Conditional Graph Information Bottleneck for Molecular Relational Learning

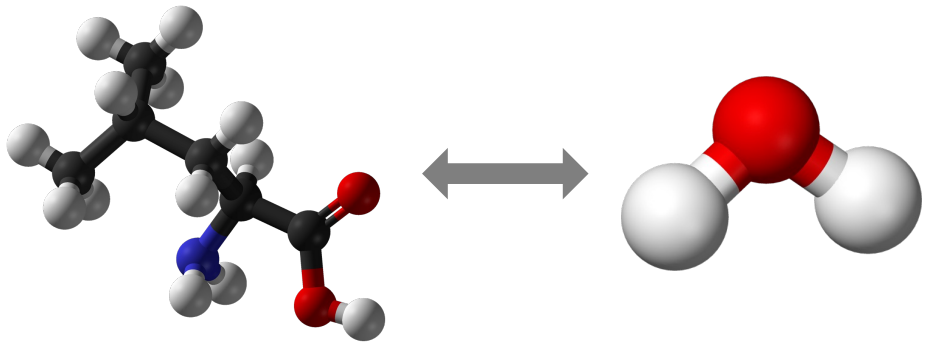
Namkyeong Lee, Dongmin Hyun, Gyoung S. Na,
Sungwon Kim, Junseok Lee, Chanyoung Park

KAIST DSAIL Data Science & Artificial Intelligence POSTECH KRICT ICML International Conference On Machine Learning

TABLE OF CONTENTS

- **Background**
  - Molecular Relational Learning
  - Functional Group
  - Information Bottleneck
- **Motivation**
- **Conditional Graph Information Bottleneck**
- **Experiments**
- **Conclusion**

BACKGROUND MOLECULAR RELATIONAL LEARNING



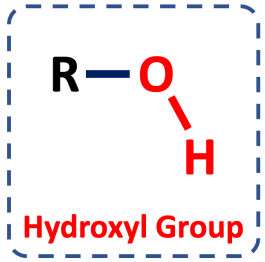
Molecular Relational Learning

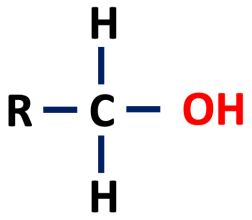
Learning the interaction behavior between a pair of molecules

Examples

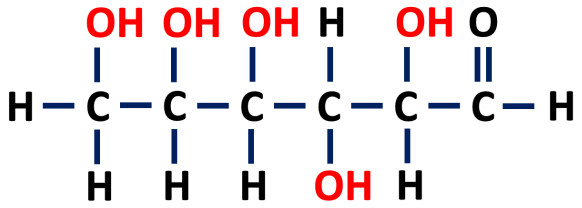
- Predicting optical properties when a **Chromophore** and **Solvent** react
- Predicting solubility when a **solute** and **solvent** react
- Predicting side effects when taking **two types of drugs** simultaneously

BACKGROUND FUNCTIONAL GROUP





Alcohol



Glucose

Functional Group

Specific atomic groups that play an important role in determining the chemical reactivity of organic compounds

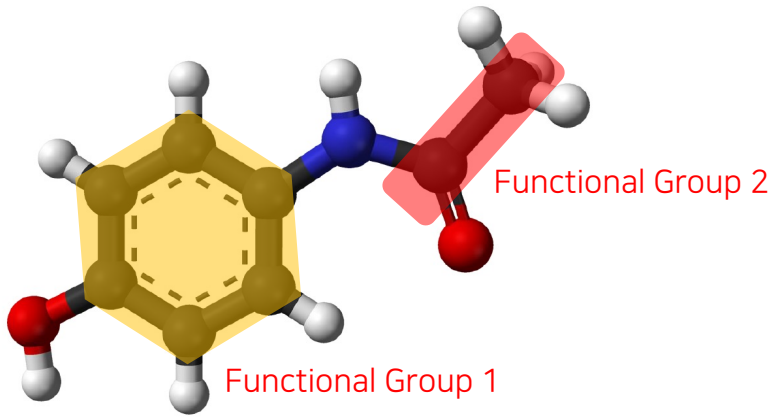
Compounds with the same functional group generally have similar properties and undergo similar chemical reactions

Examples

The hydroxyl group structure has the characteristic of increasing the pol arity of the molecule

Hence, it is important to consider functional group for molecular relational learning

BACKGROUND FUNCTIONAL GROUP



Functional Group

Specific atomic groups that play an important role in determining the chemical reactivity of organic compounds

Compounds with the same functional group generally have similar properties and undergo similar chemical reactions

Molecule can be represented as a graph
Functional group can be represented as a subgraph

Recently, information theory-based approaches have been proposed to detect important subgraph

BACKGROUND INFORMATION BOTTLENECK

Definition 2.1. (Information Bottleneck) Given random variables X and Y, the Information Bottleneck principle aims to compress X to a bottleneck random variable T, while keeping the information relevant for predicting Y:

$$\min_T -I(Y;T) + \beta I(X;T) \tag{2}$$

where β is a Lagrangian multiplier for balancing the two mutual information terms.

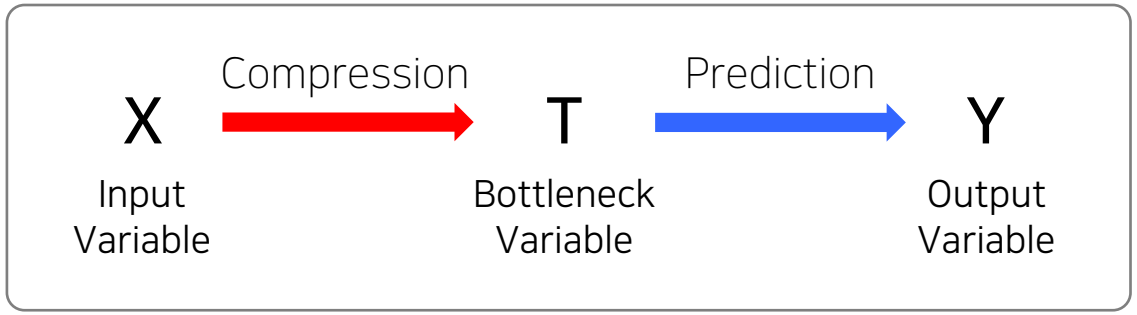
Information Bottleneck Theory

A theoretical approach to trade-off between information compression and preservation

Minimize MI between X and T
→ T should contain minimal information about X
→ **Compression**

$$\min_T -I(Y;T) + \beta I(X;T)$$

Maximize MI between T and Y
→ T should contain as much information about Y as possible
→ **Prediction**



BACKGROUND GRAPH INFORMATION BOTTLENECK

Definition 2.2. (IB-Graph) For a graph $\mathcal{G} = (\mathbf{X}, \mathbf{A})$ and its label information $\mathbf{Y}$, the optimal graph $\mathcal{G}_{\mathrm{IB}} = (\mathbf{X}_{\mathrm{IB}}, \mathbf{A}_{\mathrm{IB}})$ discovered under the IB principle is denoted as IB-Graph:

$$\mathcal{G}_{\mathrm{IB}} = \arg\min_{\mathcal{G}_{\mathrm{IB}}} -I(\mathbf{Y}; \mathcal{G}_{\mathrm{IB}}) + \beta I(\mathcal{G}; \mathcal{G}_{\mathrm{IB}}) \tag{3}$$

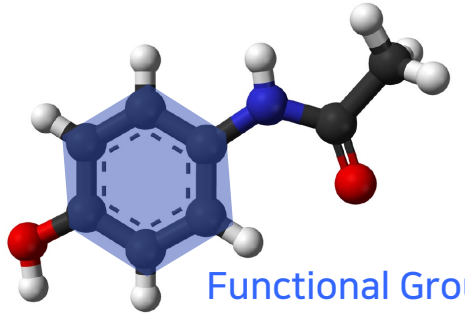
where $\mathbf{X}_{\mathrm{IB}}$ and $\mathbf{A}_{\mathrm{IB}}$ denote the task-relevant feature set and the adjacency matrix of $\mathcal{G}$, respectively.

