Self-Interpretable Graph Learning with Sufficient and Necessary Explanations

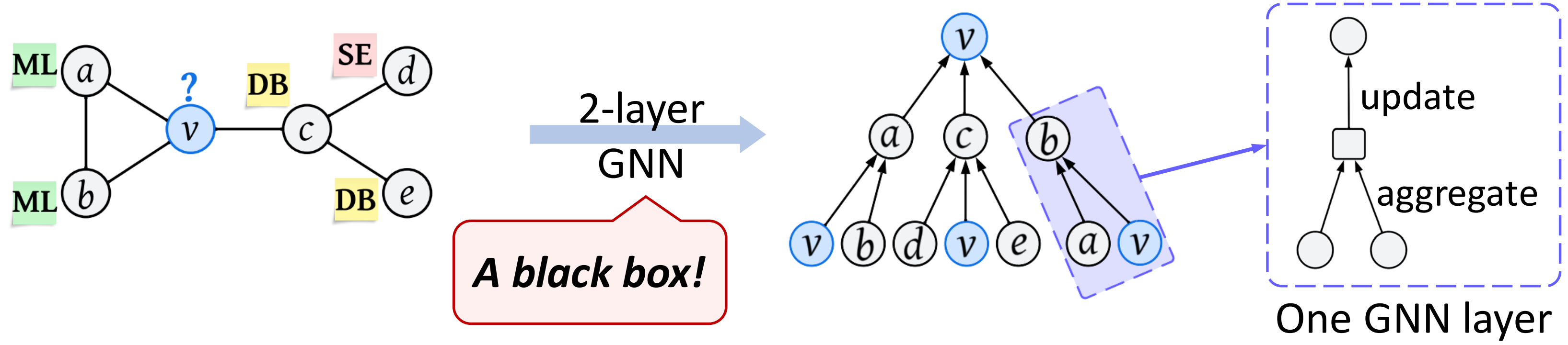
Jiale Deng, Yanyan Shen

Shanghai Jiao Tong University

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Deep Learning on Graphs

- Graphs are everywhere
  - Social networks, co-purchase graphs, paper citation networks, …
- Graph Neural Networks (GNNs) with message passing
  - Strong performance but lack transparency;
  - Learning representations by aggregating and updating information from neighbors;
  - An example (predicting the research area of an author node in coauthor network):



