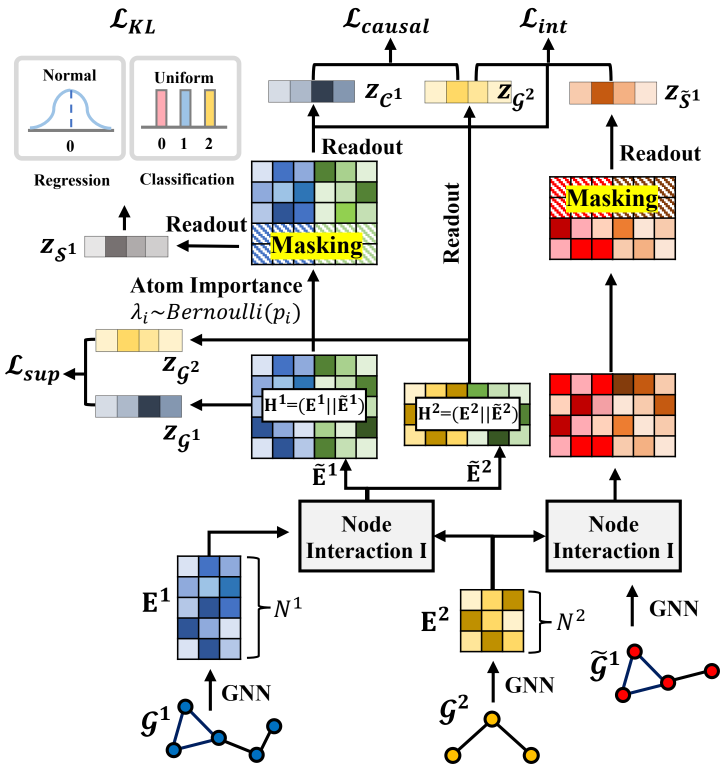
$$\mathrm{S}_i^1 = (1-\lambda_i)\mathrm{H}_i^1$$ Shortcut substructure

where

$$\lambda_i \sim \mathrm{Bernoulli}(p_i) \quad \epsilon \sim N(\mu_{\mathrm{H}^1}, \sigma_{\mathrm{H}^1}^2)$$

Gumbel sigmoid approach for differentiable optimization of p_i

METHODOLOGY CAUSAL MOLECULAR RELATIONAL LEARNER



Disentangling with Atom Representation Masks

Causal substructure $\mathcal{C}^1$

→ Cross entropy loss for classification

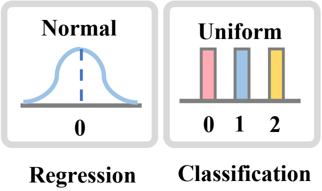
→ RMSE loss for Regression

$$\mathcal{L}_{causal}(\mathrm{Y}, z_{C^1}, z_{\mathcal{G}^2})$$

Shortcut substructure $\mathcal{S}^1$

→ Learn non informative distribution

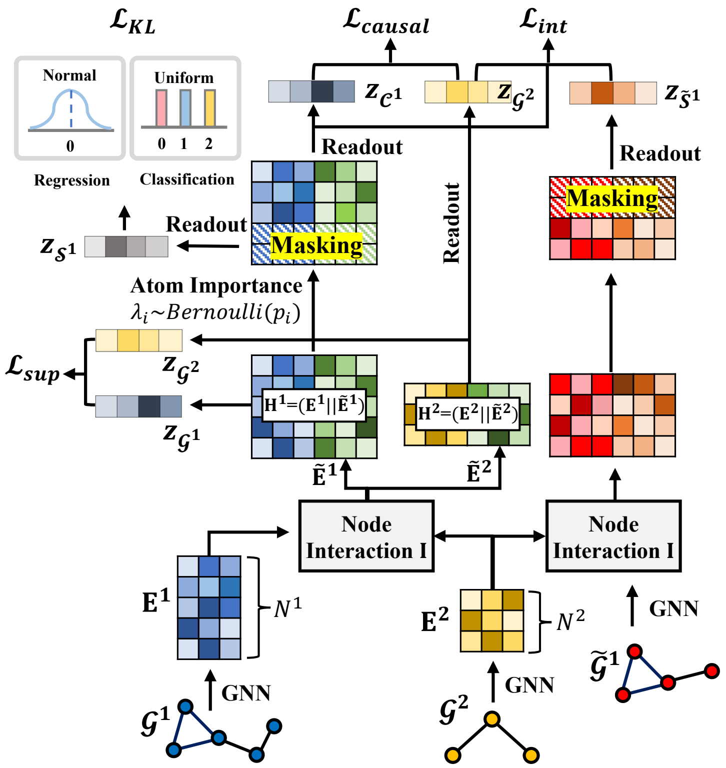
$$\mathcal{L}_{KL}(\mathrm{Y}_{rand}, z_{\mathcal{S}^1})$$



METHODOLOGY CAUSAL MOLECULAR RELATIONAL LEARNER

$$
\begin{aligned}
P(\mathrm{Y}|do(C^1),\mathcal{G}^2) &= \tilde{P}(\mathrm{Y}|C^1,\mathcal{G}^2) \\
&= \sum_s \tilde{P}(\mathrm{Y}|C^1,\mathcal{G}^2,s)\cdot\tilde{P}(s|C^1,\mathcal{G}^2) \text{ (Bayes' Rule)} \\
&= \sum_s \tilde{P}(\mathrm{Y}|C^1,\mathcal{G}^2,s)\cdot\tilde{P}(s|\mathcal{G}^2) \text{ (Independence)} \\
&= \sum_s P(\mathrm{Y}|C^1,\mathcal{G}^2,s)\cdot P(s|\mathcal{G}^2),
\end{aligned}
$$

Backdoor Adjustment



Conditional Causal Intervention via backdoor adjustment

Straightforward approach → Synthesize / Collect various molecules

Challenges

1) Expensive time/financial costs

2) Intervention space on $\boldsymbol{\mathcal{C}^1}$ should be conditioned on the paired molecule $\boldsymbol{\mathcal{G}^2}$

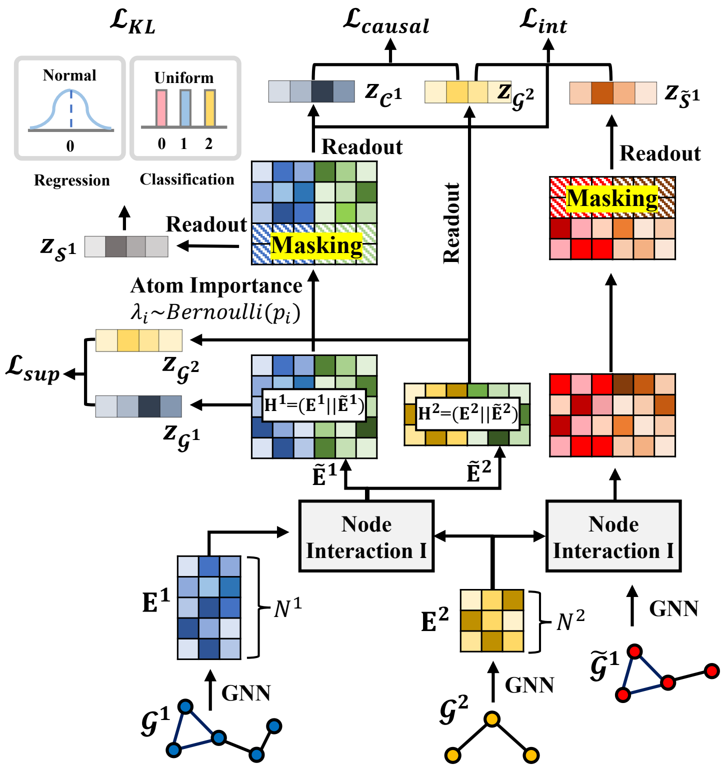
Our Solution

Obtain shortcut substructure $\tilde{\boldsymbol{\mathcal{S}}}^{\mathbf{1}}$

by modeling interaction with other molecules $\tilde{\boldsymbol{\mathcal{G}}}^{\mathbf{1}}$ and molecule $\boldsymbol{\mathcal{G}^2}$

$$\mathcal{L}_{int} = \sum_{(\mathcal{G}^1,\mathcal{G}^2)\in\mathcal{D}} \sum_{\tilde{\mathcal{S}}^1} \mathcal{L}(\mathrm{Y}, z_{\mathcal{C}^1}, z_{\mathcal{G}^2}, z_{\tilde{\mathcal{S}}^1})$$