Information Bottleneck Graph

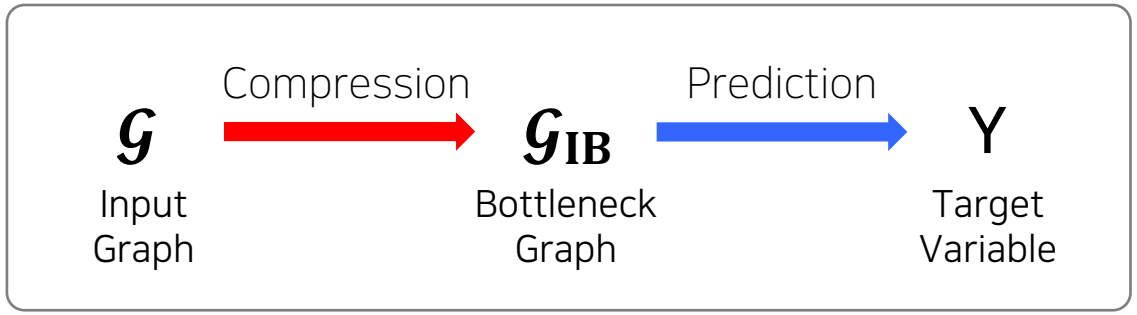
Subgraph that maximally preserves the property of the original graph

→ Motif in ordinary graphs

→ Functional group in molecules



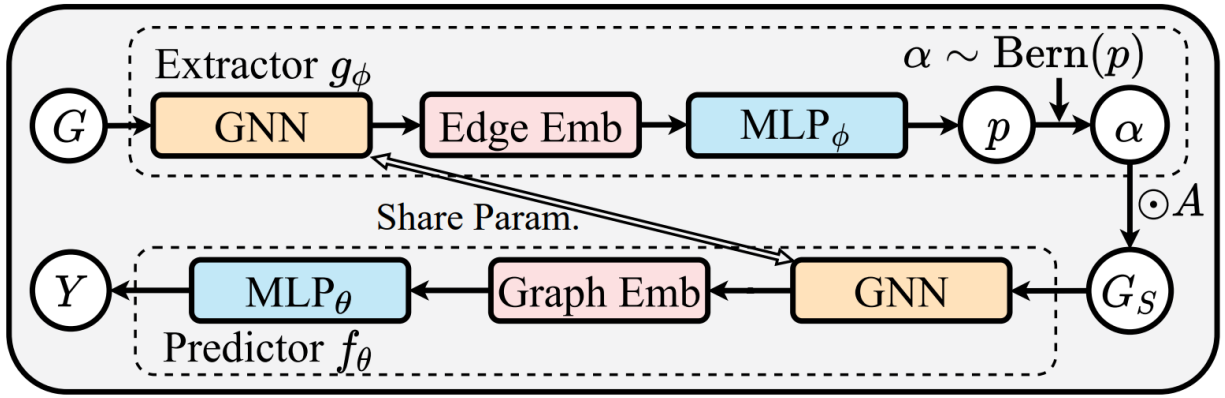
Functional Group of Molecule



BACKGROUND GRAPH INFORMATION BOTTLENECK

Extract a subgraph in terms of edges

Model an edge based Bernoulli distribution to perform graph compression



Model objective

$$\min_{\phi} -I(G_S;Y) + \beta I(G_S;G), \text{ s.t. } G_S \sim g_\phi(G).$$

Proposed variational upperbound

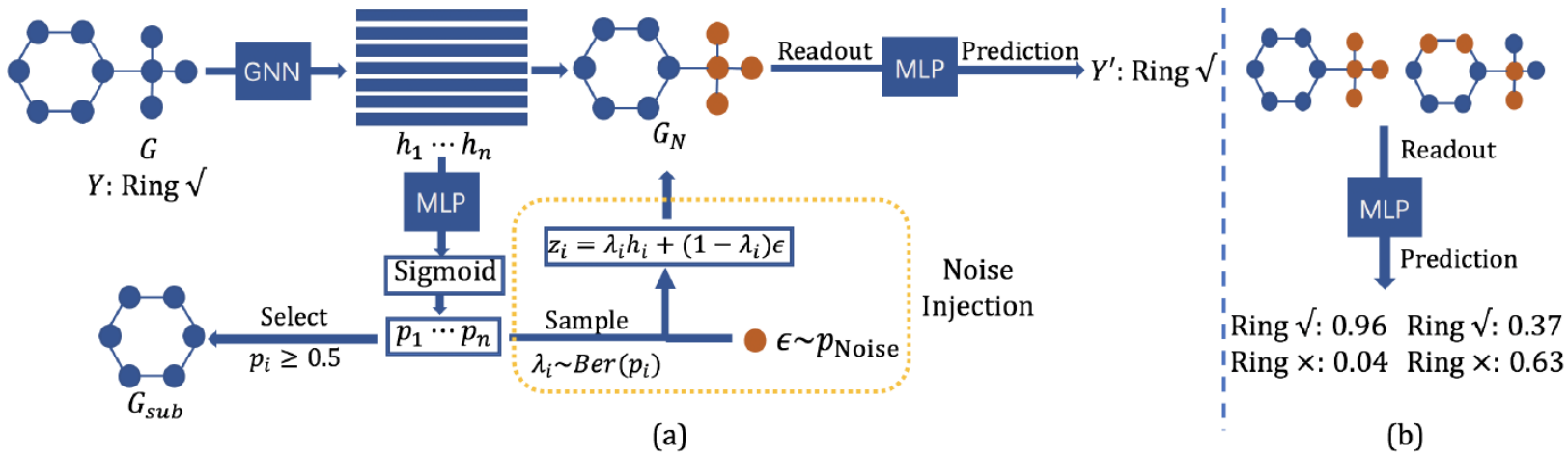
$$\min_{\theta,\phi} -\mathbb{E}\left[\log \mathbb{P}_\theta(Y|G_S)\right] + \beta \mathbb{E}\left[\mathrm{KL}(\mathbb{P}_\phi(G_S|G)||\mathbb{Q}(G_S))\right],$$
$$\text{s.t.} \quad G_S \sim \mathbb{P}_\phi(G_S|G). \tag{8}$$

Miao, Siqi, Mia Liu, and Pan Li. "Interpretable and generalizable graph learning via stochastic attention mechanism." ICML 2022

BACKGROUND GRAPH INFORMATION BOTTLENECK

Extract a subgraph in terms of nodes

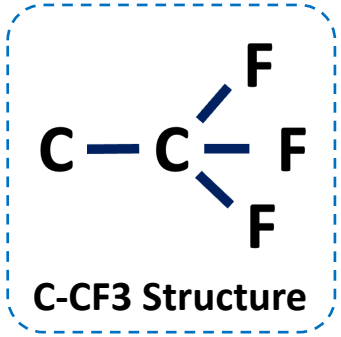
Inject noise into node embeddings to perform graph compression



Can Information bottleneck theory also benefits molecular relational learning?

Yu, Junchi, Jie Cao, and Ran He. "Improving subgraph recognition with variational graph information bottleneck." CVPR 2022.

MOTIVATION



C-CF3 Structure


Water

Decrease Solubility


Oil

Unknown

Functional Group

Specific atomic groups that play an important role in determining the chemical reactivity of organic compounds

Compounds with the same functional group generally have similar properties and undergo similar chemical reactions

On the other hand, the role of functional group varies depending on which solvent the solute reacts with!