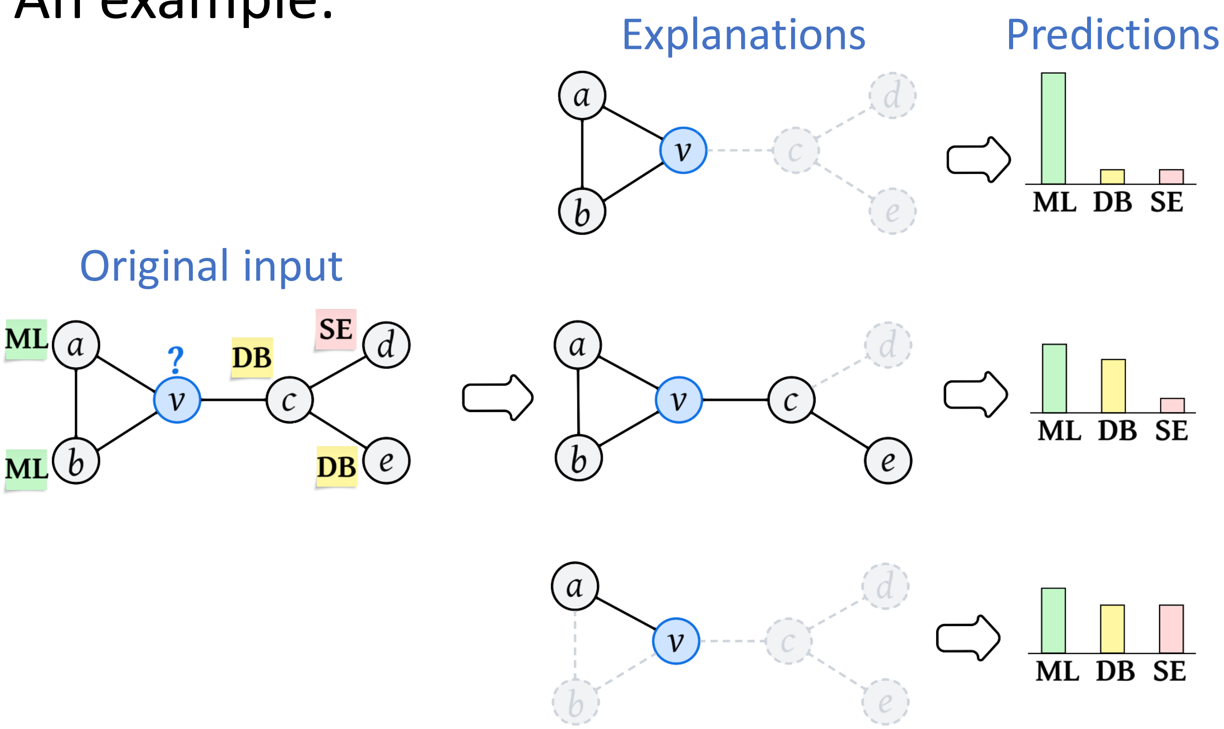
Explaining GNNs

- Post-hoc methods
  - Explain a fix GNN by extracting salient substructures from the input graph;
  - Post-hoc explanations can be biased and inconsistent[1].
    - Not directly produced by the GNN model!
- Self-interpretable methods[2]
  - Simultaneously provide predictions and built-in explanations (unbiased);
  - Consist of:
    - An explanation generator: extracts explanation from input;
    - A predictor: learns the representation from the explanation to make final prediction.

[1] *Dai, Enyan, and Suhang Wang. "Towards self-explainable graph neural network." CIKM 2021.*
[2] *Miao, Siqi, Mia Liu, and Pan Li. "Interpretable and generalizable graph learning via stochastic attention mechanism." ICML 2022.*

Issues of Self-Interpretable Methods

- The quality of explanations determines the model performance.
- An example:



Sufficient & Necessary:
Ideal explanation capturing the rationale that *v, a, b form an ML research group (clique)*.

Sufficient[1-3]:
Contains the clique while *introduce noisy edges (e,c) and (c,v)*.

Necessary[4-6]:
Missing the salient clique.

Compromised performance!

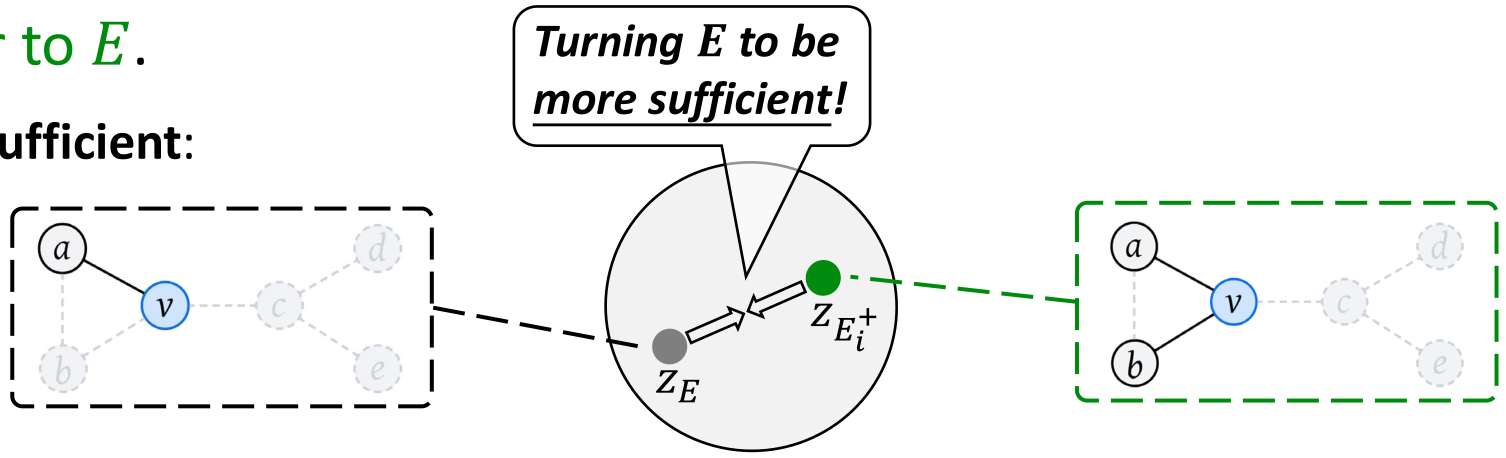
[1] Yu, Junchi, et al. "Graph Information Bottleneck for Subgraph Recognition." ICLR 2020.
[2] Dai, Enyan, et al. "Towards self-explainable graph neural network." CIKM 2021.
[3] Miao, Siqi, et al. "Interpretable and generalizable graph learning via stochastic attention mechanism." ICML 2022.
[4] Wu, Yingxin, et al. "Discovering Invariant Rationales for Graph Neural Networks." ICLR 2021.
[5] Sui, Yongduo, et al. "Causal attention for interpretable and generalizable graph classification." KDD 2022.
[6] Fan, Shaohua, et al. "Debiasing graph neural networks via learning disentangled causal substructure." NIPS 2022.