METHODOLOGY CAUSAL MOLECULAR RELATIONAL LEARNER



Final Objective

$$\mathcal{L}_{final} = \mathcal{L}_{sup} + \mathcal{L}_{causal} + \lambda_1 \cdot \mathcal{L}_{KL} + \lambda_2 \cdot \mathcal{L}_{int}$$

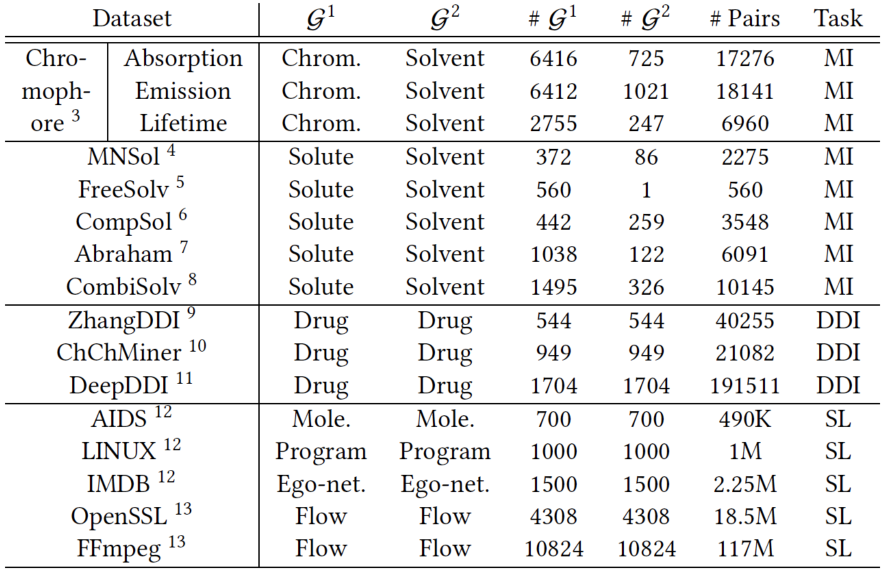
$\mathcal{L}_{sup}$: loss with paired graph $(\boldsymbol{\mathcal{G}^1}, \boldsymbol{\mathcal{G}^2})$ and target $\boldsymbol{Y}$

$\mathcal{L}_{causal}$: loss with causal substructure

$\mathcal{L}_{KL}$: loss with shortcut substructure

λ_1, λ_2: weight hyperparameters for $\mathcal{L}_{KL}$ and $\mathcal{L}_{int}$

EXPERIMENTS DATASET DESCRIPTION



Dataset		$\mathcal{G}^1$	$\mathcal{G}^2$	# $\mathcal{G}^1$	# $\mathcal{G}^2$	# Pairs	Task
Chromophore [3]	Absorption	Chrom.	Solvent	6416	725	17276	MI
	Emission	Chrom.	Solvent	6412	1021	18141	MI
	Lifetime	Chrom.	Solvent	2755	247	6960	MI
MNSol [4]		Solute	Solvent	372	86	2275	MI
FreeSolv [5]		Solute	Solvent	560	1	560	MI
CompSol [6]		Solute	Solvent	442	259	3548	MI
Abraham [7]		Solute	Solvent	1038	122	6091	MI
CombiSolv [8]		Solute	Solvent	1495	326	10145	MI
ZhangDDI [9]		Drug	Drug	544	544	40255	DDI
ChChMiner [10]		Drug	Drug	949	949	21082	DDI
DeepDDI [11]		Drug	Drug	1704	1704	191511	DDI
AIDS [12]		Mole.	Mole.	700	700	490K	SL
LINUX [12]		Program	Program	1000	1000	1M	SL
IMDB [12]		Ego-net.	Ego-net.	1500	1500	2.25M	SL
OpenSSL [13]		Flow	Flow	4308	4308	18.5M	SL
FFmpeg [13]		Flow	Flow	10824	10824	117M	SL

Molecular Interaction Dataset

→ Predicting Chromophores' Absorption max, Emission max, Lifetime

→ Predicting Solvation Free Energy of molecules (MNSol, FreeSolv, CompSol, Abraham, CombiSolv)

→ Regression Task

Drug-Drug Interaction Dataset

→ Zhang DDI, ChChMiner, DeepDDI

→ Classification Task

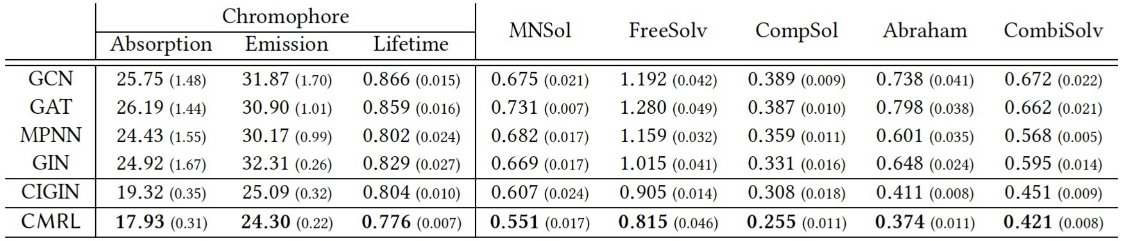
Graph Similarity Learning Dataset

→ How similar are the paired graphs? (ex. GED)