Examples: C-CF3 structure in molecules

It is important to consider the paired solvent when detecting the important substructure from solute
→ Existing approaches for information bottleneck cannot capture such a prior knowledge

METHODOLOGY CONDITIONAL GRAPH INFORMATION BOTTLENECK

Conditional Graph Information Bottleneck

Consider Graph 2 (Solvent) when detecting the important subgraph from Graph 1 (Solute)

$$\mathcal{G}_{\mathrm{IB}} = \arg\min_{\mathcal{G}_{\mathrm{IB}}} -I(\mathbf{Y}; \mathcal{G}_{\mathrm{IB}}) + \beta I(\mathcal{G}; \mathcal{G}_{\mathrm{IB}})$$

Graph Information Bottleneck

$$\mathcal{G}^1_{\mathrm{CIB}} = \arg\min_{\mathcal{G}^1_{\mathrm{CIB}}} -I(\mathbf{Y}; \mathcal{G}^1_{\mathrm{CIB}} | \mathcal{G}^2) + \beta I(\mathcal{G}^1; \mathcal{G}^1_{\mathrm{CIB}} | \mathcal{G}^2)$$

Conditional Graph Information Bottleneck

METHODOLOGY CONDITIONAL GRAPH INFORMATION BOTTLENECK

LEMMA 3.3. *(Noise Invariance) Given a pair of graphs* $(\mathcal{G}^1, \mathcal{G}^2)$ *and its label information* Y, *let* $\mathcal{G}_n^1$ *be a task irrelevant noise in the input graph* $\mathcal{G}^1$. *Then, the following inequality holds:*

$$I(\mathcal{G}_{\text{CIB}}^1; \mathcal{G}_n^1 | \mathcal{G}^2) \leq -I(\text{Y}; \mathcal{G}_{\text{CIB}}^1 | \mathcal{G}^2) + I(\mathcal{G}^1; \mathcal{G}_{\text{CIB}}^1 | \mathcal{G}^2) \quad (6)$$

Proof of Lemma 3.3

Assuming that $\mathcal{G}^1$, $\mathcal{G}_{\text{CIB}}^1$, $\mathcal{G}_n^1$, $\mathcal{G}^2$, and Y satisfy the Markov condition $(\text{Y}, \mathcal{G}_n^1, \mathcal{G}^2) \rightarrow \mathcal{G}^1 \rightarrow \mathcal{G}_{\text{CIB}}^1$, we have the following inequality due to data processing inequality:

$$\begin{aligned} I(\mathcal{G}^1; \mathcal{G}_{\text{CIB}}^1 | \mathcal{G}^2) &= I(\mathcal{G}_{\text{CIB}}^1; \mathcal{G}^1, \mathcal{G}^2) - I(\mathcal{G}_{\text{CIB}}^1; \mathcal{G}^2) \\ &\geq I(\mathcal{G}_{\text{CIB}}^1; \text{Y}, \mathcal{G}_n^1, \mathcal{G}^2) - I(\mathcal{G}_{\text{CIB}}^1; \mathcal{G}^2) \\ &= I(\mathcal{G}_{\text{CIB}}^1; \mathcal{G}_n^1, \mathcal{G}^2) + I(\mathcal{G}_{\text{CIB}}^1; \text{Y} | \mathcal{G}_n^1, \mathcal{G}^2) - I(\mathcal{G}_{\text{CIB}}^1; \mathcal{G}^2) \\ &= I(\mathcal{G}_{\text{CIB}}^1; \mathcal{G}_n^1, \mathcal{G}^2) + I(\mathcal{G}_{\text{CIB}}^1; \text{Y} | \mathcal{G}_n^1, \mathcal{G}^2) \quad (1) \end{aligned}$$

Suppose that $\mathcal{G}_n^1$ and Y, $\mathcal{G}_n^1$ and $\mathcal{G}^2$, and joint random variable $(\mathcal{G}_n^1, \mathcal{G}^2)$ and Y are independent respectively. Then, for $I(\mathcal{G}_{\text{CIB}}^1; \text{Y} | \mathcal{G}_n^1, \mathcal{G}^2)$ we have:

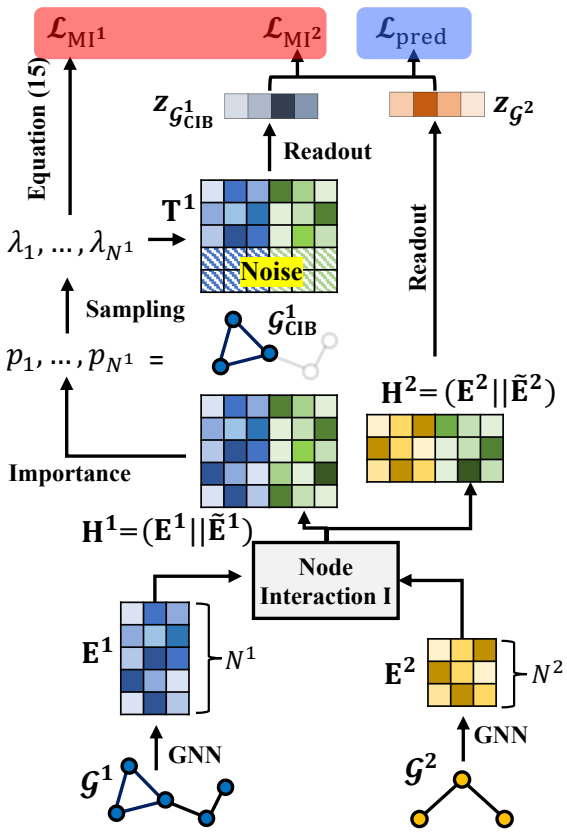
$$\begin{aligned} I(\mathcal{G}_{\text{CIB}}^1; \text{Y} | \mathcal{G}_n^1, \mathcal{G}^2) &= H(Y | \mathcal{G}_n^1, \mathcal{G}^2) - H(Y | \mathcal{G}_n^1, \mathcal{G}_{\text{CIB}}^1, \mathcal{G}^2) \\ &\geq H(Y | \mathcal{G}^2) - H(Y | \mathcal{G}_{\text{CIB}}^1, \mathcal{G}^2) \\ &= I(Y; \mathcal{G}_{\text{CIB}}^1 | \mathcal{G}^2) \quad (2) \end{aligned}$$

By plugging Equation (2) into Equation (1), we have:

$$I(\mathcal{G}^1; \mathcal{G}_{\text{CIB}}^1 | \mathcal{G}^2) \geq I(\mathcal{G}_{\text{CIB}}^1; \mathcal{G}_n^1, \mathcal{G}^2) + I(Y; \mathcal{G}_{\text{CIB}}^1 | \mathcal{G}^2)$$

By minimizing CGIB objective function,
the model learns a CIB-Graph with the smallest mutual information with task-irrelevant noise

METHODOLOGY CONDITIONAL GRAPH INFORMATION BOTTLENECK



$$\min -I(\mathrm{Y}; \mathcal{G}^1_{\text{CIB}} | \mathcal{G}^2) + \beta I(\mathcal{G}^1; \mathcal{G}^1_{\text{CIB}} | \mathcal{G}^2)$$

Overall procedure

Decompose the conditional MI based on the chain rule of MI, and then derive the upper bound of the decomposed terms

$$-I(\mathrm{Y}; \mathcal{G}^1_{\text{CIB}} | \mathcal{G}^2) = -I(\mathrm{Y}; \mathcal{G}^1_{\text{CIB}}, \mathcal{G}^2) + I(\mathrm{Y}; \mathcal{G}^2)$$

$$-I(\mathrm{Y}; \mathcal{G}^1_{\text{CIB}}, \mathcal{G}^2) \le \mathbb{E}_{\mathcal{G}^1_{\text{CIB}}, \mathcal{G}^2, \mathrm{Y}}[-\log p_\theta(\mathrm{Y} | \mathcal{G}^1_{\text{CIB}}, \mathcal{G}^2)]$$

Prediction Loss

$$I(\mathcal{G}^1; \mathcal{G}^1_{\text{CIB}} | \mathcal{G}^2) = I(\mathcal{G}^1_{\text{CIB}}; \mathcal{G}^1, \mathcal{G}^2) - I(\mathcal{G}^1_{\text{CIB}}; \mathcal{G}^2)$$

$$I(\mathcal{G}^1_{\text{CIB}}; \mathcal{G}^1, \mathcal{G}^2) \le \mathbb{E}_{\mathcal{G}^1, \mathcal{G}^2}[-\frac{1}{2}\log A + \frac{1}{2N^1}A + \frac{1}{2N^1}B^2]$$

