## [Re] GNNInterpreter: A probabilistic generative model-level explanation for Graph Neural Networks

#### Ana-Maria Vasilcoiu, Batu Helvacioğlu, Thies Kersten, Thijs Stessen





## Outline

- Introduction and Background
- GNNInterpreter model
- Scope of reproducibility
- Methodology
- Results from reproducing the original paper
- Results beyond the original paper
- Discussion and Main Takeaways

#### Introduction & Background

- <u>Why</u> interpret GNNs?
  - GNNs demonstrate strong performance on graph-based tasks, but their complexity challenges interpretability, which is critical in high-stakes domains (e.g. chemistry or biomedicine).
- Existing state-of-the-art solution XGNN
  - Uses reinforcement learning to generate representative graphs for each class.
  - Limitations: requires domain-specific rules and can't handle continuous features.

### GNNInterpreter (1)

- Explanation method that works with any GNN model.
- Generates graphs that highlight the key patterns the GNN uses for its predictions.
- Learning objective with 2 goals:

$$\max_{G} L(G) = \max_{A,Z,X} L(A, Z, X) = \max_{A,Z,X} \phi_c(A, Z, X) + \mu \cdot \sin_{\cos}(\psi(A, Z, X), \bar{\psi}_c)$$
Maximize the likelihood of explanation graphs being predicted as the target class by the GNN
Confine explanation graph distribution within domain-specific boundaries

## GNNInterpreter (2)

- Continuous relaxation: converts discrete graph structures to continuous form for gradient-based optimization.
- Reparameterization trick: enables differentiable sampling over the relaxed graph.
- Regularization
  - L1 & L2: prevent overfitting and reduce gradient saturation.
  - Budget penalty: limits graph size for concise explanations.
  - Connectivity incentive: promote correlation.

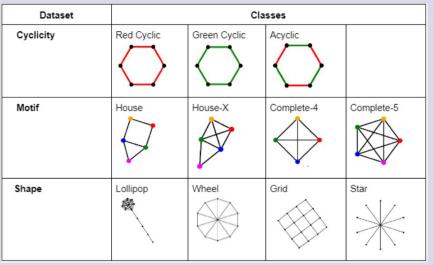
#### Scope of Reproducibility

- <u>Claim 1</u>: The explanations generated by GNNInterpreter are *faithful and realistic*. Additionally, GNNInterpreter *doesn't require domain-specific knowledge* to achieve that.
- <u>Claim 2</u>: GNNInterpreter is a general approach that performs well with *different types of node and edge features*.
- <u>Claim 3</u>: The explanations generated by GNNInterpreter are *more representative regarding the target class* compared to XGNN.
- <u>Claim 4</u>: The *time complexity* for training GNNInterpreter is *much lower* than for XGNN.

#### Methodology

#### Datasets

#### Synthetic datasets



#### Real-world datasets

| Dataset       | Classes           |                 |  |  |  |  |
|---------------|-------------------|-----------------|--|--|--|--|
| MUTAG         | Mutagen           | Nonmutagen      |  |  |  |  |
| Reddit Binary | Online-Discussion | Question-Answer |  |  |  |  |