→ AIDS, LINUX, IMDB, OpenSSL, Ffmpeg

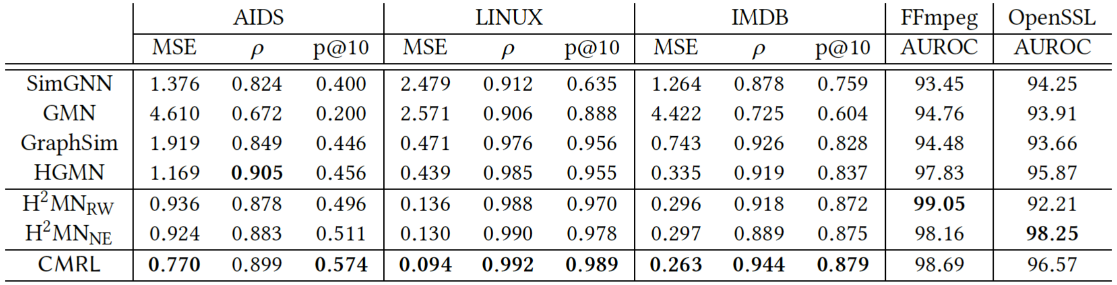
→ Regression Task / Classification Task

EXPERIMENTS OVERALL PERFORMANCE



	Chromophore			MNSol	FreeSolv	CompSol	Abraham	CombiSolv
	Absorption	Emission	Lifetime					
GCN	25.75 (1.48)	31.87 (1.70)	0.866 (0.015)	0.675 (0.021)	1.192 (0.042)	0.389 (0.009)	0.738 (0.041)	0.672 (0.022)
GAT	26.19 (1.44)	30.90 (1.01)	0.859 (0.016)	0.731 (0.007)	1.280 (0.049)	0.387 (0.010)	0.798 (0.038)	0.662 (0.021)
MPNN	24.43 (1.55)	30.17 (0.99)	0.802 (0.024)	0.682 (0.017)	1.159 (0.032)	0.359 (0.011)	0.601 (0.035)	0.568 (0.005)
GIN	24.92 (1.67)	32.31 (0.26)	0.829 (0.027)	0.669 (0.017)	1.015 (0.041)	0.331 (0.016)	0.648 (0.024)	0.595 (0.014)
CIGIN	19.32 (0.35)	25.09 (0.32)	0.804 (0.010)	0.607 (0.024)	0.905 (0.014)	0.308 (0.018)	0.411 (0.008)	0.451 (0.009)
CMRL	**17.93** (0.31)	**24.30** (0.22)	**0.776** (0.007)	**0.551** (0.017)	**0.815** (0.046)	**0.255** (0.011)	**0.374** (0.011)	**0.421** (0.008)

Performance on molecular interaction prediction task



	AIDS			LINUX			IMDB			FFmpeg	OpenSSL
	MSE	ρ	p@10	MSE	ρ	p@10	MSE	ρ	p@10	AUROC	AUROC
SimGNN	1.376	0.824	0.400	2.479	0.912	0.635	1.264	0.878	0.759	93.45	94.25
GMN	4.610	0.672	0.200	2.571	0.906	0.888	4.422	0.725	0.604	94.76	93.91
GraphSim	1.919	0.849	0.446	0.471	0.976	0.956	0.743	0.926	0.828	94.48	93.66
HGMN	1.169	**0.905**	0.456	0.439	0.985	0.955	0.335	0.919	0.837	97.83	95.87
H^2MN_{RW}	0.936	0.878	0.496	0.136	0.988	0.970	0.296	0.918	0.872	**99.05**	92.21
H^2MN_{NE}	0.924	0.883	0.511	0.130	0.990	0.978	0.297	0.889	0.875	98.16	**98.25**
CMRL	**0.770**	0.899	**0.574**	**0.094**	**0.992**	**0.989**	**0.263**	**0.944**	**0.879**	98.69	96.57

Performance on graph similarity learning task

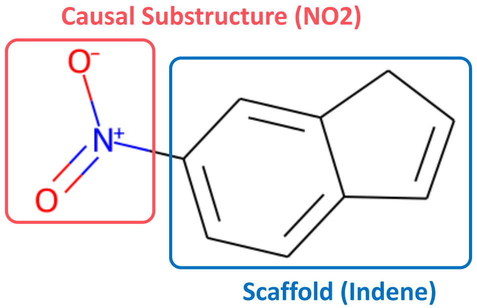
Observations

1. CMRL outperforms all other baseline methods
→ It is crucial to discover causally related substructure in molecules

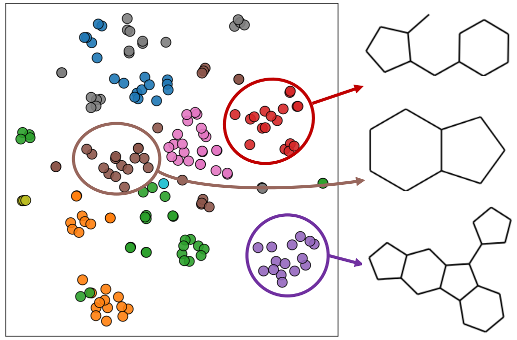
2. Wide applicability of CMRL beyond molecules
→ Performs well in dataset that contains core substructure

EXPERIMENTS OUT-OF-DISTRIBUTION PERFORMANCE

In out-of-distribution experiment, we assess the model’s performance on molecules belonging to new scaffold classes

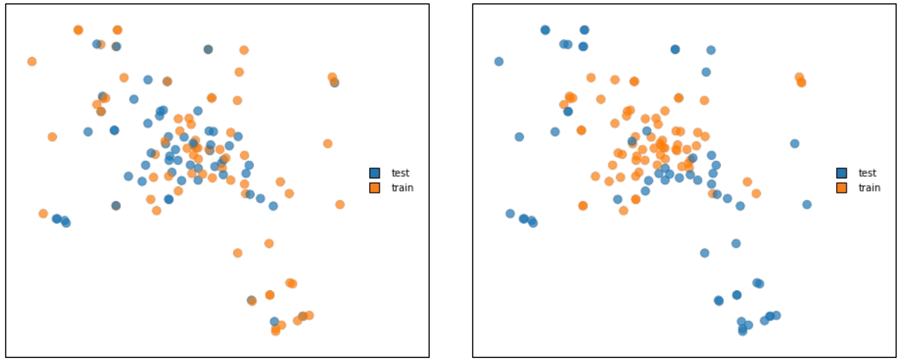


Molecule: 6-nitro-1H-indene



TSNE embeddings

Different scaffolds exhibit totally different distribution

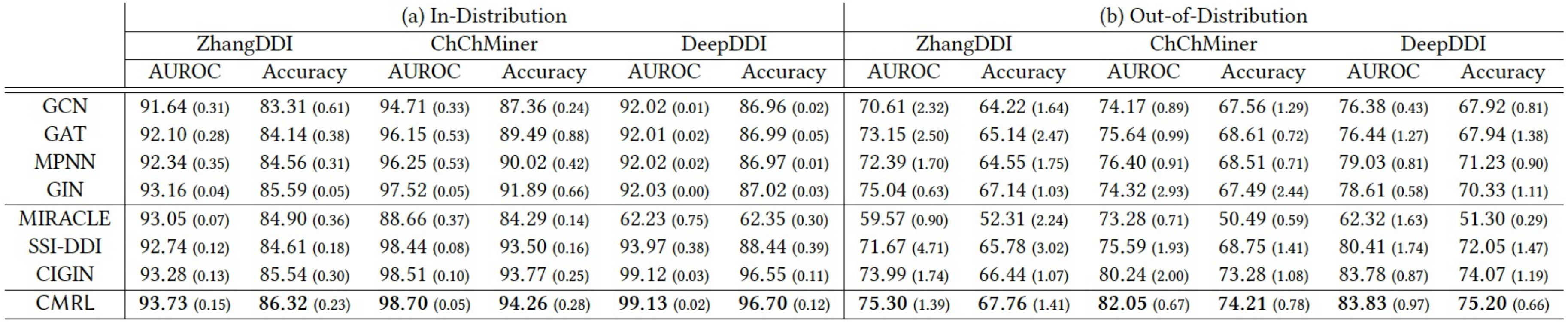


Random Split **Scaffold Split**

TSNE on splitted data (Train / Test)

EXPERIMENTS OUT-OF-DISTRIBUTION PERFORMANCE

In out-of-distribution experiment, we assess the model's performance on molecules belonging to new scaffold classes