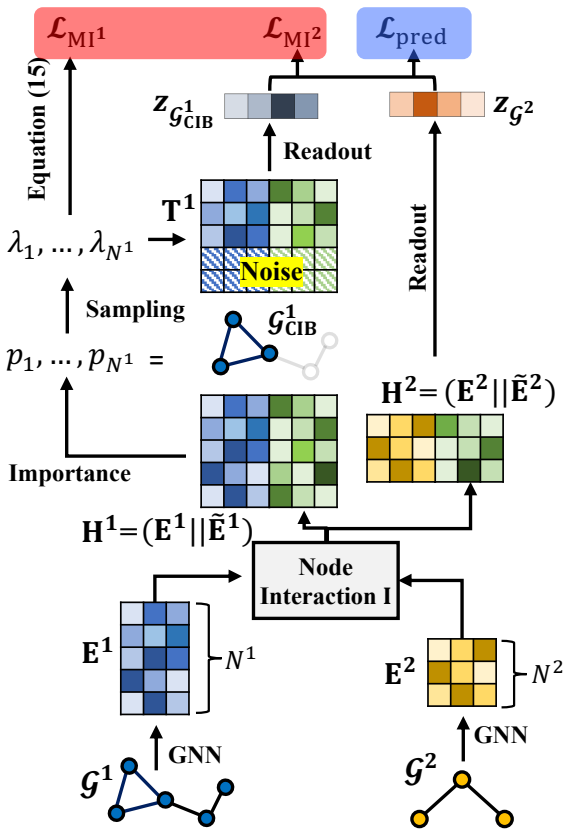
$$:= \mathcal{L}_{\text{MI}^1}(\mathcal{G}^1_{\text{CIB}}, \mathcal{G}^1, \mathcal{G}^2)$$

$$-I(\mathcal{G}^1_{\text{CIB}}; \mathcal{G}^2) \le \mathbb{E}_{\mathcal{G}^1_{\text{CIB}}, \mathcal{G}^2}[-\log p_\xi(\mathcal{G}^2 | \mathcal{G}^1_{\text{CIB}})]$$

$$:= \mathcal{L}_{\text{MI}^2}(\mathcal{G}^1_{\text{CIB}}, \mathcal{G}^2)$$

Compression Loss

METHODOLOGY CONDITIONAL GRAPH INFORMATION BOTTLENECK



$$-I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2) = -I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2) + I(\mathrm{Y};\mathcal{G}^2) \qquad \because \text{Chain rule of mutual information}$$

Direct calculation of mutual information is intractable;
Instead, we minimize the upper bound

Proposition. (Upper bound of $-\mathrm{I}(Y;\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$) Given a pair of graph $(\mathcal{G}^1,\mathcal{G}^2)$, its label information Y, and the learned CIB-graph $\mathcal{G}^1_{\mathrm{CIB}}$, we have:

$$-\mathrm{I}(Y;\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2) \le \mathbb{E}_{Y,\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2}\left[\log p_\theta\left(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2\right)\right]$$

where $p_\theta(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$ is variational approximation of $p(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$.

Proof. By the definition of mutual information and introducing variational approximation $p_\theta(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$ of intractable distribution $p(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$, we have:

$$I\left(Y;\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2\right) = \mathbb{E}_{Y,\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2}\left[\log\frac{p\left(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2\right)}{p(Y)}\right]$$

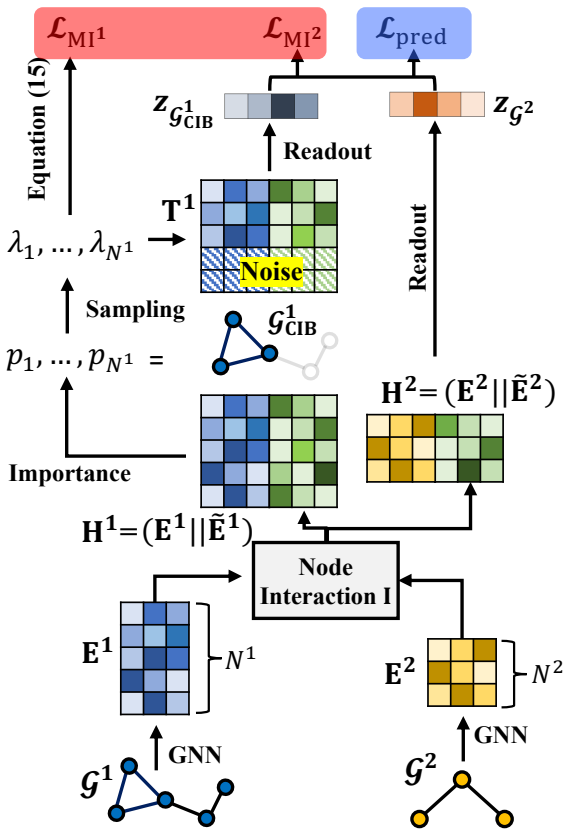
$$= \mathbb{E}_{Y,\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2}\left[\log\frac{p_\theta\left(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2\right)}{p(Y)}\right] + \mathbb{E}_{\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2}\left[p\left(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2\right)||p_\theta\left(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2\right)\right]$$

$$\ge \mathbb{E}_{Y,\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2}\left[\log\frac{p_\theta\left(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2\right)}{p(Y)}\right] \qquad \because \text{Non-negativity of KL divergence}$$

$$= \mathbb{E}_{Y,\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2}\left[\log p_\theta\left(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2\right)\right] + H(Y)$$

$$\min -I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)+\beta\, I(\mathcal{G}^1;\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)$$

METHODOLOGY CONDITIONAL GRAPH INFORMATION BOTTLENECK



$$-I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2) = -I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2) + I(\mathrm{Y};\mathcal{G}^2) \qquad \because \text{Chain rule of mutual information}$$

Direct calculation of mutual information is intractable;
Instead, we minimize the upper bound

Proposition. (Upper bound of $-\mathrm{I}(Y;\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$) Given a pair of graph $(\mathcal{G}^1,\mathcal{G}^2)$, its label information Y, and the learned CIB-graph $\mathcal{G}^1_{\mathrm{CIB}}$, we have:

$$-\mathrm{I}(Y;\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2) \le \mathbb{E}_{Y,\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2}\left[\log p_\theta\left(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2\right)\right]$$

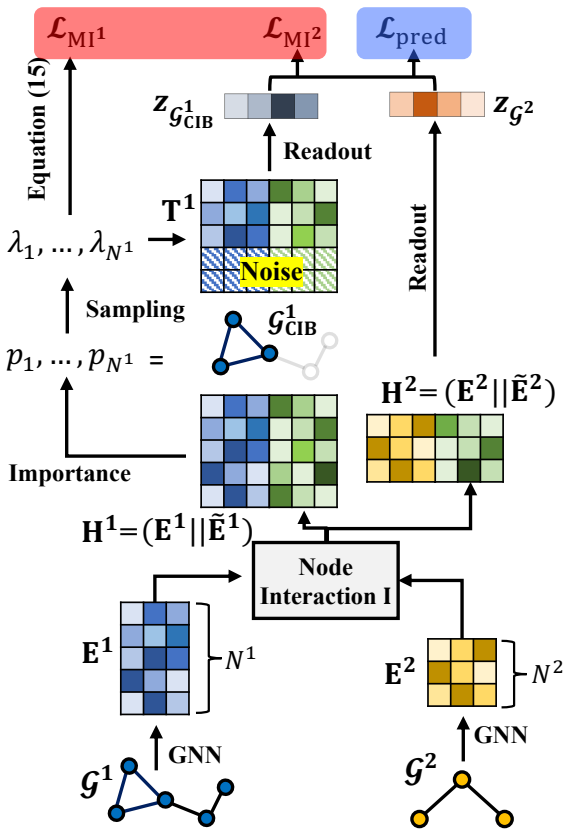
where $p_\theta(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$ is variational approximation of $p(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$.

Implementation

- Consider $p_\theta(Y|\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$ as a predictor parameterezied by θ, which outputs the model prediction Y based on the input pair $(\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$
- The upper bound is minimized by minimizing the prediction loss $\mathcal{L}_{\mathrm{pred}}(\mathrm{Y},\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$

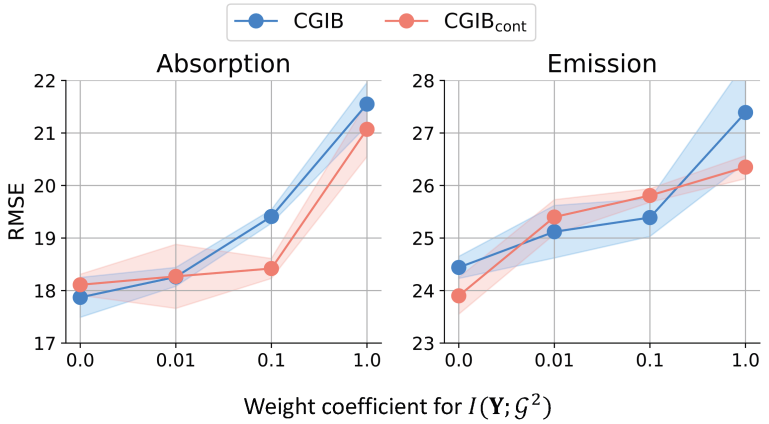
$$\min -I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)+\beta\, I(\mathcal{G}^1;\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)$$