• GNN architectures - GCN and NNConv

|                  |                 | Average of          | Best              | Worst             | Percentage of | Percentage of | Training |
|------------------|-----------------|---------------------|-------------------|-------------------|---------------|---------------|----------|
|                  |                 | all Models          | Model             | Model             | good models   | bad models    | time (s) |
| MUTAG            | Mutagen         | $0.987 \pm 0.100$   | -                 | -                 | -             | -             | 38.83    |
| (XGNN)           | Nonmutagen      | $0.999 \pm 0.002$   | -                 | -                 | -             | -             | 30.03    |
| MUTAG            | Mutagen         | $0.999 \pm 0.006$   | $1.000\pm0.000$   | $0.921{\pm}0.254$ | 1.00          | 0.00          | 0.79     |
| (GNNInterpreter) | Nonmutagen      | $0.943 \pm 0.068$   | $1.000\pm0.000$   | $0.330 \pm 0.429$ | 0.87          | 0.00          | 0.19     |
| Cyclicity        | Red Cyclic      | $0.926 \pm 0.0677$  | $1.000\pm0.000$   | $0.000\pm0.000$   | 0.84          | 0.02          |          |
| (GNNInterpreter) | Green Cyclic    | $0.665\pm0.372$     | $1.000\pm0.000$   | $0.101 \pm 0.290$ | 0.22          | 0             | 24.85    |
| (Givinterpreter) | Acyclic         | $0.525{\pm}0.120$   | $1.000\pm0.000$   | $0.000\pm0.000$   | 0.37          | 0.40          |          |
|                  | House           | $0.787 \pm 0.220$   | $0.991 \pm 0.006$ | $0.000\pm0.000$   | 0.41          | 0.08          |          |
| Motif            | House-X         | $0.276 {\pm} 0.085$ | $0.999 \pm 0.009$ | $0.000\pm0.000$   | 0.11          | 0.63          | 19.17    |
| (GNNInterpreter) | Complete-4      | $0.077 {\pm} 0.202$ | $0.995\pm0.052$   | $0.000\pm0.000$   | 0.06          | 0.91          | 13.17    |
|                  | Complete-5      | $0.131 {\pm} 0.034$ | $0.997 \pm 0.053$ | $0.000\pm0.000$   | 0.07          | 0.82          |          |
|                  | Lollipop        | $0.222 {\pm} 0.294$ | $0.43{\pm}0.374$  | $0.096\pm0.199$   | 0.00          | 0.01          |          |
| Shape            | Wheel           | $0.84 \pm 0.279$    | $0.997\pm0.056$   | $0.058 \pm 0.231$ | 0.45          | 0.02          | 23.48    |
| (GNNInterpreter) | Grid            | $0.782\pm0.327$     | $0.911 \pm 0.216$ | $0.612\pm0.408$   | 0.02          | 0.00          | 23.40    |
|                  | Star            | $1.000\pm0.001$     | $1.000\pm0.000$   | $0.987{\pm}0.109$ | 1.00          | 0.00          |          |
| Reddit-Binary    | Question-Answer | $0.8454 \pm 0.019$  | 0.89199           | 0.72159           | -             | -             | 25.774   |
| (GNNInterpreter) | Discussion      | $0.989 \pm 0.000$   | 0.9889            | 0.9889            | -             | -             | 20.114   |

