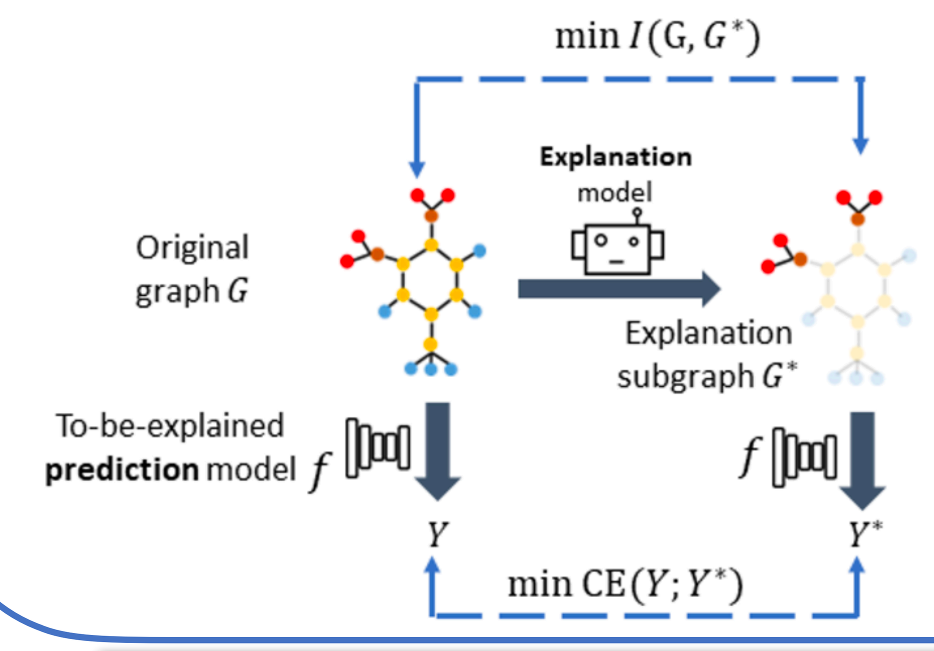
# RegExplainer: Generating Explanations for Graph Neural Networks in Regression Tasks

Jiaxing Zhang[1], Zhuomin Chen[2], Hao Mei[3], Longchao Da[3], Dongsheng Luo[2], Hua Wei[3]

[1]New Jersey Institute of Technology, [2]Florida International University, [3]Arizona State University

## Explaining GNNs

### Post-hoc Instance-level Explanation




Discover subgraphs that explain the prediction of a trained GNN.

$$\underset{G^*}{\operatorname{argmin}} I(G; G^*) + \alpha H(Y|G^*)$$

## Graph Information Bottleneck(GIB) in previous classification tasks couldn't be trivially used in explaining regression tasks.

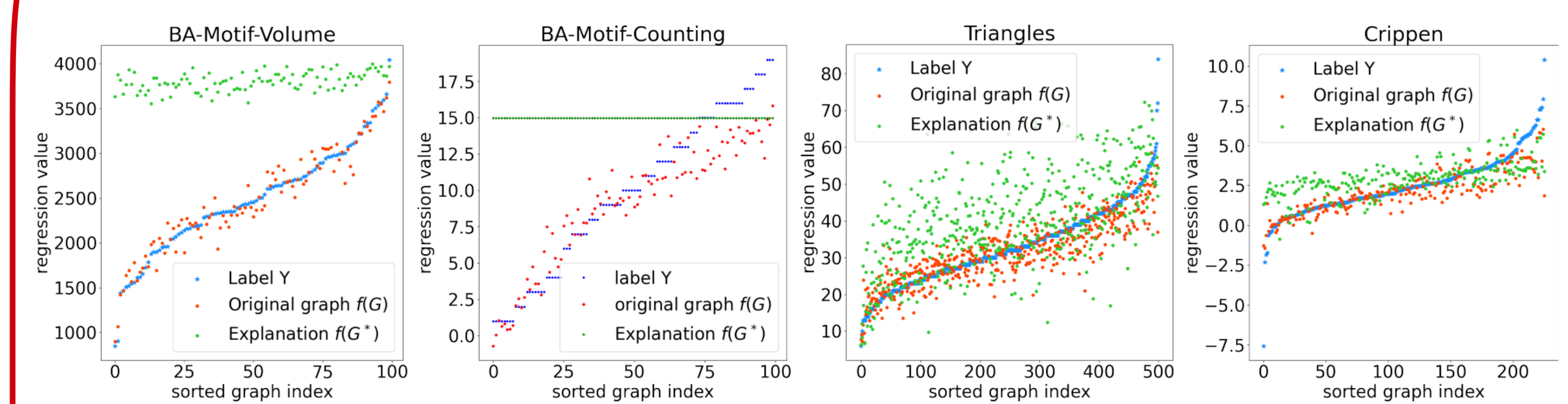
Mutual Information $I(G^*; Y) = H(Y) - H(Y|G^*)$