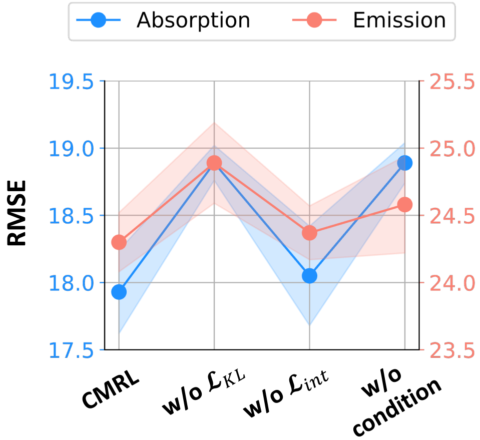
	(a) In-Distribution						(b) Out-of-Distribution					
	ZhangDDI		ChChMiner		DeepDDI		ZhangDDI		ChChMiner		DeepDDI	
	AUROC	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC	Accuracy
GCN	91.64 (0.31)	83.31 (0.61)	94.71 (0.33)	87.36 (0.24)	92.02 (0.01)	86.96 (0.02)	70.61 (2.32)	64.22 (1.64)	74.17 (0.89)	67.56 (1.29)	76.38 (0.43)	67.92 (0.81)
GAT	92.10 (0.28)	84.14 (0.38)	96.15 (0.53)	89.49 (0.88)	92.01 (0.02)	86.99 (0.05)	73.15 (2.50)	65.14 (2.47)	75.64 (0.99)	68.61 (0.72)	76.44 (1.27)	67.94 (1.38)
MPNN	92.34 (0.35)	84.56 (0.31)	96.25 (0.53)	90.02 (0.42)	92.02 (0.02)	86.97 (0.01)	72.39 (1.70)	64.55 (1.75)	76.40 (0.91)	68.51 (0.71)	79.03 (0.81)	71.23 (0.90)
GIN	93.16 (0.04)	85.59 (0.05)	97.52 (0.05)	91.89 (0.66)	92.03 (0.00)	87.02 (0.03)	75.04 (0.63)	67.14 (1.03)	74.32 (2.93)	67.49 (2.44)	78.61 (0.58)	70.33 (1.11)
MIRACLE	93.05 (0.07)	84.90 (0.36)	88.66 (0.37)	84.29 (0.14)	62.23 (0.75)	62.35 (0.30)	59.57 (0.90)	52.31 (2.24)	73.28 (0.71)	50.49 (0.59)	62.32 (1.63)	51.30 (0.29)
SSI-DDI	92.74 (0.12)	84.61 (0.18)	98.44 (0.08)	93.50 (0.16)	93.97 (0.38)	88.44 (0.39)	71.67 (4.71)	65.78 (3.02)	75.59 (1.93)	68.75 (1.41)	80.41 (1.74)	72.05 (1.47)
CIGIN	93.28 (0.13)	85.54 (0.30)	98.51 (0.10)	93.77 (0.25)	99.12 (0.03)	96.55 (0.11)	73.99 (1.74)	66.44 (1.07)	80.24 (2.00)	73.28 (1.08)	83.78 (0.87)	74.07 (1.19)
CMRL	**93.73** (0.15)	**86.32** (0.23)	**98.70** (0.05)	**94.26** (0.28)	**99.13** (0.02)	**96.70** (0.12)	**75.30** (1.39)	**67.76** (1.41)	**82.05** (0.67)	**74.21** (0.78)	**83.83** (0.97)	**75.20** (0.66)

Performance on drug-drug interaction task

Observation

CMRL outperforms previous work on out-of-distribution scenarios
→ Learning causal substructure enhances the generalization ability of the model

EXPERIMENTS MODEL ANALYSIS

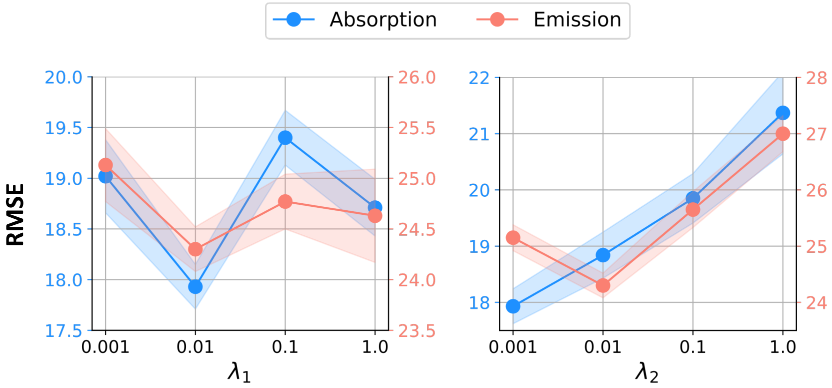


Observations in Ablation Studies

Naïve intervention whose confounders are not conditioned on paired molecule $\mathcal{G}^2$

→ Performs worse than the model without intervention

→ Wideness of intervention space introduces noisy signal during model training



$\lambda_2 = 0.001$ for Absorption
$\lambda_2 = 0.01$ for Emission

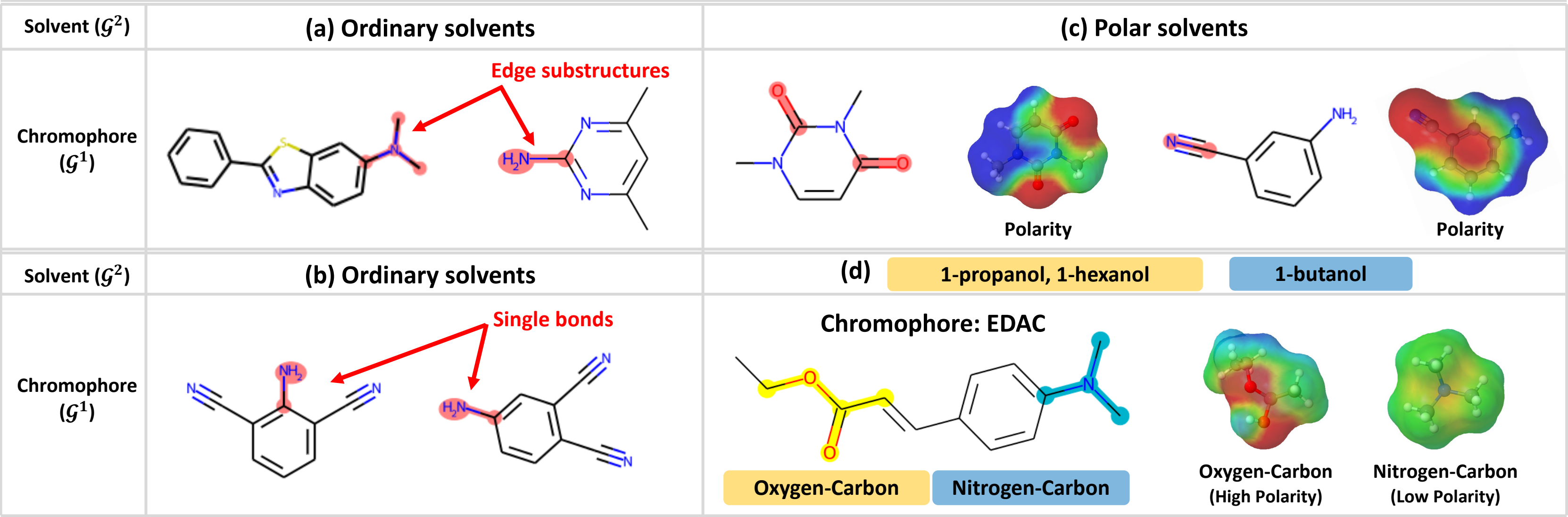
$\lambda_1 = 0.01$ for Absorption
$\lambda_1 = 0.01$ for Emission