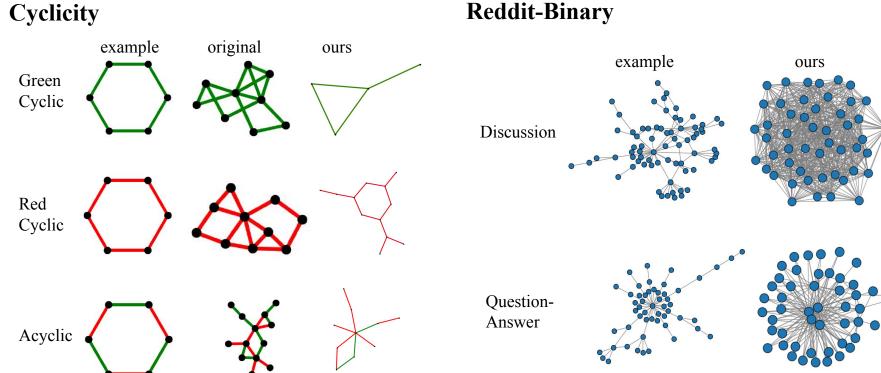
|                 | Average of   | D  | XX7 4   | D ( C  | D  | <b>m</b> · · .   |
|-----------------|--|--|---|--|--|--|
|                 |  | Best   | Worst   | Percentage of  | Percentage of  | Training   |
|                 | and the second s | Model  | Model   | good models  | bad models   | time $(s)$   |
| Mutagen         | $0.987 \pm 0.100$  | -  | -   | -  |  | 38.83  |
| Nonmutagen      | $0.999 \pm 0.002$  | -  | -   | -  | · -  | 00.00  |
| Mutagen         | $0.999\pm0.006$  | $1.000\pm0.000$  | $0.921{\pm}0.254$   | 1.00   | 0.00   | 0.79   |
| Nonmutagen      | $0.943\pm0.068$  | $1.000\pm0.000$  | $0.330 \pm 0.429$   | 0.87   | 0.00   | 0.19   |
| Red Cyclic      | $0.926 \pm 0.0677$   | $1.000\pm0.000$  | $0.000\pm0.000$   | 0.84   | 0.02   |  |
| Green Cyclic    | $0.665\pm0.372$  | $1.000\pm0.000$  | $0.101 \pm 0.290$   | 0.22   | 0  | 24.85  |
| Acyclic         | $0.525{\pm}0.120$  | $1.000\pm0.000$  | $0.000\pm0.000$   | 0.37   | 0.40   |  |
| House           | $0.787\pm0.220$  | $0.991 \pm 0.006$  | $0.000\pm0.000$   | 0.41   | 0.08   |  |
| House-X         | $0.276 {\pm} 0.085$  | $0.999 \pm 0.009$  | $0.000\pm0.000$   | 0.11   | 0.63   | 19.17  |
| Complete-4      | $0.077{\pm}0.202$  | $0.995 \pm 0.052$  | $0.000\pm0.000$   | 0.06   | 0.91   | 19.17  |
| Complete-5      | $0.131 {\pm} 0.034$  | $0.007 \pm 0.052$  | $0.000\pm0.000$   | 0.07   | 0.82   |  |
| Lollipop        | $0.222{\pm}0.294$  | $0.43{\pm}0.374$   | $0.096 \pm 0.199$   | 0.00   | 0.01   |  |
| Wheel           | $0.84 \pm 0.279$   | $0.997 \pm 0.000$  | $0.058\pm0.231$   | 0.45   | 0.02   | 09.40  |
| Grid            | $0.782\pm0.327$  | $0.911 \pm 0.216$  | $0.612 \pm 0.408$   | 0.02   | 0.00   | 23.48  |
| Star            | $1.000\pm0.001$  | $1.000\pm0.000$  | $0.987 {\pm} 0.109$   | 1.00   | 0.00   |  |
| Question-Answer | $0.8454 \pm 0.019$   | 0.89199  | 0.72159   | -  | -  | 95 774   |
| Discussion      | $0.989 \pm 0.000$  | 0.9889   | 0.9889  | -  | -  | 25.774   |
|                 | Nonmutagen<br>Mutagen<br>Nonmutagen<br>Red Cyclic<br>Green Cyclic<br>Acyclic<br>House<br>House-X<br>Complete-4<br>Complete-5<br>Lollipop<br>Wheel<br>Grid<br>Star<br>uestion-Answer  | all ModelsMutagen $0.987 \pm 0.100$ Nonmutagen $0.999 \pm 0.002$ Mutagen $0.999 \pm 0.006$ Mutagen $0.999 \pm 0.006$ Nonmutagen $0.943 \pm 0.068$ Red Cyclic $0.926 \pm 0.0677$ Green Cyclic $0.665 \pm 0.372$ Acyclic $0.525 \pm 0.120$ House $0.787 \pm 0.220$ House-X $0.276 \pm 0.085$ Complete-4 $0.077 \pm 0.202$ Complete-5 $0.131 \pm 0.034$ Lollipop $0.222 \pm 0.294$ Wheel $0.84 \pm 0.279$ Grid $0.782 \pm 0.327$ Star $1.000 \pm 0.001$ uestion-Answer $0.8454 \pm 0.019$ | all Models         Model           Mutagen         0.987 ± 0.100         -           Nonmutagen         0.999 ± 0.002         -           Mutagen         0.999 ± 0.002         1.000 ± 0.000           Mutagen         0.999 ± 0.066         1.000 ± 0.000           Nonmutagen         0.926 ± 0.0677         1.000 ± 0.000           Red Cyclic         0.665 ± 0.372         1.000 ± 0.000           Green Cyclic         0.665 ± 0.372         1.000 ± 0.000           Acyclic         0.525 ± 0.120         1.000 ± 0.000           House         0.787 ± 0.220         0.991 ± 0.006           House-X         0.276 ± 0.085         0.999 ± 0.009           Complete-4         0.077 ± 0.202         0.995 ± 0.052           Complete-5         0.131 ± 0.034         0.907 ± 0.053           