Key Insight: Overview

- A self-interpretable graph learning framework with SUfficient aNd NecessarY explanations (SUNNY-GNN).
- Our goal
  - Promote the quality of explanations toward both sufficiency and necessity directions, thus encourage the explanations to improve model performance.
- How to promote?
  - Perform augmentations on explanations and employ a contrastive loss to supervise the explanation generator for producing sufficient and necessary explanations.

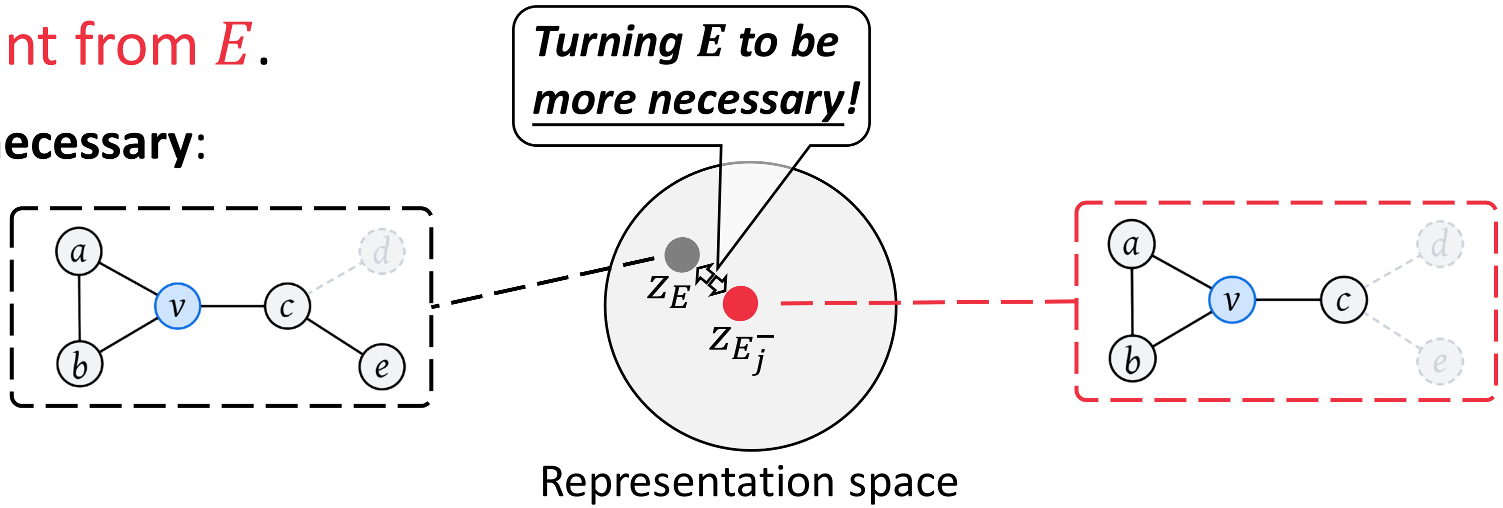
Key Insight: Augmentations and Contrastive Loss