Observations in Sensitivity Analysis

1. Optimal point for λ_2 exist balancing the noisiness and robustness
2. No certain relationship between model performance and λ_1

Training objective $\mathcal{L}_{final} = \mathcal{L}_{sup} + \mathcal{L}_{causal} + \lambda_1 \cdot \mathcal{L}_{KL} + \lambda_2 \cdot \mathcal{L}_{int}$

EXPERIMENTS QUALITATIVE ANALYSIS



Observations

1. Discovered causal substructure aligns to well-known chemical domain knowledge
- (a) CMRL selects edge substructure → Chemical reactions usually happen around ionized atoms
- (b) CMRL concentrates on single-bonded substructure → Single-bonded substructures are more likely to undergo chemical reactions

2. (c) When reacting with polar solvents, CMRL focuses on the edge substructures of high polarity

3. (d) Selected important substructures of chromophore varies as the solvent varies

CONCLUSION

This paper proposed a method for tackling relation learning tasks, which are prevalent in various scientific field

Keyword: Conditional causal intervention

→ Crucial to narrow down intervention space by conditioning on paired molecule $\mathcal{G}^2$

Extensive experiments demonstrating the superiority and interpretability of CMRL

→ Makes CMRL highly practical for real-world scientific discovery

[Full Paper] https://arxiv.org/abs/2305.18451

[Source Code] https://github.com/Namkyeong/CMRL

[Author Email] namkyeong96@kaist.ac.kr

Paper

Code

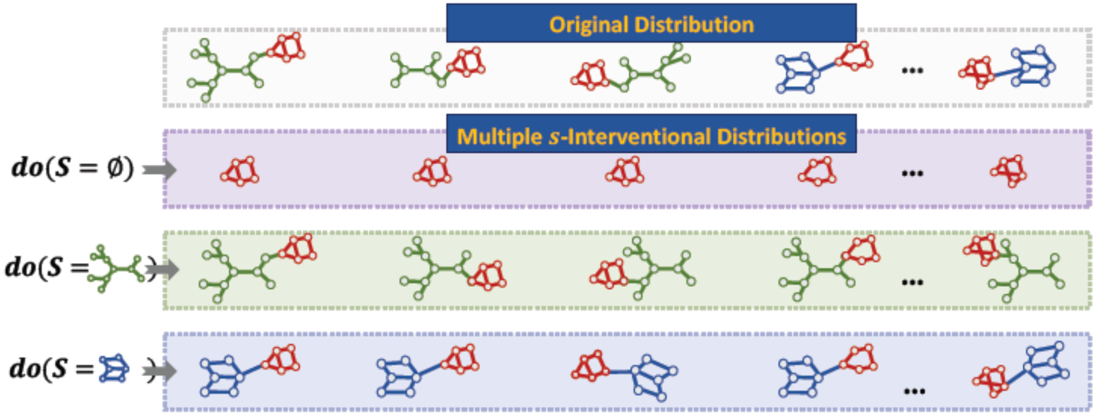
Appendix

RELATED WORKS

Task: Rationalization for GNNs → "What knowledge drives the GNNs to make certain predictions?"

Invariant Learning

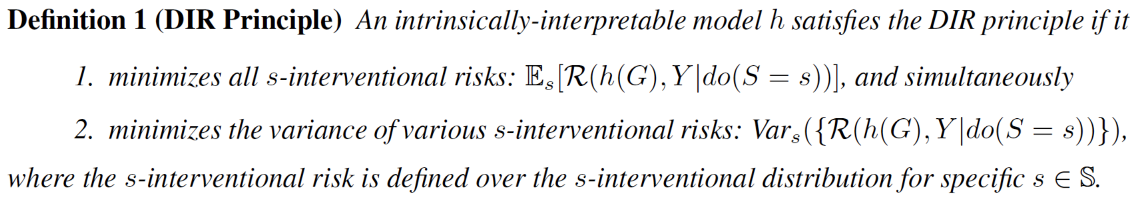
→ Constructs different environments to infer the invariant features or predictors



Generate s-interventional distribution by doing intervention on $\boldsymbol{S}$

Discovering Invariant Rationales for Graph Neural Networks. ICLR 2022

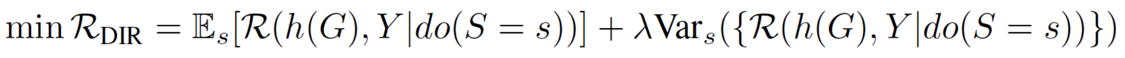
RELATED WORKS



Definition 1 (DIR Principle) *An intrinsically-interpretable model h satisfies the DIR principle if it*

1. *minimizes all s-interventional risks:* $\mathbb{E}_s[\mathcal{R}(h(G), Y|do(S=s))]$, *and simultaneously*
2. *minimizes the variance of various s-interventional risks:* $Var_s(\{\mathcal{R}(h(G), Y|do(S=s))\})$,

where the s-interventional risk is defined over the s-interventional distribution for specific $s \in \mathbb{S}$.

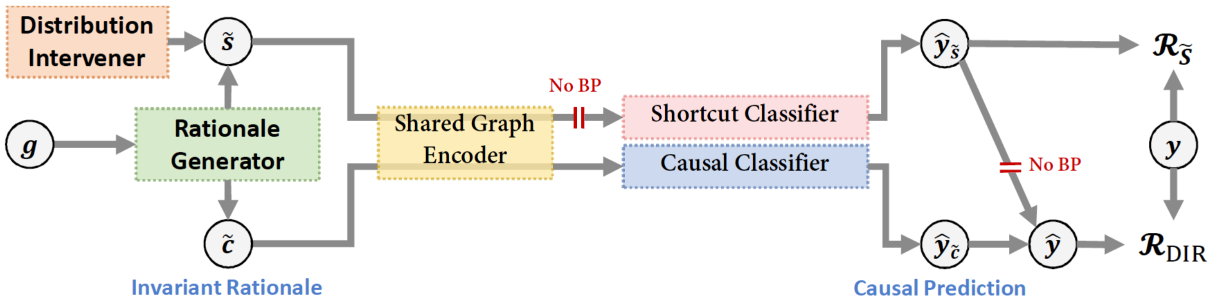


$$\min \mathcal{R}_{\mathrm{DIR}} = \mathbb{E}_s[\mathcal{R}(h(G), Y|do(S=s))] + \lambda \mathrm{Var}_s(\{\mathcal{R}(h(G), Y|do(S=s))\})$$

1. Minimize the risk under all s-interventional distributions
2. Minimize variance of risk over different s-interventional distributions

Discovering Invariant Rationales for Graph Neural Networks. ICLR 2022

RELATED WORKS



Model Architecture

Rationale Generator

Split the input graph instance $g = (\mathcal{V}, \mathcal{E})$ into two subgraphs: causal part $\tilde{c}$ and non-causal part $\tilde{s}$

Distribution Intervener

Collects non-causal part of all instances into a memory bank as $\widetilde{\mathbb{S}}$

Samples memory $\tilde{s}_i \in \widetilde{\mathbb{S}}$ to conduct intervention $do(S = \tilde{s}_i)$, constructing an intervened pair $(\tilde{c}_j, \tilde{s}_i)$