Lollipop         0.222 ± 0.294         0.43 ± 0.374           Wheel         0.84 ± 0.279         0.911 ± 0.216           Star         1.000 ± 0.001         1.000 ± 0.000           uestion-Answer         0.8454 ± 0.019         0.89199 | all Models         Model         Model           Mutagen         0.987±0.100         -         -           Nonmutagen         0.999±0.002         -         -           Mutagen         0.999±0.006         1.000±0.000         0.921±0.254           Mutagen         0.943±0.068         1.000±0.000         0.330±0.429           Red Cyclic         0.926±0.0677         1.000±0.000         0.101±0.290           Green Cyclic         0.665±0.372         1.000±0.000         0.101±0.290           Acyclic         0.525±0.120         1.000±0.000         0.000±0.000           House         0.787±0.220         0.991±0.006         0.000±0.000           House-X         0.276±0.085         0.999±0.009         0.000±0.000           Complete-4         0.077±0.202         0.991±0.006         0.000±0.000           Complete-5         0.131±0.034         0.999±0.052         0.000±0.000           Lollipop         0.222±0.294         0.43±0.374         0.096±0.199           Wheel         0.84±0.279         0.997±0.050         0.058±0.231           Grid         0.782±0.327         0.911±0.216         0.5058±0.231           Grid         0.784±0.019         0.89199         0.72159 | all ModelsModelModelgood modelsMutagen $0.987 \pm 0.100$ Nonmutagen $0.999 \pm 0.002$ Mutagen $0.999 \pm 0.006$ $1.000 \pm 0.000$ $0.921 \pm 0.254$ $1.00$ Nonmutagen $0.943 \pm 0.068$ $1.000 \pm 0.000$ $0.330 \pm 0.429$ $0.87$ Red Cyclic $0.926 \pm 0.0677$ $1.000 \pm 0.000$ $0.000 \pm 0.000$ $0.84$ Green Cyclic $0.665 \pm 0.372$ $1.000 \pm 0.000$ $0.101 \pm 0.290$ $0.222$ Acyclic $0.525 \pm 0.120$ $1.000 \pm 0.000$ $0.000 \pm 0.000$ $0.37$ House $0.787 \pm 0.220$ $0.991 \pm 0.066$ $0.000 \pm 0.000$ $0.41$ House-X $0.276 \pm 0.085$ $0.999 \pm 0.052$ $0.000 \pm 0.000$ $0.11$ Complete-4 $0.077 \pm 0.202$ $0.995 \pm 0.052$ $0.000 \pm 0.000$ $0.06$ Complete-5 $0.131 \pm 0.034$ $0.097 \pm 0.053$ $0.000 \pm 0.000$ $0.07$ Lollipop $0.222 \pm 0.294$ $0.997 \pm 0.053$ $0.000 \pm 0.199$ $0.00$ Wheel $0.84 \pm 0.279$ $0.991 \pm 0.056$ $0.058 \pm 0.231$ $0.45$ Grid $0.782 \pm 0.327$ $0.911 \pm 0.216$ $0.612 \pm 0.408$ $0.02$ Star $1.000 \pm 0.01$ $1.000 \pm 0.000$ $0.987 \pm 0.109$ $1.00$ uestion-Answer $0.8454 \pm 0.019$ $0.89199$ $0.72159$ $-$ | all ModelsModelModelgood modelsbad modelsMutagen $0.987 \pm 0.100$ Nonmutagen $0.999 \pm 0.002$ 1.000 $\pm 0.000$ $0.921 \pm 0.254$ 1.000.00Mutagen $0.999 \pm 0.006$ $1.000 \pm 0.000$ $0.330 \pm 0.429$ $0.87$ 0.00Nonmutagen $0.943 \pm 0.068$ $1.000 \pm 0.000$ $0.330 \pm 0.429$ $0.87$ 0.00Red Cyclic $0.926 \pm 0.0677$ $1.000 \pm 0.000$ $0.000 \pm 0.000$ $0.844$ $0.02$ Green Cyclic $0.665 \pm 0.372$ $1.000 \pm 0.000$ $0.101 \pm 0.290$ $0.222$ $0$ Acyclic $0.525 \pm 0.120$ $1.000 \pm 0.000$ $0.000 \pm 0.000$ $0.37$ $0.40$ House $0.787 \pm 0.220$ $0.991 \pm 0.066$ $0.000 \pm 0.000$ $0.11$ $0.08$ House-X $0.276 \pm 0.085$ $0.999 \pm 0.052$ $0.000 \pm 0.000$ $0.11$ $0.63$ Complete-4 $0.077 \pm 0.202$ $0.995 \pm 0.052$ $-0.000 \pm 0.000$ $0.01$ $0.82$ Lollipop $0.222 \pm 0.294$ $0.995 \pm 0.052$ $-0.000 \pm 0.000$ $0.07$ $0.82$ Lollipop $0.222 \pm 0.294$ $0.991 \pm 0.006$ $-0.058 \pm 0.231$ $0.455$ $0.02$ Grid $0.782 \pm 0.327$ $0.911 \pm 0.216$ $0.612 \pm 0.408$ $0.02$ $0.00$ Star $1.000 \pm 0.001$ $0.89199$ $0.72159$ $ -$ |