METHODOLOGY CONDITIONAL GRAPH INFORMATION BOTTLENECK



$$-I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2) = -I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2) + I(\mathrm{Y};\mathcal{G}^2) \qquad \because \text{ Chain rule of mutual information}$$

The 2nd term is empirically found to be not helpful



Following VIB, we treat $r(\mathrm{Y})$ as fixed spherical Gaussian,

$I(Y;\mathcal{G}^2) \leq \mathbb{E}_{\mathcal{G}^2}\left[KL(p_\xi(Y|\mathcal{G}^2)||r(Y))\right]$

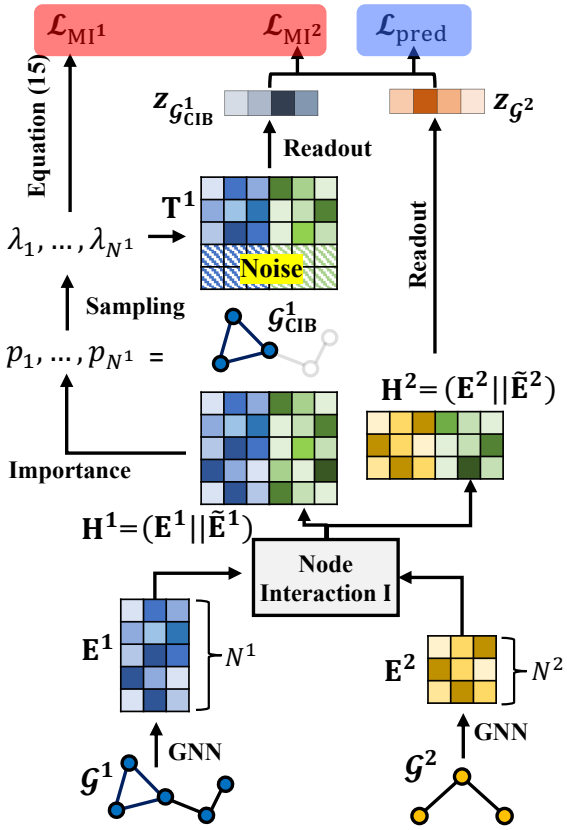
where $r(Y)\sim N(\mathrm{Y}|0,1)$

Increasing the contribution of this term deteriorates the model performance

We remove the term $I(\mathrm{Y},\mathcal{G}^2)$ from the model

$$\min -I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)+\beta\, I(\mathcal{G}^1;\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)$$

METHODOLOGY CONDITIONAL GRAPH INFORMATION BOTTLENECK



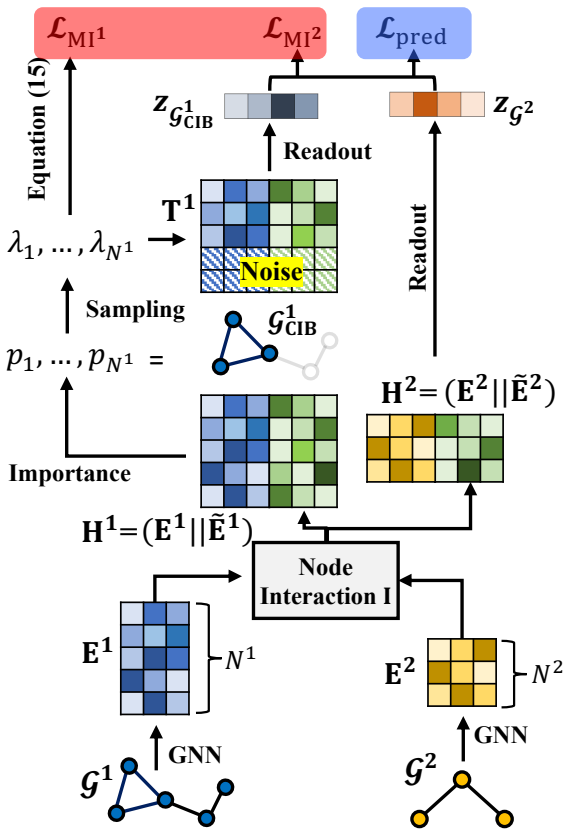
$$I(\mathcal{G}^1;\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2) = \underbrace{I(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^1,\mathcal{G}^2)}_{\mathcal{L}_{\mathrm{MI}^1}} - \underbrace{I(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^2)}_{\mathcal{L}_{\mathrm{MI}^2}} \quad \because \text{Chain rule of mutual information}$$

$\mathcal{L}_{\mathrm{MI}^1}$: **Compression through Noise injection**
→ Injecting noise into unimportant nodes

$\mathcal{L}_{\mathrm{MI}^2}$: **Solute Prediction**
→ Encourage $\mathcal{G}^1_{\mathrm{CIB}}$ to contain as much information about $\mathcal{G}^2$ as possible
→ The term that arises from the Conditional Mutual Information
→ Key to the success of CGIB! Enables the conditional information compression of CGIB

$$\min -I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)+\beta\, I(\mathcal{G}^1;\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)$$

METHODOLOGY CONDITIONAL GRAPH INFORMATION BOTTLENECK



$$I(\mathcal{G}^1;\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)=\underbrace{I(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^1,\mathcal{G}^2)}_{\mathcal{L}_{\mathrm{MI}^1}}-I(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^2) \quad \because \text{Chain rule of mutual information}$$

Compression through Noise Injection

* Injecting noise into unimportant nodes

$\mathrm{H}^1_{\mathrm{i}}$: Representation of node i of $\mathcal{G}^1$ that contains information about both $\mathcal{G}^1, \mathcal{G}^2$

$p_i = \mathrm{MLP}(\mathrm{H}^1_{\mathrm{i}})$: Important of node i of $\mathcal{G}^1$

$\mathrm{T}^1_{\mathrm{i}} = \lambda_{\mathrm{i}}\mathrm{H}^1_{\mathrm{i}} + (1-\lambda_{\mathrm{i}})\varepsilon$ where $\lambda_{\mathrm{i}} \sim \mathrm{Bernoulli}(\mathrm{p_i})$ and $\varepsilon \sim \mathrm{N}(\mu_{\mathrm{H}^1}, \sigma^2_{\mathrm{H}^1})$

Intuition) Unimportant nodes would not affect the model performance even if they are replaced with noise

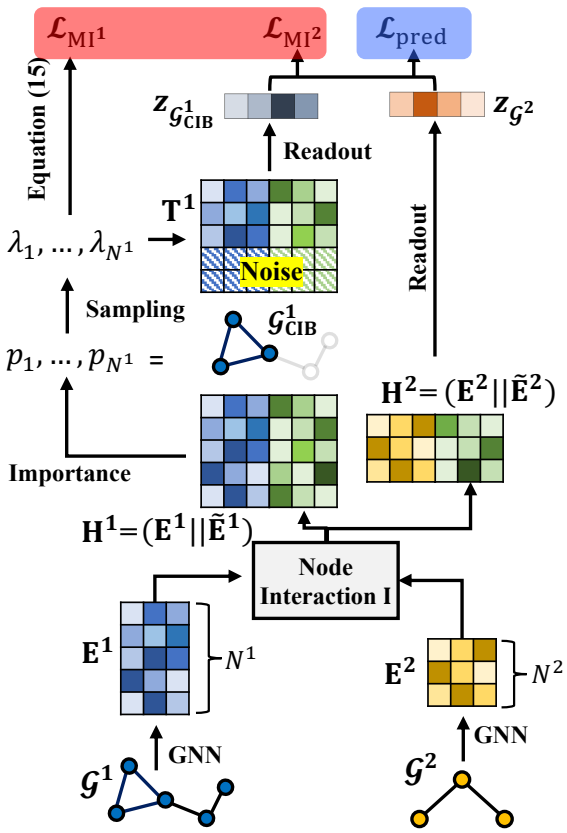
Upper bound of $\mathrm{I}(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^1,\mathcal{G}^2)$

$$\mathrm{I}(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^1,\mathcal{G}^2) \le \mathbb{E}_{\mathcal{G}^1,\mathcal{G}^2}\left[-\frac{1}{2}\log A+\frac{1}{2N^1}A+\frac{1}{2N^1}B^2\right] \quad \text{where } A=\sum_{j=1}^{N^1}(1-\lambda_j)^2 \text{ and } B=\frac{\sum_{j=1}^{N^1}\lambda_j(\mathrm{H}^1_j-\mu_{\mathrm{H}^1})^2}{\sigma_{\mathrm{H}^1}}$$

$$:= \mathcal{L}_{MI^1}(\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^1,\mathcal{G}^2)$$

$$\min -I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)+\beta I(\mathcal{G}^1;\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)$$