Model Prediction

$$\hat{y} = \hat{y}_{\tilde{c}} \odot \sigma(\hat{y}_{\tilde{s}})$$

Optimization

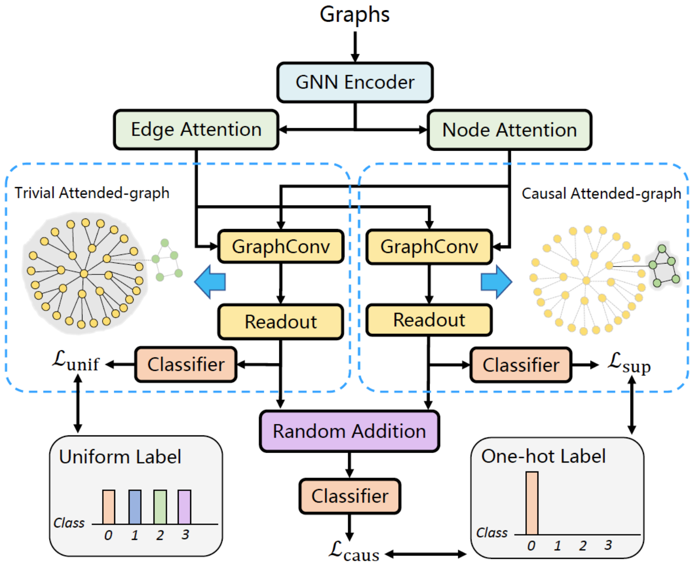
$$\mathcal{R}(h(G), Y | do(S = \tilde{s})) = \mathbb{E}_{(g,y)\in\mathcal{O}, S=\tilde{s}, C=h_{\tilde{C}}(g)} l(\hat{y}, y)$$

$$\mathcal{R}_{\tilde{S}} = \mathbb{E}_{(g,y)\in\mathcal{O}, \tilde{s}=g/h_{\tilde{C}}(g)} l(\hat{y}_{\tilde{s}}, y)$$

Discovering Invariant Rationales for Graph Neural Networks. ICLR 2022

RELATED WORKS

Task: Graph Classification → "How to classify biased graph datasets?"



Model Architecture

Soft Mask Estimation

Separate the causal and shortcut features from the full graphs

Disentanglement

Separate the causal and shortcut features from the full graphs

Causal graph $\mathbf{h}_{\mathcal{G}_c} = f_{\text{readout}}(\text{GConv}_c(\mathbf{A} \odot \mathbf{M}_a, \mathbf{X} \odot \mathbf{M}_x)), \quad \mathbf{z}_{\mathcal{G}_c} = \Phi_c(\mathbf{h}_{\mathcal{G}_c})$

Trivial graph $\mathbf{h}_{\mathcal{G}_t} = f_{\text{readout}}(\text{GConv}_t(\mathbf{A} \odot \overline{\mathbf{M}}_a, \mathbf{X} \odot \overline{\mathbf{M}}_x)), \quad \mathbf{z}_{\mathcal{G}_t} = \Phi_t(\mathbf{h}_{\mathcal{G}_t})$

$$\mathcal{L}_{\text{sup}} = -\frac{1}{|\mathcal{D}|} \sum_{\mathcal{G} \in \mathcal{D}} \mathbf{y}_{\mathcal{G}}^{\top} \log(\mathbf{z}_{\mathcal{G}_c})$$

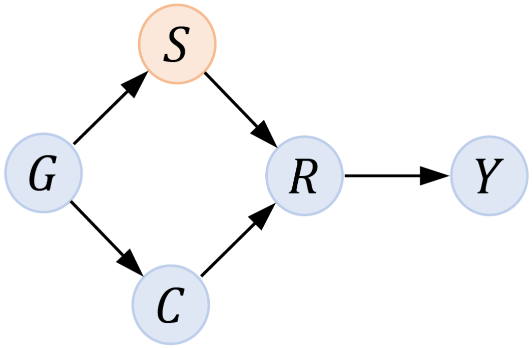
Causal graph → Ground truth label prediction

$$\mathcal{L}_{\text{unif}} = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{G} \in \mathcal{D}} \text{KL}(\mathbf{y}_{\text{unif}}, \mathbf{z}_{\mathcal{G}_t})$$

Trivial graph → Random label prediction

Causal Attention for Interpretable and Generalizable Graph Classification. KDD 2022

RELATED WORKS



G : graph data
C : causal feature
S : shortcut feature
R : representation
Y : prediction

Structure Causal Model (SCM)

$$
\begin{aligned}
P(Y|do(C)) &= P_m(Y|C) \\
&= \sum\nolimits_{s\in\mathcal{T}} P_m(Y|C,s)P_m(s|C) \quad (Bayes\ Rule) \\
&= \sum\nolimits_{s\in\mathcal{T}} P_m(Y|C,s)P_m(s) \quad (Independency) \\
&= \sum\nolimits_{s\in\mathcal{T}} P(Y|C,s)P(s),
\end{aligned}
$$

Confounder Set

Backdoor Adjustment

<u>Causal Intervention via Backdoor adjustment</u>

Challenges

1) Confounder set $\mathcal{T}$ is commonly unobservable and hard to obtain
2) Difficult to directly manipulate graph data (∵Discrete nature)

Let's make implicit intervention on representation level!

$$\mathbf{z}_{\mathcal{G}'} = \Phi(\mathbf{h}_{\mathcal{G}_c} + \mathbf{h}_{\mathcal{G}_{t'}})$$

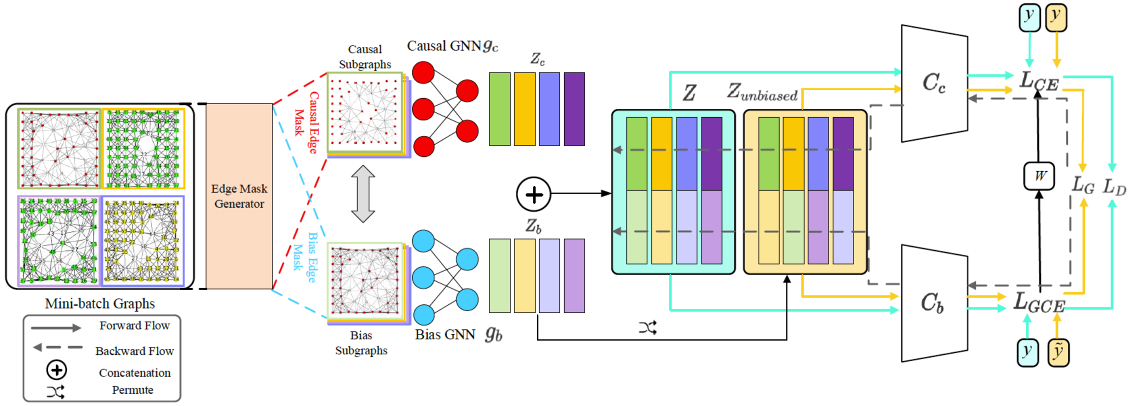
Trivial graph from different graphs

$$\mathcal{L}_{\text{caus}} = -\frac{1}{|\mathcal{D}|\cdot|\hat{\mathcal{T}}|}\sum_{\mathcal{G}\in\mathcal{D}}\sum_{t'\in\hat{\mathcal{T}}} \mathbf{y}_{\mathcal{G}}^{\top}\log\left(\mathbf{z}_{\mathcal{G}'}\right)$$

Causal Attention for Interpretable and Generalizable Graph Classification. KDD 2022

RELATED WORKS

Task: Graph Classification → "How to classify biased graph datasets?"