|                 |  |   | 1   | -   |  |   |
|-----------------|--|---|---|---|--|---|
|                 |  | Best  | Worst   | Percentage of   | Percentage of  | Training  |
|                 | all Models   | Model   | Model   | good models   | bad models   | time $(s)$  |
| Mutagen         | $0.987 \pm 0.100$  | -   | -   | -   | -  | 38.83   |
| Nonmutagen      | $0.999 \pm 0.002$  | -   |   | -   | -  | 00.00   |
| Mutagen         | $0.999\pm0.006$  | $1.000\pm0.000$   | $0.921{\pm}0.254$   | 1.00  | 0.00   | 0.79  |
| Nonmutagen      | $0.943 \pm 0.068$  | $1.000\pm0.000$   | $0.330\pm0.429$   | 0.87  | 0.00   | 0.19  |
| Red Cyclic      | $0.926 \pm 0.0677$   | $1.000\pm0.000$   | $0.000\pm0.000$   | 0.84  | 0.02   |   |
| Green Cyclic    | $0.665\pm0.372$  | $1.000\pm0.000$   | $0.101 \pm 0.290$   | 0.22  | 0  | 24.85   |
| Acyclic         | $0.525{\pm}0.120$  | $1.000\pm0.000$   | $0.000\pm0.000$   | 0.37  | 0.40   |   |
| House           | $0.787\pm0.220$  | $0.991 \pm 0.006$   | $0.000 \pm 0.000$   | 0.41  | 0.08   | 19.17   |
| House-X         | $0.276 {\pm} 0.085$  | $0.999 \pm 0.009$   | $0.000\pm0.000$   | 0.11  | 0.63   |   |
| Complete-4      | $0.077{\pm}0.202$  | $0.995 \pm 0.052$   | $0.000\pm0.000$   | 0.06  | 0.91   | 19.17   |
| Complete-5      | $0.131{\pm}0.034$  | $0.997 \pm 0.053$   | $0.000\pm0.000$   | 0.07  | 0.82   |   |
| Lollipop        | $0.222{\pm}0.294$  | $0.43{\pm}0.374$  | $0.096 \pm 0.199$   | 0.00  | 0.01   |   |
| Wheel           | $0.84 \pm 0.279$   | $0.997 \pm 0.056$   | $0.058 \pm 0.231$   | 0.45  | 0.02   | 23.48   |
| Grid            | $0.782\pm0.327$  | $0.911 \pm 0.216$   | $0.612 \pm 0.408$   | 0.02  | 0.00   | 23.40   |
| Star            | $1.000\pm0.001$  | $1.000\pm0.000$   | $0.987{\pm}0.109$   | 1.00  | 0.00   |   |
| Question-Answer | $0.8454 \pm 0.019$   | 0.89199   | 0.72159   | -   | -  | 25.774  |
| Discussion      | $0.989 \pm 0.000$  | 0.9889  | 0.9889  | -   | -  | 20.114  |
|                 | Nonmutagen<br>Mutagen<br>Nonmutagen<br>Red Cyclic<br>Green Cyclic<br>Acyclic<br>House-X<br>House-X<br>Complete-4<br>Complete-5<br>Lollipop<br>Wheel<br>Grid<br>Star<br>Question-Answer | Nonmutagen $0.999 \pm 0.002$ Mutagen $0.999 \pm 0.006$ Nonmutagen $0.943 \pm 0.068$ Red Cyclic $0.926 \pm 0.0677$ Green Cyclic $0.665 \pm 0.372$ Acyclic $0.525 \pm 0.120$ House $0.787 \pm 0.220$ House-X $0.276 \pm 0.085$ Complete-4 $0.077 \pm 0.202$ Complete-5 $0.131 \pm 0.034$ Lollipop $0.222 \pm 0.294$ Wheel $0.84 \pm 0.279$ Grid $0.782 \pm 0.327$ Star $1.000 \pm 0.001$ Question-Answer $0.8454 \pm 0.019$ | all Models         Model           Mutagen         0.987±0.100         -           Nonmutagen         0.999±0.002         -           Mutagen         0.999±0.006         1.000±0.000           Nonmutagen         0.943±0.068         1.000±0.000           Nonmutagen         0.943±0.068         1.000±0.000           Red Cyclic         0.926±0.0677         1.000±0.000           Green Cyclic         0.665±0.372         1.000±0.000           Acyclic         0.525±0.120         1.000±0.000           House         0.787±0.220         0.991±0.006           House         0.276±0.085         0.999±0.009           Complete-4         0.077±0.202         0.995±0.052           Complete-5         0.131±0.034         0.997±0.056           Muheel         0.84±0.279         0.997±0.056           Grid         0.782±0.327         0.911±0.216           Star         1.000±0.001         1.000±0.000           Question-Answer         0.8454±0.019         0.89199 | all Models         Model         Model           