METHODOLOGY CONDITIONAL GRAPH INFORMATION BOTTLENECK



Upper bound of $\mathrm{I}(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^1,\mathcal{G}^2)$

$$\mathrm{I}(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^1,\mathcal{G}^2) \le \mathbb{E}_{\mathcal{G}^1,\mathcal{G}^2}\left[-\frac{1}{2}\log A+\frac{1}{2N^1}A+\frac{1}{2N^1}B^2\right] \quad \text{where } A=\sum_{j=1}^{N^1}(1-\lambda_j)^2 \text{ and } B=\frac{\sum_{j=1}^{N^1}\lambda_j(\mathrm{H}^1_j-\mu_{\mathrm{H}^1})^2}{\sigma_{\mathrm{H}^1}}$$

$$:= \mathcal{L}_{MI^1}(\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^1,\mathcal{G}^2)$$

Proof. Given the perturbed graph $\mathcal{G}^1_{\mathrm{CIB}}$ and its representation $z_{\mathcal{G}^1_{\mathrm{CIB}}}$, we assume there is no information loss during the readout process, i.e., $\mathrm{I}\left(z_{\mathcal{G}^1_{\mathrm{CIB}}};\mathcal{G}^1,\mathcal{G}^2\right)=\mathrm{I}(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^1,\mathcal{G}^2)$.

$$\mathrm{I}\left(z_{\mathcal{G}^1_{\mathrm{CIB}}};\mathcal{G}^1,\mathcal{G}^2\right)=\mathbb{E}_{z_{\mathcal{G}^1_{\mathrm{CIB}}},\mathcal{G}^1,\mathcal{G}^2}\left[-\log\frac{p_\phi\left(z_{\mathcal{G}^1_{\mathrm{CIB}}}|\mathcal{G}^1,\mathcal{G}^2\right)}{p\left(z_{\mathcal{G}^1_{\mathrm{CIB}}}\right)}\right]$$

$$=\mathbb{E}_{\mathcal{G}^1,\mathcal{G}^2}\left[-\log\frac{p_\phi\left(z_{\mathcal{G}^1_{\mathrm{CIB}}}|\mathcal{G}^1,\mathcal{G}^2\right)}{q\left(z_{\mathcal{G}^1_{\mathrm{CIB}}}\right)}\right]-\mathbb{E}_{z_{\mathcal{G}^1_{\mathrm{CIB}}},\mathcal{G}^1,\mathcal{G}^2}\left[KL\left(p\left(z_{\mathcal{G}^1_{\mathrm{CIB}}}\right)||q\left(z_{\mathcal{G}^1_{\mathrm{CIB}}}\right)\right)\right]$$

$$\le \mathbb{E}_{z_{\mathcal{G}^1_{\mathrm{CIB}}},\mathcal{G}^1,\mathcal{G}^2}\left[KL\left(p_\phi\left(z_{\mathcal{G}^1_{\mathrm{CIB}}}|\mathcal{G}^1,\mathcal{G}^2\right)||q\left(z_{\mathcal{G}^1_{\mathrm{CIB}}}\right)\right)\right] \quad (1) \qquad \because \text{Non-negativity of KL divergence}$$

Assuming that $q\left(z_{\mathcal{G}^1_{\mathrm{CIB}}}\right)$ is Gaussian distribution.
The noise $\varepsilon\sim N(\mu_{\mathbf{H}^1},\sigma_{\mathbf{H}^1})$ is sampled from Gaussian distribution where $\mu_{\mathbf{H}^1}$ and $\sigma_{\mathbf{H}^1}$ are mean and variance of $\mathbf{H}^1$.

Thus, $q\left(z_{\mathcal{G}^1_{\mathrm{CIB}}}\right)=N(N^1\mu_{\mathbf{H}^1},N^1\sigma_{\mathbf{H}^1})$ (2) $\because$ Summation of Gaussian is Gaussian

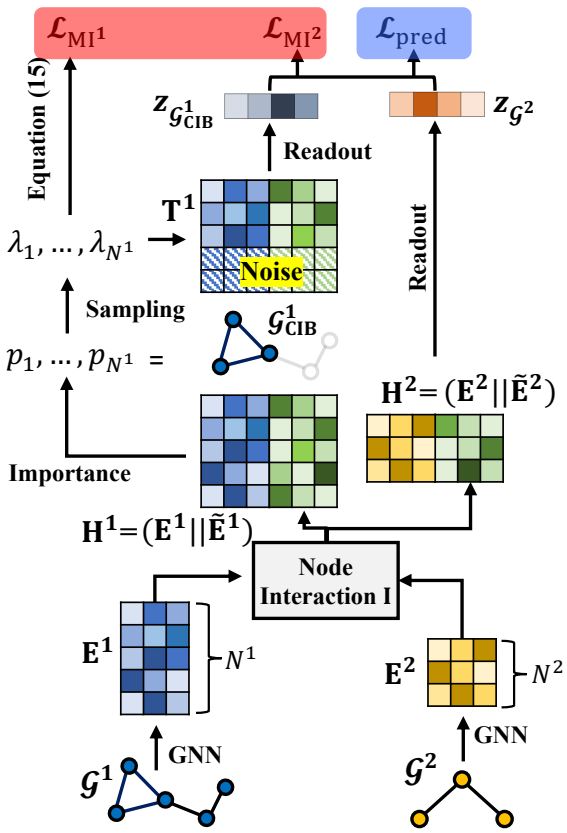
And, $p\left(z_{\mathcal{G}^1_{\mathrm{CIB}}}|\mathcal{G}^1,\mathcal{G}^2\right)=N(N^1\mu_{\mathbf{H}^1}+\sum_{j=1}^{N^1}\lambda_j\mathbf{H}^1_j-\sum_{j=1}^{N^1}\lambda_j\mu_{\mathbf{H}^1},\sum_{j=1}^{N^1}(1-\lambda_j)^2\sigma^2_{\mathbf{H}^1})$ (3)

By plugging Equation (2) and (3) into (1), we have:

$$-\mathrm{I}(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^1,\mathcal{G}^2) \le \mathbb{E}_{\mathcal{G}^1,\mathcal{G}^2}\left[-\frac{1}{2}\log A+\frac{1}{2N^1}A+\frac{1}{2N^1}B^2\right]+C \quad \text{where } A=\sum_{j=1}^{N^1}(1-\lambda_j)^2 \text{ and } B=\frac{\sum_{j=1}^{N^1}\lambda_j(\mathrm{H}^1_j-\mu_{\mathrm{H}^1})^2}{\sigma_{\mathrm{H}^1}}$$

$$\min -I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)+\beta\, I(\mathcal{G}^1;\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)$$

METHODOLOGY CONDITIONAL GRAPH INFORMATION BOTTLENECK



$$I(\mathcal{G}^1;\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2) = I(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^1,\mathcal{G}^2) - I(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^2) \quad \because \text{Chain rule of mutual information}$$

Solute Prediction

Encourage $\mathcal{G}^1_{\mathrm{CIB}}$, which is compressed conditioned on $\mathcal{G}^2$, to contain as much information about $\mathcal{G}^2$ as possible

Intuition) Make use of $\mathcal{G}^2$ when detecting $\mathcal{G}^1_{\mathrm{CIB}}$

1) Variational IB-based approach

Derive upper bound similar to the prediction loss

$$-I(\mathcal{G}^1_{\mathrm{CIB}};\mathcal{G}^2) \le \mathbb{E}_{\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2}[-\log p_\xi(\mathcal{G}^2|\mathcal{G}^1_{\mathrm{CIB}})] := \mathcal{L}_{\mathrm{MI}^2}(\mathcal{G}^1_{\mathrm{CIB}},\mathcal{G}^2)$$