$$\underset{G^*}{\operatorname{argmin}} I(G; G^*) - \alpha I(G^*; Y)$$

Intractability of $I(G^*; Y)$

$$\underset{G^*}{\operatorname{argmin}} I(G; G^*) - \alpha I(f(G^*); Y) \quad [2][3]$$

### Diverging Distributions between $f(G^*)$ and $Y$ !




**Prediction of ground truth explanation diverges from the prediction of original graph**

| Dataset | $(f(G), Y)$ | $(f(G^*), Y)$ | $(f(G), f(G^*))$ |
|---|---|---|---|
| BA-Motif-Volume | 131.42 | 1432.07 | 1427.07 |
| BA-Motif-Counting | 2.06 | 7.43 | 7.22 |
| Triangles | 5.28 | 12.38 | 12.40 |
| Crippen | 1.13 | 1.54 | 1.17 |

Prediction shifting study on the RMSE.

| | BA-Motif-Volume | BA-Motif-Counting | Triangles | Crippen |
|---|---|---|---|---|
| $\mathrm{COS}(v_g, v_e)$ | 0.95 | 0.80 | 0.97 | 0.89 |
| $\mathrm{COS}(v_g, v_m)$ | 0.98 | 0.89 | 0.99 | 0.92 |
| $\mathrm{EUC}(v_g, v_e)$ | 0.46 | 0.68 | 0.19 | 0.67 |
| $\mathrm{EUC}(v_g, v_m)$ | 0.37 | 0.52 | 0.08 | 0.63 |
| $\mathrm{RMSE}(p_g, p_e)$ | 1427.07 | 7.22 | 12.40 | 1.17 |
| $\mathrm{RMSE}(p_g, p_m)$ | 393.26 | 2.73 | 8.22 | 0.68 |

**Mixed explanation could alleviate this distribution shifting problem. [1]**

### We adopt the GIB objective with following properties:

**Property 1: $I(Y^*, Y)$ is the lower bound of $I(G^*, Y)$**

$$\underset{G^*}{\operatorname{argmin}} I(G, G^*) - \alpha I(G^*, Y) \rightarrow \underset{G^*}{\operatorname{argmin}} I(G, G^*) - \alpha I(Y^*, Y)$$

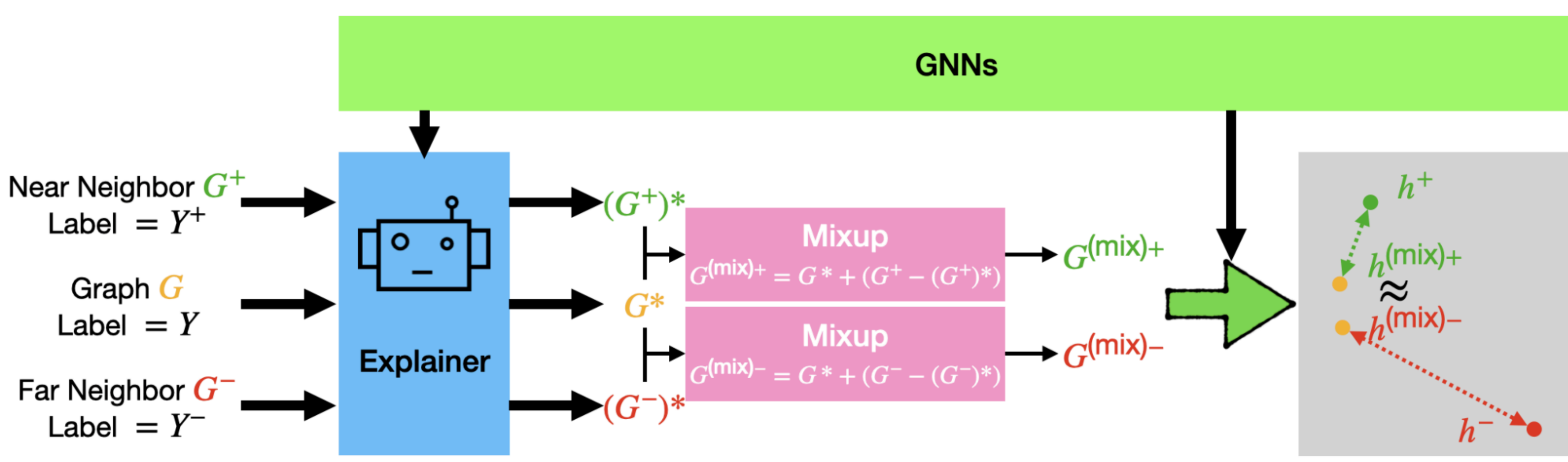
**Property 2: InfoNCE loss is the lower bound of $I(Y^*, Y)$**