- Augmentations on an explanation E
  - Positive samples E^+: add information (e.g., edges) to E;
  - Negative samples E^-: remove information (e.g., edges) from E;
  - Construct contrastive loss $\mathcal{L}_{cts}$ with E^+ and E^-;
- Minimizing $\mathcal{L}_{cts}$ means
  - Pulling E^+ closer to E.
    - When E is **not sufficient**:



Key Insight: Augmentations and Contrastive Loss

- Augmentations on an explanation E
  - Positive samples E^+: add edges to E;
  - Negative samples E^-: delete edges in E;
  - Construct Contrastive Loss $\mathcal{L}_{cts}$ with E^+ and E^-;
- Minimizing $\mathcal{L}_{cts}$ means
  - Pushing E^- distant from E.
    - When E is **not necessary**:



- Workflow of SUNNY-GNN*

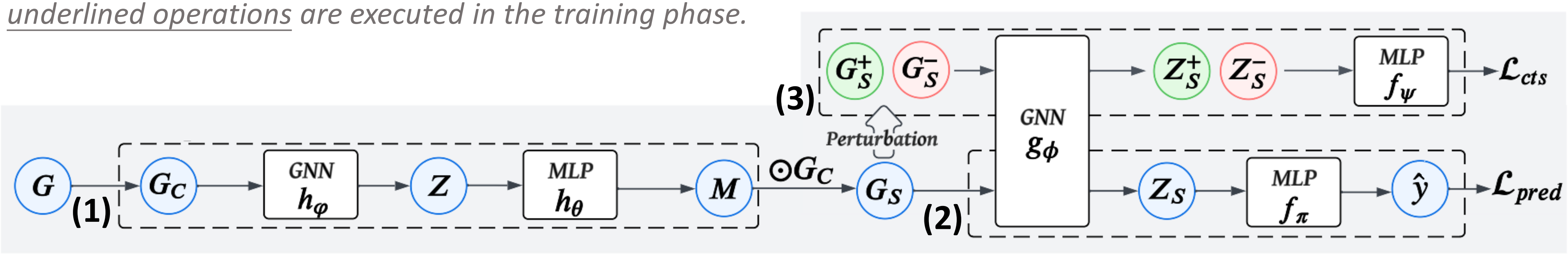
  (1) Explanation generation: Extract the explanation G_S from input G;

  (2) Prediction: Encode G_S and yield the prediction $\hat{y}$, compute the prediction loss $\mathcal{L}_{pred}$;

  (3) Augmentation: Perturb G_S to get a set of positive and negative samples, compute the contrastive loss $\mathcal{L}_{cts}$.

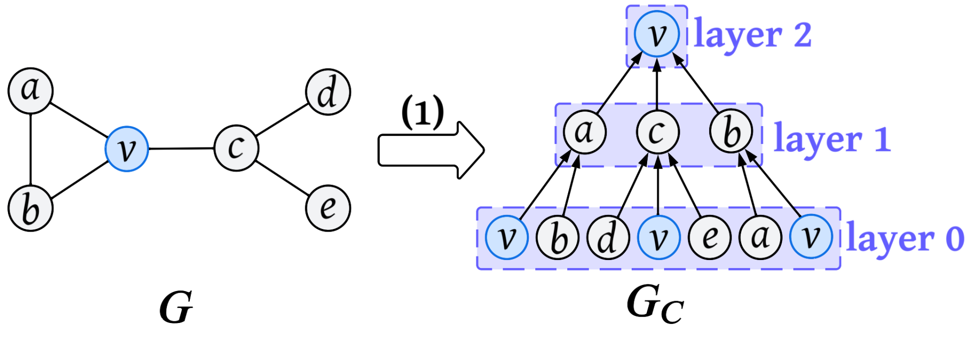
  - SUNNY-GNN is trained in an end-to-end manner
    - The update of Parameters $\Theta = \{\varphi, \theta, \phi, \pi, \psi\}$ are supervised by $\mathcal{L}_{pred}$ and $\mathcal{L}_{cts}$.