2) Contrastive Learning-based approach

- Minimizing the contrastive loss is proven to be equivalent to maximizing the mutual information
- Hence, minimize $-\mathrm{I}(\mathrm{G}^1_{\mathrm{CIB}};\mathrm{G}^2)$ by minimizing the contrastive loss → $\mathrm{CGIB}_{\mathrm{cont}}$

$$\mathcal{L}_{\mathrm{MI}^2} = -\frac{1}{K}\sum_{i=1}^{K}\log\frac{\exp(\mathrm{sim}(\mathbf{z}_{\mathcal{G}^1_{\mathrm{CIB},i}},\mathbf{z}_{\mathcal{G}^2_i})/\tau)}{\sum_{j=1,j\neq i}^{K}\exp(\mathrm{sim}(\mathbf{z}_{\mathcal{G}^1_{\mathrm{CIB},i}},\mathbf{z}_{\mathcal{G}^2_j})/\tau)}$$

EXPERIMENTS DATASET

Dataset		$\mathcal{G}^1$	$\mathcal{G}^2$	# $\mathcal{G}^1$	# $\mathcal{G}^2$	# Pairs	Task
Chromophore [1]	Absorption	Chrom.	Solvent	6416	725	17276	reg.
	Emission	Chrom.	Solvent	6412	1021	18141	reg.
	Lifetime	Chrom.	Solvent	2755	247	6960	reg.
MNSol [2]		Solute	Solvent	372	86	2275	reg.
FreeSolv [3]		Solute	Solvent	560	1	560	reg.
CompSol [4]		Solute	Solvent	442	259	3548	reg.
Abraham [5]		Solute	Solvent	1038	122	6091	reg.
CombiSolv [6]		Solute	Solvent	1495	326	10145	reg.
ZhangDDI [7]		Drug	Drug	544	544	40255	cls.
ChChMiner [8]		Drug	Drug	949	949	21082	cls.

Dataset statistics

Chromophore dataset
Absorption max, Emission max, Lifetime

Solvation Free Energy dataset
- MNSol
- FreeSolv
- CompSol
- Abraham
- CombiSolv

Drug-Drug Interaction dataset
- ZhangDDI
- ChChMiner

EXPERIMENTS MAIN TABLE

	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)
CGIB	**17.87** (0.38)	24.44 (0.21)	0.796 (0.010)	0.568 (0.013)	**0.831**(0.012)	0.277 (0.008)	0.396 (0.009)	0.428 (0.009)
$CGIB_{cont}$	18.11 (0.20)	**23.90** (0.35)	**0.771** (0.005)	**0.538** (0.007)	0.852 (0.022)	**0.276** (0.017)	**0.390** (0.006)	**0.422** (0.005)

Performance on Molecular Interaction

Observations

Outperforms baselines on both Molecular Interaction / Drug-Drug Interaction tasks

Evaluation on drugs unseen during training

	(a) Transductive				(b) Inductive			
	ZhangDDI		ChChMiner		ZhangDDI		ChChMiner	
	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)	68.39 (1.85)	63.78 (1.55)	73.63 (0.44)	67.07 (0.66)
GAT	92.10 (0.28)	84.14 (0.38)	96.15 (0.53)	89.49 (0.88)	69.99 (2.95)	64.41 (1.39)	75.72 (1.66)	68.77 (1.48)
MPNN	92.34 (0.35)	84.56 (0.31)	96.25 (0.53)	90.02 (0.42)	71.54 (1.24)	65.12 (1.14)	75.45 (0.32)	68.24 (1.42)
GIN	93.16 (0.04)	85.59 (0.05)	97.52 (0.05)	91.89 (0.66)	72.74 (1.32)	66.16 (1.21)	74.63 (0.48)	67.80 (0.46)
SSI-DDI	92.74 (0.12)	84.61 (0.18)	98.44 (0.08)	93.50 (0.16)	73.29 (2.23)	66.53 (1.31)	78.24 (1.29)	70.69 (1.47)
MIRACLE	93.05 (0.07)	84.90 (0.36)	88.66 (0.37)	84.29 (0.14)	73.23 (3.32)	50.00 (0.00)	60.25 (0.56)	50.09 (0.11)
CIGIN	93.28 (0.13)	85.54 (0.30)	98.51 (0.10)	93.77 (0.25)	74.02 (0.10)	66.81 (0.09)	79.23 (0.51)	71.56 (0.38)
CGIB	**94.27** (0.47)	**86.88** (0.56)	98.80 (0.04)	**94.69** (0.16)	74.59 (0.88)	**67.65** (1.07)	81.14 (1.20)	72.47 (0.16)
$CGIB_{cont}$	93.78 (0.62)	86.36 (0.75)	**98.84** (0.31)	94.52 (0.38)	**75.08** (0.34)	67.31 (0.82)	**81.51** (0.67)	**74.29** (0.14)

Performance on Drug-Drug Interaction

Observations

Improvement gap is larger in inductive setting

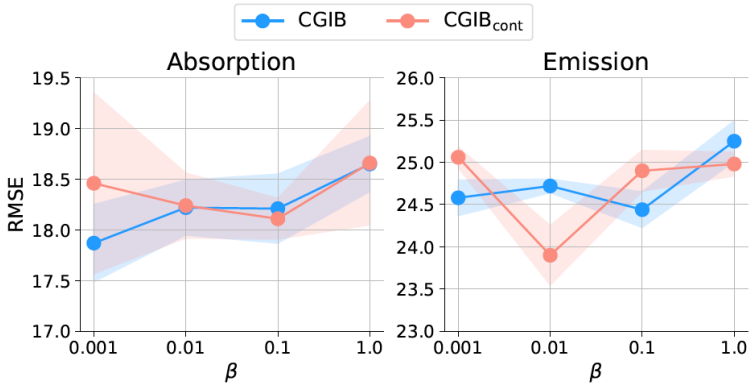
∵ By detecting function group that is basic in nature → helps generalization

EXPERIMENTS SENSITIVITY ANALYSIS

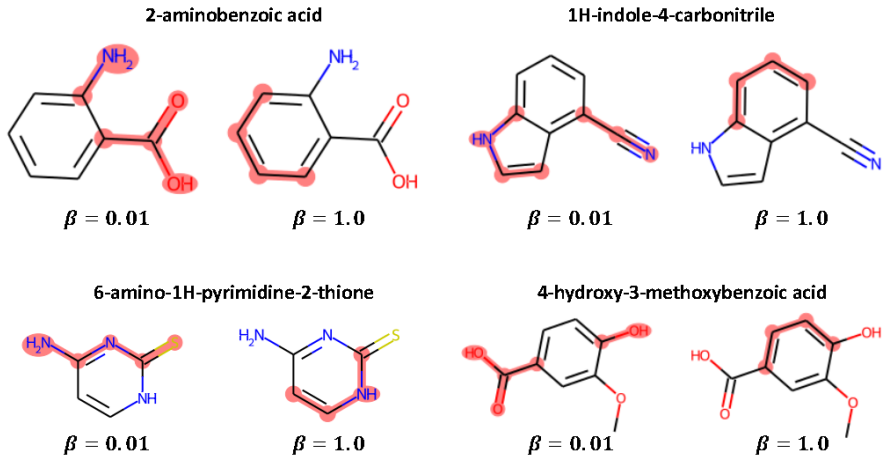
$$\min -I(\mathrm{Y};\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2) + \beta I(\mathcal{G}^1;\mathcal{G}^1_{\mathrm{CIB}}|\mathcal{G}^2)$$

- β Controls Trade-off btw prediction and compression

As β increases, Compression > Prediction



Sensitivity Analysis on beta



Qualitative Analysis on beta

Observations

- $\beta = 1.0$: Poor performance in general (focus on compression)
- However, the model fails to detect functional group when β is too small

→ poor generalization

→ Hence, finding an appropriate β is crucial

Observations

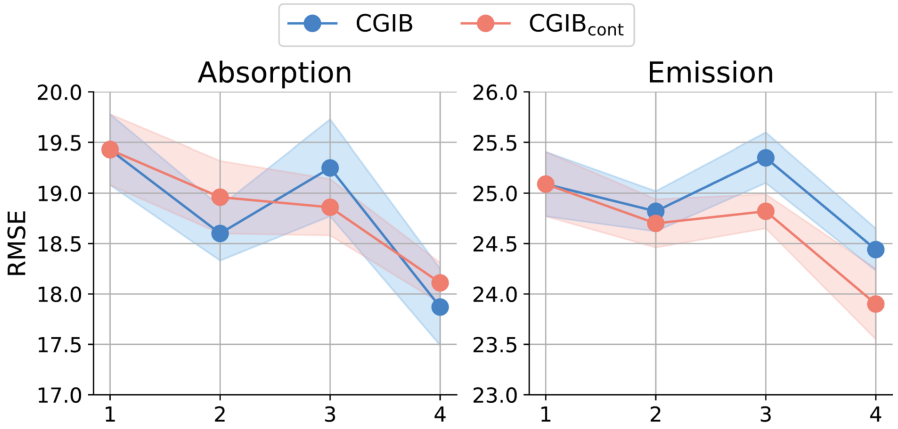
- $\beta = 1.0$ → CGIB focuses on compression

e.g., CGIB focuses an aromatic ring, which is not relevant to chemical reactions

- $\beta = 0.01$ → CGIB focuses on prediction

e.g., CGIB focuses on external part, which generally more relevant to chemical reactions