$$\underset{G^*}{\operatorname{argmin}} I(G, G^*) - \alpha I(Y^*, Y) \rightarrow \underset{G^*}{\operatorname{argmin}} I(G; G^*) - \alpha \mathbb{E}_{\mathbb{H}} \left[ \log \frac{sim(h^*, h)}{\frac{1}{|\mathbb{H}|} \sum_{h' \in \mathbb{H}} sim(h^*, h')} \right]$$

## Implementation

Overall loss functions:

1. $\mathscr{L}_{\text{contr}}(G, G^+, G^-) = -\log \dfrac{\exp((h^{(\text{mix})+})^T h^+)}{\exp((h^{(\text{mix})+})^T h^+) + \exp((h^{(\text{mix})-})^T h^-)}$
2. $\mathscr{L}_{\text{GIB}} = \mathscr{L}_{\text{size}}(G, G^*) - \alpha \mathscr{L}_{\text{contr}}(G, G^+, G^-)$
3. $\mathscr{L} = \mathscr{L}_{\text{GIB}} + \mathscr{L}_{\text{MSE}} = \mathscr{L}_{\text{GIB}} + \beta \mathscr{L}_{\text{MSE}}(f(G), f(G^{(\text{mix})+}))$




$G$ is the to-be-explained graph, $G^+$ is the randomly sampled positive graph and $G^-$ is the randomly sampled negative graph. The explanation of the graph is produced by the explainer model. Then graph $G$ is mixed with $G^+$ and $G^-$ respectively to produce $G^{(\text{mix})+}$ and $G^{(\text{mix})-}$. Then the graphs are fed into the trained GNN model to retrieve the embedding vectors $h^+$, $h^-$, $h^{(\text{mix})+}$ and $h^{(\text{mix})-}$. We use contrastive loss to minimize the distance between $G^{(\text{mix})+}$ and the positive sample and maximize the distance between $G^{(\text{mix})-}$ and the negative sample. The explainer is trained with the GIB objective and contrastive loss.

**Algorithm 1** Graph Mixup Algorithm

**Input:** Graph $G_a = (X_a, A_a)$, a set of graphs $\mathcal{G}$, the number of random connections $\eta$, explanation model $g$.