** The underlined operations are executed in the training phase.*



SUNNY-GNN: Explanation generation

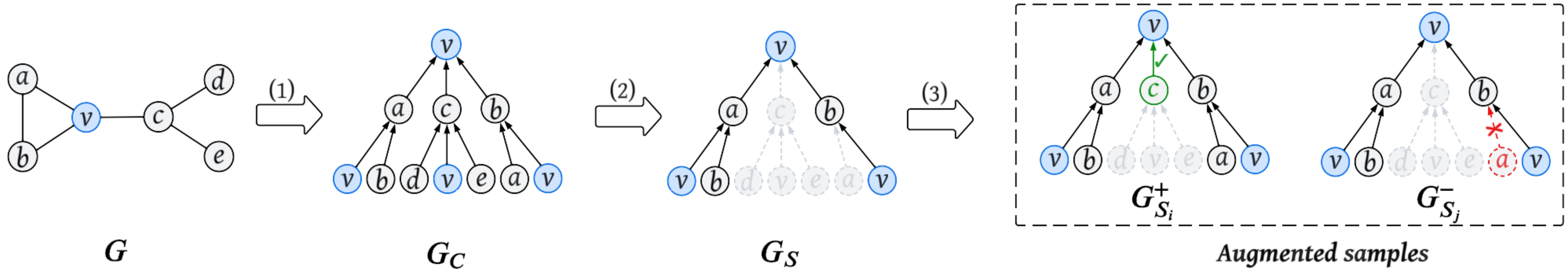
- Generate edge mask M
  - $m_{ij} = h_\theta(z_i^{(0)}||z_j^{(1)}||z_v^{(2)})$ if (i, j) connects nodes between layer 0 and 1, e.g., edge (e, c);
  - $m_{ij} = h_\theta(z_i^{(1)}||z_j^{(2)}||z_v^{(2)})$ if (i, j) connects nodes between layer 1 and 2, e.g., edge (c, v).



- Get explainable subgraph G_S
  - Training phase: $G_{S_{att}} = G_C \odot M$ (*differentiable*)
  - Testing phase: $G_S \sim \mathrm{Bern}(G_{S_{att}})$ (*not differentiable*)

SUNNY-GNN: Augmentation

- Positive samples
  - Sample edges from $G_C \backslash G_S$ and add them to G_S.
- Negative samples
  - Sample edges from G_S and remove them.



SUNNY-GNN: Augmentation

- How to sample?
  - A simple way: by edge mask, but may lead to
    - Trivial samples that impair the self-supervision signals;
    - Unreliable samples that mislead the GNN training.
  - Distance coefficients → Enhance contrastive signal
    - Intuition: *perturbations on edges closer to the target node v tend to have a greater impact to v than those on farther edges*.
    - $$\delta^+ = 1 - \alpha \cdot \exp(\mathrm{d}) \in \mathbb{R}^{|\mathcal{E}_\mathrm{C}|}$$ $$\delta^- = \alpha \cdot \exp(\mathrm{d}) \in \mathbb{R}^{|\mathcal{E}_\mathrm{C}|}$$, where $\alpha \in \mathbb{R}^+$ is a positive constant and $\mathrm{d} \in \{1, ..., L\}^{|\mathcal{E}_C|}$.
    - Get positive and negative samples by sampling edges with $M \odot \delta^+$ and $M \odot \delta^-$, respectively.

SUNNY-GNN: Augmentation

- How to sample?
  - A simple way: by edge mask, but may lead to
    - Trivial samples that impair the self-supervision signals;
    - Unreliable samples that mislead the GNN training.
  - Confidence coefficients → Filter out unreliable samples with labels
    - Intuition: introducing noisy edges and removing irrelevant edges forms *untrustworthy* positive and negative samples, respectively.
    - $\eta^{+} = \mathrm{SoftMax}(f_\pi(g_\phi(G_S^{+}))y_{vt}) \in \mathbb{R}^{n^{+}}$, $\eta^{-} = (1 - \mathrm{SoftMax}(f_\pi(g_\phi(G_S^{-}))y_{vt})) \in \mathbb{R}^{n^{-}}$, where $f_\pi(g_\phi(G_S^{+}))y_{vt}$ the prediction probability of samples in the truth label t of target node v.
    - Reweight the augmented samples in the contrastive loss by η^{+} and η^{-}.