EXPERIMENTS ABLATION STUDIES



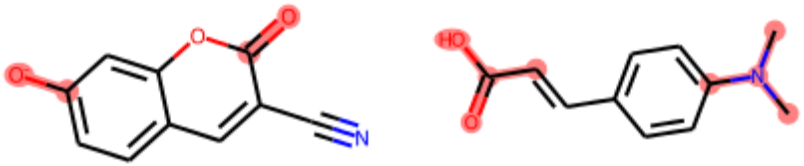
1. **Without IB** → $\min -I(Y;\mathcal{G}^1,\mathcal{G}^2)$ **(Same as CIGIN)**
2. $I(\mathcal{G}^1_{\text{CIB}};\mathcal{G}^1)$ → $\min -I(Y;\mathcal{G}^1_{\text{CIB}},\mathcal{G}^2)+I(\mathcal{G}^1_{\text{CIB}};\mathcal{G}^1)$ **(Same as VGIB)**
3. $I(\mathcal{G}^1_{\text{CIB}};\mathcal{G}^1,\mathcal{G}^2)$ → $\min -I(Y;\mathcal{G}^1_{\text{CIB}},\mathcal{G}^2)+I(\mathcal{G}^1_{\text{CIB}};\mathcal{G}^1,\mathcal{G}^2)$
4. $I(\mathcal{G}^1;\mathcal{G}^1_{\text{CIB}}|\mathcal{G}^2)$ → $\min -I(Y;\mathcal{G}^1_{\text{CIB}},\mathcal{G}^2)+I(\mathcal{G}^1;\mathcal{G}^1_{\text{CIB}}|\mathcal{G}^2)$ **(Same as CGIB)**

Observations

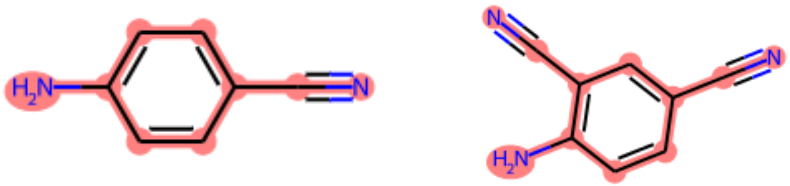
- Considering conditional MI is the key for success in relational learning
- A naïve consideration of $\mathcal{G}^1$ and $\mathcal{G}^2$ rather performs worse than considering $\mathcal{G}^1$ only

EXPERIMENTS QUALITATIVE ANALYSIS

(a) Ordinary solvents



(b) Liquid oxygen solvent



Observations

(a) Chromophore ($\mathcal{G}^1$) interact with ordinary solvents ($\mathcal{G}^2$)

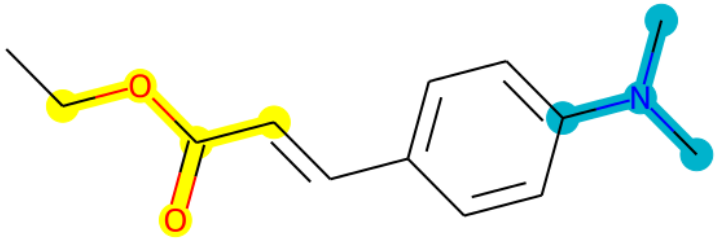
Focus on external parts → Aligns with domain knowledge

(b) Chromophore ($\mathcal{G}^1$) interact with liquid oxygen solvents ($\mathcal{G}^2$)

Focus on all parts → Aligns with domain knowledge

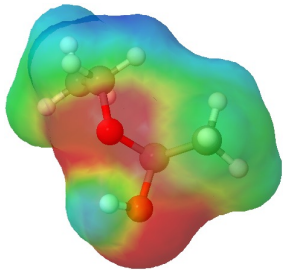
EXPERIMENTS QUALITATIVE ANALYSIS

(c) Ethanol, THF, 1-Hexanol, 1-Butanol | Benzene

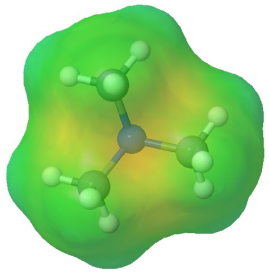


Oxygen-Carbon | Nitrogen-Carbon

(d) Polarity of the structure



Oxygen-Carbon
(High Polarity)



Nitrogen-Carbon
(Low Polarity)

Observations

(c) Chromophore ($\mathcal{G}^1$) interacts with various solvents ($\mathcal{G}^2$)
(e.g., Trans-ethyl p-(dimethylamino) cinnamate (EDAC))
Detected parts in chromophore depend on the polarity of solvent

- Case 1: High polarity solvent (Ethanol, THF, 1-hexanol, 1-butanol)
Structure with high polarity is detected (e.g., Oxygen-carbon)
→ Interact with high polarity solvent

- Case 2: Low polarity solvent (Benzene solvent)
Structure with low polarity is detected (e.g., Nitrogen-Carbon)
→ Interact with low polarity solvent

Detected structure of Chromophore (G^1) depends on the paired solvents (G^2)

CONCLUSION

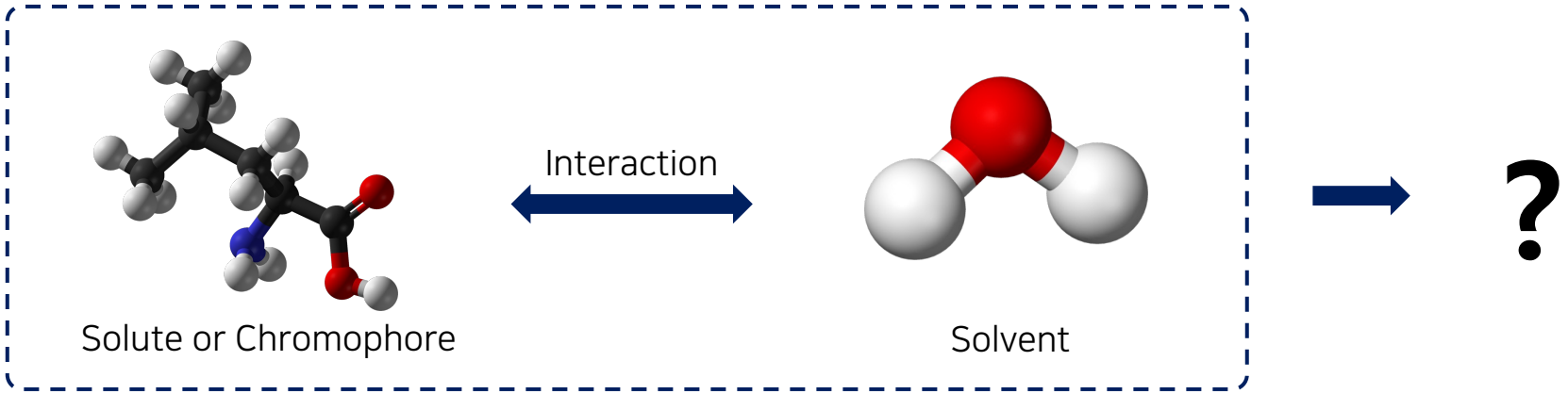
Proposed a method for tackling relation learning tasks, which are crucial for scientific discovery

- Based on Conditional Information Bottleneck

It is crucial to consider Graph 2 (Solvent) when detecting the important subgraph from Graph 1 (Chromophore)

- i.e., Make use of $\mathcal{G}^2$ when detecting $\mathcal{G}^1_{\text{CIB}}$ of $\mathcal{G}^1$

CGIB has interpretability, which makes it highly practical



THANK YOU!

[Full Paper] https://openreview.net/forum?id=5hz3GV4lPq

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

[Author Email] namkyeong96@kaist.ac.kr