Mutagen         0.987 ± 0.100         -         -           Nonmutagen         0.999 ± 0.002         I         I         -           Mutagen         0.999 ± 0.002         I.000 ± 0.000         I.921 ± 0.254           Nonmutagen         0.999 ± 0.068         I.000 ± 0.000         I.330 ± 0.429           Mutagen         0.943 ± 0.068         I.000 ± 0.000         I.330 ± 0.429           Red Cyclic         0.926 ± 0.0677         I.000 ± 0.000         I.000 ± 0.001           Green Cyclic         0.665 ± 0.372         I.000 ± 0.000         I.011 ± 0.290           Acyclic         0.525 ± 0.120         I.000 ± 0.000         I.000 ± 0.000           House         0.787 ± 0.220         0.991 ± 0.068         I.000 ± 0.000           House-X         0.276 ± 0.032         I.999 ± 0.029         I.000 ± 0.000           Complete-4         0.077 ± 0.202         I.995 ± 0.052         I.000 ± 0.000           Complete-5         0.131 ± 0.034         I.997 ± 0.53         I.000 ± 0.001           Lollipop         0.222 ± 0.294         0.43 ± 0.374         I.096 ± 0.199           Wheel         0.84 ± 0.279         I.997 ± 0.056         I.612 ± 0.408           Star         I.00 | all Models         Model         Model         good models           Mutagen         0.987 ± 0.100         -         -         -           Nonmutagen         0.999 ± 0.002         -         -         -           Mutagen         0.999 ± 0.002         1.000 ± 0.000         0.921 ± 0.254         1.00           Nonmutagen         0.943 ± 0.068         1.000 ± 0.000         0.330 ± 0.429         0.87           Red Cyclic         0.926 ± 0.0677         1.000 ± 0.000         0.000 ± 0.000         0.84           Green Cyclic         0.665 ± 0.372         1.000 ± 0.000         0.101 ± 0.290         0.222           Acyclic         0.525 ± 0.120         1.000 ± 0.000         0.000 ± 0.000         0.41           House         0.787 ± 0.220         0.991 ± 0.066         0.000 ± 0.000         0.41           House-X         0.276 ± 0.085         0.999 ± 0.052         0.000 ± 0.000         0.11           Complete-4         0.077 ± 0.202         0.995 ± 0.52         0.000 ± 0.000         0.06           Complete-5         0.131 ± 0.34         0.997 ± 0.53         0.000 ± 0.000         0.07           Lollipop         0.222 ± 0.294         0.43 ± 0.374         0.096 ± 0.199         0.00           Wheel         0.84 ± | all Models         Model         Model         good models         bad models           Mutagen         0.987±0.100         -         -         -         -           Nonmutagen         0.999±0.002         I         0         -         -         -           Mutagen         0.999±0.006         1.000±0.000         0.921±0.254         1.00         0.00           Nonmutagen         0.943±0.068         1.000±0.000         0.330±0.429         0.87         0.00           Red Cyclic         0.926±0.0677         1.000±0.000         0.000±0.000         0.84         0.02           Green Cyclic         0.665±0.372         1.000±0.000         0.000±0.000         0.37         0.40           House         0.787±0.220         0.991±0.006         0.000±0.000         0.41         0.08           House-X         0.276±0.085         0.999±0.002         0.000±0.000         0.11         0.63           Complete-4         0.077±0.202         0.995±0.052         0.000±0.000         0.01         0.02           Complete-5         0.131±0.034         0.997±0.056         0.000±0.000         0.07         0.822           Lollipop         0.22±0.294         0.43±0.374         0.096±0.199         0.00         0.0 |