SUNNY-GNN: Prediction and Optimization

- Prediction: $z_S = g_\phi(G_S),\ \hat{y} = f_\pi(z_S)$
- Optimization: $\min_{\Theta} \mathcal{L}_{pred} + \gamma \mathcal{L}_{cts}$
  - Prediction loss:
    $$\mathcal{L}_{pred} = -\frac{1}{|\mathcal{V}_{train}|} \sum_{v \in \mathcal{V}_{train}} \sum_{t=1}^{\mathcal{T}} \mathcal{Y}_{vt} \log \hat{\mathcal{Y}}_{vt}$$
  - Contrastive loss
    - For one node v: $\mathcal{L}_{cts}(v) = \mathbb{E}\left[-\log \frac{\eta_i^+ \exp(z_S^\top z_{S_i}^+/\tau)}{\sum_j \eta_j^+ \exp(z_S^\top z_{S_j}^+/\tau) + \sum_k \eta_k^- \exp(z_S^\top z_{S_k}^-/\tau)}\right]$;
    - Contrastive loss over all training nodes: $\mathcal{L}_{cts} = \frac{1}{|\mathcal{V}_{train}|} \sum_{v \in \mathcal{V}_{train}} \mathcal{L}_{cts}(v)$

Evaluations of Prediction Performance

	Citeseer	Cora	Pubmed	Amazon	Coauthor-CS	Coauthor-Physics
GCN	69.84±0.7	81.20±0.7	77.68±0.7	90.18±0.3	83.52±0.4	92.46±0.2
+ GSAT	**70.90±1.1**	81.48±0.7	77.44±0.3	88.36±1.3	83.76±0.6	92.14±0.5
+ CAL	65.60±1.1	75.72±1.2	73.66±0.8	84.32±1.7	82.12±1.2	91.26±0.7
+ SE-GNN	68.90±0.9	80.72±0.1	77.56±0.3	-	83.14±0.8	-
+ ProtGNN	66.30±2.1	77.48±8.7	74.18±3.3	82.46±1.4	79.50±3.7	88.80±3.3
+ SUNNY-GNN	70.72±0.8	**81.68±0.9**	**78.68±0.2**	**90.43±0.4**	**85.03±1.1**	**93.10±0.8**
Average impro. (%)	3.6 ↑	3.1 ↑	3.4 ↑	4.8 ↑	3.2 ↑	3.0 ↑
GAT	69.68±1.2	81.22±0.7	77.50±0.4	89.08±1.8	84.42±0.8	92.30±0.5
+ GSAT	69.42±0.8	81.20±0.7	77.04±0.3	89.73±0.4	84.37±0.7	91.90±0.8
+ CAL	67.64±1.5	76.64±1.1	74.74±0.7	84.86±11.5	78.69±3.8	78.24±5.1
+ SE-GNN	68.18±1.1	79.46±0.4	75.88±0.4	-	83.71±0.5	-
+ ProtGNN	69.90±1.5	80.40±0.9	76.84±0.8	86.52±0.3	80.95±1.2	90.42±2.3
+ SUNNY-GNN	**71.30±0.7**	**82.18±1.3**	**78.14±0.3**	**90.78±0.4**	**85.13±0.5**	**93.06±0.6**
Average impro. (%)	3.2 ↑	3.1 ↑	2.3 ↑	3.0 ↑	3.4 ↑	6.0 ↑

Table 2: Classification Acc(%). The best and second-best results are bolded and underlined, respectively.

SUNNY-GNN outperforms all baselines by 3.5% on average and up to 6.0%.