Model Architecture

Causal and Bias Substructure Generator

Measure the edge importance between node v_i and v_j

$$\alpha_{ij} = \mathrm{MLP}([\mathbf{x}_i, \mathbf{x}_j]) \xrightarrow{\text{Edge in causal subgraph}} c_{ij} = \sigma(\alpha_{ij})$$

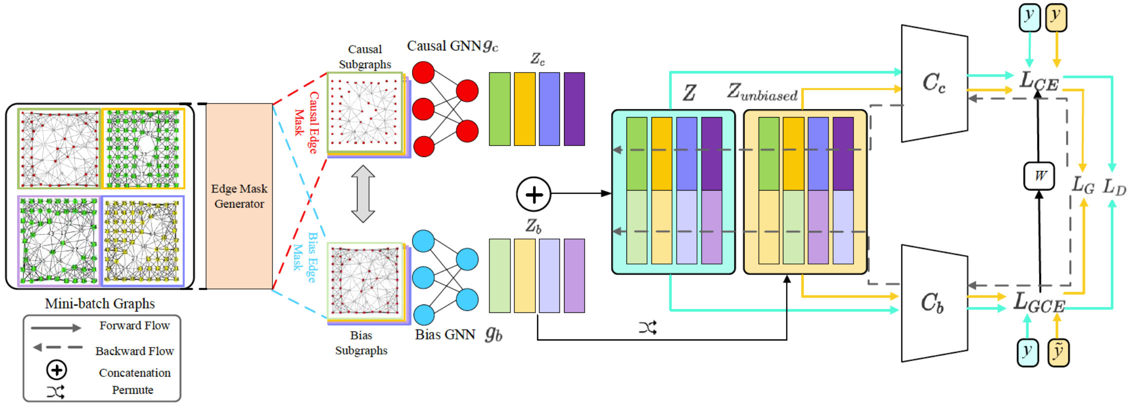
Learning Disentangled Graph Representations

Bias GNN → Generalized cross entropy loss

Causal GNN → Weighted cross entropy loss

Debiasing Graph Neural Networks via Learning Disentangled Causal Substructure. NeurIPS 2022

RELATED WORKS



Model Architecture

Counterfactual Unbiased Sample Generation

How to make causal variable z_c and bias variable z_b uncorrelated?
Swapping z_b with randomly selected different graphs

$$z_{unbiased} = [z_c; \hat{z}_b] \leftarrow \text{From different graphs}$$

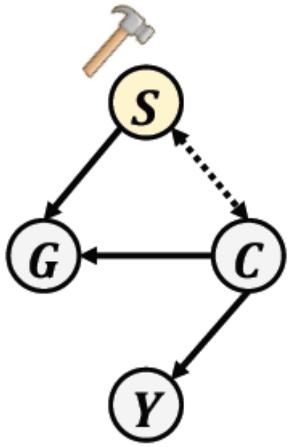
$$L_G = W(z)CE(C_c(z_{unbiased}), y) + GCE(C_b(z_{unbiased}), \hat{y})$$

Can be considered as Backdoor adjustment!

Debiasing Graph Neural Networks via Learning Disentangled Causal Substructure. NeurIPS 2022

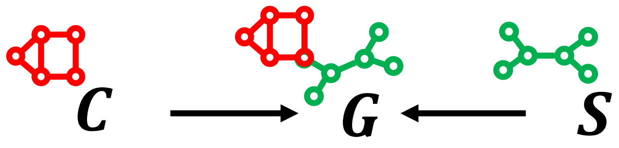
BACKGROUND CAUSAL INFERENCE FOR GRAPH STRUCTURED DATA

Causal view of data-generating process

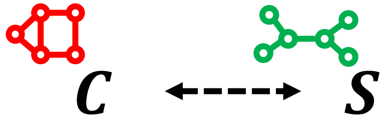


Structure Causal Model (SCM)

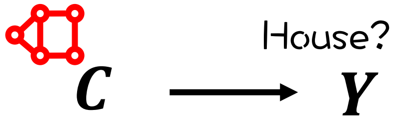
Input graph $\boldsymbol{G}$ consists of two disjoint part:
Causal part $\boldsymbol{C}$ and Non-causal part $\boldsymbol{S}$



Create spurious correlation between $\boldsymbol{S}$ and $\boldsymbol{Y}$



Causal part $\boldsymbol{C}$ only determines target value $\boldsymbol{Y}$



Discovering Invariant Rationales for Graph Neural Networks. ICLR 2022

THEORETICAL ANALYSIS

Training objective of CMRL $-\ell = -\sum_{i=1}^{n} \log q(\mathrm{Y}_i | C_i^1, \mathcal{G}_i^2)$

Expand by multiplying and dividing q

$$
\begin{aligned}
-\ell &= \sum_{i=1}^{n} \log \frac{p(\mathrm{Y}_i | C_i^1, \mathcal{G}_i^2)}{q(\mathrm{Y}_i | C_i^1, \mathcal{G}_i^2)} + \sum_{i=1}^{n} \log \frac{p(\mathrm{Y}_i | \mathcal{G}_i^1, \mathcal{G}_i^2)}{p(\mathrm{Y}_i | C_i^1, \mathcal{G}_i^2)} - \sum_{i=1}^{n} \log p(\mathrm{Y}_i | \mathcal{G}_i^1, \mathcal{G}_i^2) \\
&= \mathbb{E}\left[\log \frac{p(\mathrm{Y} | C^1, \mathcal{G}^2)}{q(\mathrm{Y} | C^1, \mathcal{G}^2)}\right] + \mathbb{E}\left[\log \frac{p(\mathrm{Y} | \mathcal{G}^1, \mathcal{G}^2)}{p(\mathrm{Y} | C^1, \mathcal{G}^2)}\right] - \mathbb{E}\left[\log p(\mathrm{Y} | \mathcal{G}^1, \mathcal{G}^2)\right],
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[\log \frac{p(\mathrm{Y} | \mathcal{G}_i^1, \mathcal{G}_i^2)}{p(\mathrm{Y} | C_i^1, \mathcal{G}_i^2)}\right] &= \mathbb{E}\left[\log \frac{p(\mathrm{Y} | C_i^1, \mathcal{S}_i^1, \mathcal{G}_i^2)}{p(\mathrm{Y} | C_i^1, \mathcal{G}_i^2)}\right] \\
&= \sum_{i=1}^{n} p(\mathcal{G}_i^1, \mathcal{G}_i^2, \mathrm{Y}_i) \log \frac{p(\mathrm{Y}_i | C_i^1, \mathcal{S}_i^1, \mathcal{G}_i^2)}{p(\mathrm{Y}_i | C_i^1, \mathcal{G}_i^2)} \\
&= \sum_{i=1}^{n} p(\mathcal{G}_i^1, \mathcal{G}_i^2, \mathrm{Y}_i) \log \frac{p(\mathrm{Y}_i | C_i^1, \mathcal{S}_i^1, \mathcal{G}_i^2)}{p(\mathrm{Y}_i | C_i^1, \mathcal{G}_i^2)} \frac{p(\mathcal{S}_i^1 | C_i^1, \mathcal{G}_i^2)}{p(\mathcal{S}_i^1 | C_i^1, \mathcal{G}_i^2)} \\
&= \sum_{i=1}^{n} p(\mathcal{G}_i^1, \mathcal{G}_i^2, \mathrm{Y}_i) \log \frac{p(\mathcal{S}_i^1, \mathrm{Y}_i | C_i^1, \mathcal{G}_i^2)}{p(\mathrm{Y}_i | C_i^1, \mathcal{G}_i^2) \cdot p(\mathcal{S}_i^1 | C_i^1, \mathcal{G}_i^2)} \\
&= I(\mathcal{S}^1; \mathrm{Y} | C^1, \mathcal{G}^2)
\end{aligned}
$$

$$
\min \mathbb{E}\left[\log \frac{p(\mathrm{Y} | C^1, \mathcal{G}^2)}{q(\mathrm{Y} | C^1, \mathcal{G}^2)}\right] + I(\mathcal{S}^1; \mathrm{Y} | C^1, \mathcal{G}^2) + H(\mathrm{Y} | \mathcal{G}^1, \mathcal{G}^2)
$$