**Output:** Graph $G^{(\text{mix})}$.

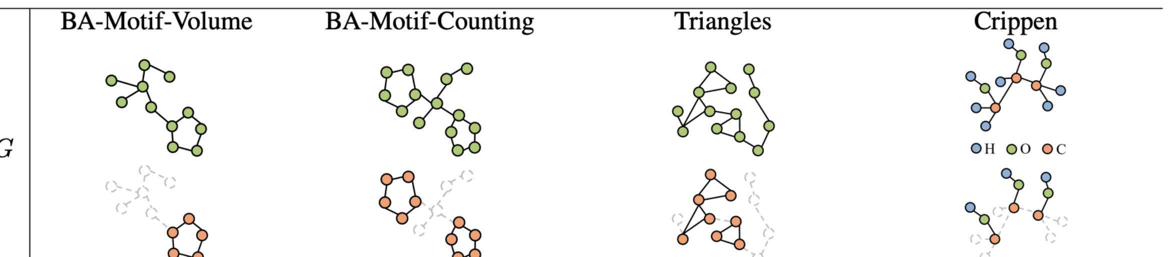
1: Randomly sample a graph $G_b = (A_b, X_b)$ from $\mathcal{G}$
2: Generate mask matrix $M_a = g(G_a)$
3: Generate mask matrix $M_b = g(G_b)$
4: Sample $\eta$ random connections between $G_a$ and $G_b$ as $A_c$
5: Mixup adjacency matrix $A^{(\text{mix})}$ with Eq. (10)
6: Mixup edge mask $M^{(\text{mix})}$ with Eq. (11)
7: Mixup node features $X^{(\text{mix})} = [X_a; X_b]$
8: **return** $G^{(\text{mix})} = (X^{(\text{mix})}, M^{(\text{mix})} \odot A^{(\text{mix})})$

$$A^{(\text{mix})} = \begin{bmatrix} A_a & A_c \\ A_c^T & A_b \end{bmatrix} \qquad M_a^{(\text{mix})} = \begin{bmatrix} \lambda M_a & M_c \\ M_c^T & A_b - \lambda M_b \end{bmatrix}$$

Graph Mix-up Algorithm

## Experiment Results

Table 1: Illustration of the graph regression datasets together with the explanation faithfulness in terms of AUC-ROC on edges under four datasets on RegExplainer and other baselines. The original graph row visualizes the structure of the complete graph, the explanation row highlights the explanation sub-graph of the corresponding original graph. In the Crippen dataset, different colors of the node represent different kinds of atoms and the node feature is a one-hot vector to encode the atom type.

| Dataset | BA-Motif-Volume | BA-Motif-Counting | Triangles | Crippen |
|---|---|---|---|---|
| Original Graph $G$ / Explanation $G^*$ |  | | | |
| Node Feature | Random Float Vector | Fixed Ones Vector | Fixed Ones Vector | One-hot Vector |
| Regression Label | Sum of Motif Value | Number of Motifs | Number of Triangles | Chemical Property Value |
| Explanation Type | Fix Size Sub-Graph | Dynamic Size Sub-graph | Dynamic Size Sub-graph | Dynamic Size Sub-graph |
| Explanation AUC | | | | |
| GRAD | $0.418 \pm 0.000$ | $0.527 \pm 0.000$ | $0.479 \pm 0.000$ | $0.426 \pm 0.000$ |
| ATT | $0.512 \pm 0.005$ | $0.521 \pm 0.003$ | $0.441 \pm 0.004$ | $0.502 \pm 0.006$ |
| MixupExplainer | $0.471 \pm 0.0291$ | $0.868 \pm 0.127$ | $0.663 \pm 0.110$ | $0.499 \pm 0.002$ |
| GNNExplainer | $0.501 \pm 0.009$ | $0.505 \pm 0.004$ | $0.500 \pm 0.002$ | $0.497 \pm 0.005$ |
| **+RegExplainer** | $0.588 \pm 0.017$ | $0.629 \pm 0.001$ | $0.537 \pm 0.003$ | $0.541 \pm 0.011$ |
| PGExplainer | $0.470 \pm 0.057$ | $0.798 \pm 0.133$ | $0.511 \pm 0.028$ | $0.448 \pm 0.005$ |
| **+RegExplainer** | $\mathbf{0.758 \pm 0.177}$ | $\mathbf{0.989 \pm 0.003}$ | $\mathbf{0.739 \pm 0.008}$ | $\mathbf{0.553 \pm 0.013}$ |

### Conclusion

1. Contrastive loss could be applied while explaining the graph regression tasks.
2. Mix explanation with sampled base-graph could help address the distribution shifting issue.

CODE PAPER

References:

[1]. Zhang, et al., "MixupExplainer: Generalizing Explanations for Graph Neural Networks with Data Augmentation." SIGKDD 2023.

[2]. Miao et al., Interpretable and generalizable graph learning via stochastic attention mechanism. ICML 2022.

[3]. Luo, et al., Parameterized explainer for graph neural network. NeurIPS 2020.