Evaluations of Explainability Performance

	Citeseer		Cora	
	$fid_+ \uparrow$	$fid_- \downarrow$	$fid_+ \uparrow$	$fid_- \downarrow$
GCN				
+ GNNExplainer	72.27±4.2	9.31±3.4	38.29±4.1	<u>1.08±0.4</u>
+ PGExplainer	82.09±7.2	0.92±2.6	87.47±0.9	1.42±0.3
+ ReFine	83.01±7.1	<u>0.78±0.5</u>	<u>88.19±0.6</u>	**0.00±0.0**
+ GSAT	<u>86.75±5.7</u>	2.72±1.1	76.11±16.0	1.82±1.0
+ CAL	86.44±4.3	12.25±3.9	82.15±9.1	5.78±0.8
+ SUNNY-GNN	**87.29±5.3**	**0.25±0.4**	**90.24±0.3**	**0.00±0.0**
GAT				
+ GNNExplainer	52.95±16.9	8.61±10.1	36.44±21.0	1.43±2.2
+ PGExplainer	76.80±0.3	1.71±2.8	88.25±6.1	<u>0.17±0.2</u>
+ ReFine	<u>77.78±2.8</u>	**0.32±1.3**	<u>88.79±3.2</u>	**0.00±0.0**
+ GSAT	72.52±8.1	1.55±1.3	77.43±8.7	1.03±0.5
+ CAL	77.78±6.5	11.46±1.4	85.23±9.6	5.34±1.5
+ SUNNY-GNN	**79.25±2.4**	<u>0.46±0.5</u>	**91.79±3.2**	**0.00±0.0**

Table 4: Explainability performance (%). The best and second-best results are bolded and underlined, respectively.

- Metrics of explanation quality:

$$fid_- = \frac{1}{N}\sum_{i=1}^{N} |\mathbb{1}(\hat{y}_i = y_i) - \mathbb{1}(\hat{y}_i^{G_S} = y_i)|,$$

$$fid_+ = \frac{1}{N}\sum_{i=1}^{N} |\mathbb{1}(\hat{y}_i = y_i) - \mathbb{1}(\hat{y}_i^{G_C \setminus G_S} = y_i)|$$

  - Smaller fid_- indicates better sufficiency;
  - Larger fid_+ means better necessity.
- SUNNY-GNN generates explanations satisfying both good sufficiency and necessity. It outperforms the baselines by 13.1% on average and up to 33.5%.

Case Studies and Ablation Studies

- Case Studies:

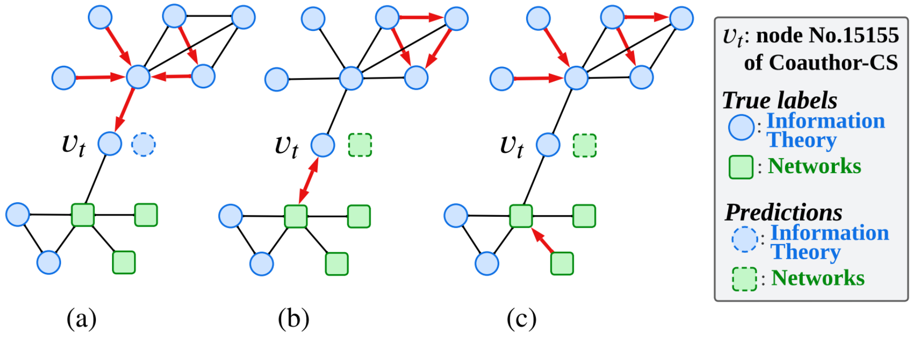


Figure 5: The visualization of explanations generated by (a) SUNNY-GNN, (b) GSAT and (c) CAL in Coauthor-CS.

- SUNNY-GNN highlights the most salient input information, while other baselines fail to.

- Ablation Studies:

	Acc ↑	fid_+ ↑	fid_- ↓
SUNNY-GNN	71.26±0.7	79.25±2.4	0.46±0.5
w/o $\mathcal{L}_{cts}$	69.01±1.2	76.50±1.4	1.78±0.4
w/o δ	70.78±0.7	79.13±1.8	0.47±0.5
w/o η	70.48±0.8	77.83±1.9	1.28±1.0

Table 5: Individual contributions of proposed modules.

- The contrastive training paradigm plays a key role in improving the prediction and explainability performance of SUNNY-GNN.

Takeaways

- We illustrate the importance of generating sufficient and necessary explanations for improving the performance of self-interpretable graph learning methods;
- We propose SUNNY-GNN, to generate such explanations while improving prediction performance empowered by contrastive loss;
- Future work: More complex graph mining scenarios such as heterogenous settings or multimodal settings.
- Code: https://github.com/SJTU-Quant/SUNNY-GNN
- Contact us: jialedeng@sjtu.edu.cn

Thanks!