1. Likelihood ratio between true distribution and predicted distribution
2. Conditional Mutual Information
3. Irreducible constant inherent in the datasets

We can explain the behavior of CMRL in two perspective

THEORETICAL ANALYSIS

$$\min \mathbb{E}\left[\log \frac{p(\mathrm{Y}|C^1,\mathcal{G}^2)}{q(\mathrm{Y}|C^1,\mathcal{G}^2)}\right] + I(\mathcal{S}^1;\mathrm{Y}|C^1,\mathcal{G}^2) + H(\mathrm{Y}|\mathcal{G}^1,\mathcal{G}^2)$$

<u>Perspective 1. CMRL learns informative causal substructure</u>

Minimize $I(\mathcal{S}^1;\mathrm{Y}|C^1,\mathcal{G}^2)$

Disentangle the shortcut substructure $\boldsymbol{S}^{\mathbf{1}}$ that are no longer needed in predicting the label $\boldsymbol{Y}$ when the context $\boldsymbol{\mathcal{C}}^{\mathbf{1}}$ and $\boldsymbol{\mathcal{G}}^{\mathbf{2}}$ given.

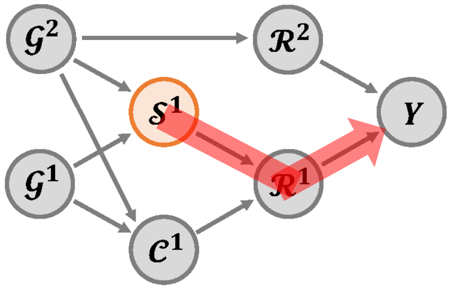
Chain rule of MI $I(\mathcal{S}^1;\mathrm{Y}|C^1,\mathcal{G}^2) = I(\mathcal{G}^1,\mathcal{G}^2;\mathrm{Y}) - I(C^1,\mathcal{G}^2;\mathrm{Y})$

Encourages the causal substructure $\boldsymbol{\mathcal{C}}^{\mathbf{1}}$ and paired molecule $\boldsymbol{\mathcal{G}}^{\mathbf{2}}$ to contain enough information on target $\boldsymbol{Y}$.

THEORETICAL ANALYSIS

$$\min \mathbb{E}\left[\log \frac{p(\mathrm{Y}|C^1,\mathcal{G}^2)}{q(\mathrm{Y}|C^1,\mathcal{G}^2)}\right] + I(\mathcal{S}^1;\mathrm{Y}|C^1,\mathcal{G}^2) + H(\mathrm{Y}|\mathcal{G}^1,\mathcal{G}^2)$$

Perspective 2. CMRL reduces model bias with causal view



$\mathcal{G}^1$: Molecule 1
$\mathcal{G}^2$: Molecule 2
$\mathcal{C}^1$: Causal Substructure in Molecule 1
$\mathcal{S}^1$: Shortcut Substructure in Molecule 1
$\mathcal{R}^1$: Molecule 1 Representation
$\mathcal{R}^2$: Molecule 2 Representation
Y : Target Value

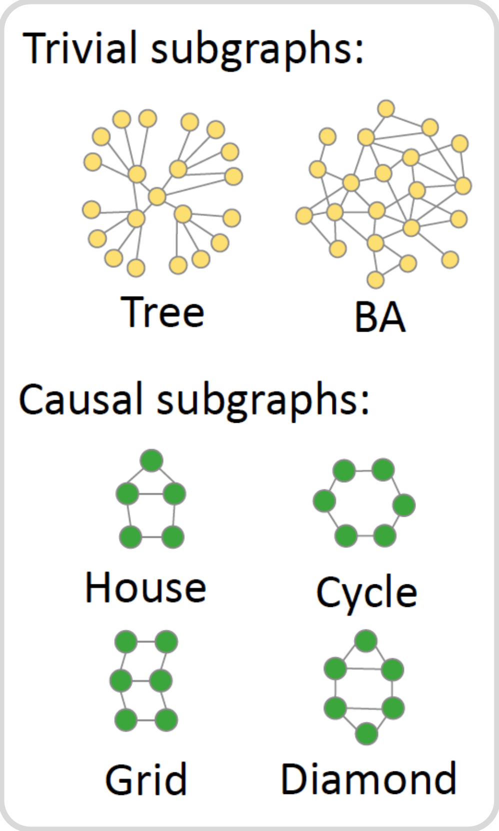
➡ Model bias

Based on information leakage,
Model bias can be quantified based on mutual information

Again, several backdoor paths are blocked by conditioning on $\mathcal{C}^1$ and $\mathcal{G}^2$
→ Enable the direct measure of model bias!
→ Finally, Loss term minimize the model bias

EXPERIMENTS SYNTHETIC DATASET EXPERIMENTS

In synthetic dataset experiment, we assess the model's performance on various levels of bias in datasets



Positive pair
a pair that shares the same causal substructure
{House, House} → Positive

Negative pair
a pair that each graph has a different causal substructure
{House, Cycle} → Negative

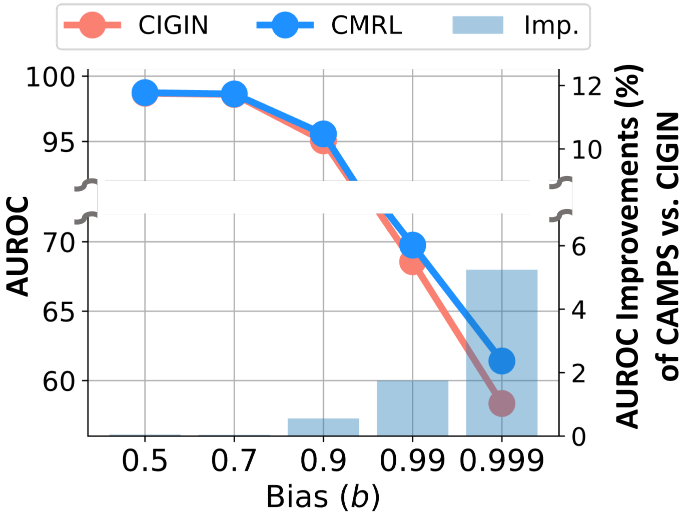
Dataset bias
the ratio of the positive pairs containing "BA." shortcut substructures

$$\begin{aligned}\text{bias}(b) &= \frac{\text{Number of positive pairs with BA substructure}}{\text{Number of positive pairs}}\\ &= \frac{\#\{\text{Causal-BA, Causal-BA}\}}{\#\{\text{Causal-Tree, Causal-Tree}\} + \#\{\text{Causal-BA, Causal-BA}\}}\end{aligned}$$

Bias level b increases
→ "BA." substructures dominates model prediction

EXPERIMENTS SYNTHETIC DATASET EXPERIMENTS

In synthetic dataset experiment, we assess the model's performance on various levels of bias in datasets



Observations

1. Models' performance degrades as the bias gets severe
→ "BA." shortcut confound the model

2. Performance gap between CMRL and CIGIN gets larger as the bias gets severe
→ Importance of learning causality between the substructure and target