|                  |                 | Average of          | Best              | Worst             | Percentage of | Percentage of | Training   |
|------------------|-----------------|---------------------|-------------------|-------------------|---------------|---------------|------------|
|                  |                 | all Models          | Model             | Model             | good models   | bad models    | time $(s)$ |
| MUTAG            | Mutagen         | $0.987 \pm 0.100$   | -                 | -                 | -             | -             | 38.83      |
| (XGNN)           | Nonmutagen      | $0.999 \pm 0.002$   | -                 | -                 | -             | -             | 00.00      |
| MUTAG            | Mutagen         | $0.999 \pm 0.006$   | $1.000\pm0.000$   | $0.921{\pm}0.254$ | 1.00          | 0.00          | 0.79       |
| (GNNInterpreter) | Nonmutagen      | $0.943\pm0.068$     | $1.000\pm0.000$   | $0.330 \pm 0.429$ | 0.87          | 0.00          | 0.19       |
| Cyclicity        | Red Cyclic      | $0.926 \pm 0.0677$  | $1.000\pm0.000$   | $0.000\pm0.000$   | 0.01          | 0.02          |            |
| (GNNInterpreter) | Green Cyclic    | $0.665\pm0.372$     | $1.000\pm0.000$   | $0.101 \pm 0.290$ | 0.22          | 0             | 24.85      |
| (Givinterpreter) | Acyclic         | $0.525{\pm}0.120$   | $1.000\pm0.000$   | $0.000\pm0.000$   | 0.37          | 0.40          |            |
|                  | House           | $0.787 \pm 0.220$   | $0.991 \pm 0.006$ | $0.000\pm0.000$   | 0.41          | 0.08          |            |
| Motif            | House-X         | $0.276 {\pm} 0.085$ | $0.999 \pm 0.009$ | $0.000\pm0.000$   | 0.11          | 0.63          | 19.17      |
| (GNNInterpreter) | Complete-4      | $0.077 {\pm} 0.202$ | $0.995\pm0.052$   | $0.000\pm0.000$   | 0.06          | 0.91          | 13.17      |
|                  | Complete-5      | $0.131{\pm}0.034$   | $0.997 \pm 0.053$ | $0.000\pm0.000$   | 0.07          | 0.82          |            |
|                  | Lollipop        | $0.222 {\pm} 0.294$ | $0.43{\pm}0.374$  | $0.096\pm0.199$   | 0.00          | 0.01          |            |
| Shape            | Wheel           | $0.84 \pm 0.279$    | $0.997\pm0.056$   | $0.058 \pm 0.231$ | 0.45          | 0.02          | 23.48      |
| (GNNInterpreter) | Grid            | $0.782\pm0.327$     | $0.911 \pm 0.216$ | $0.612\pm0.408$   | 0.02          | 0.00          | 23.40      |
|                  | Star            | $1.000\pm0.001$     | $1.000\pm0.000$   | $0.987{\pm}0.109$ | 1.00          | 0.00          |            |
| Reddit-Binary    | Question-Answer | $0.8454 \pm 0.019$  | 0.89199           | 0.72159           | -             |               | 25.774     |
| (GNNInterpreter) | Discussion      | $0.989 \pm 0.000$   | 0.9889            | 0.9889            | -             | -             | 20.114     |

|                   |                 | Average of          | Best              | Worst             | Percentage of | Percentage of  | Training |
|-------------------|-----------------|---------------------|-------------------|-------------------|---------------|----------------|----------|
|                   |                 | all Models          | Model             | Model             | good models   | bad models     | time (s) |
| MUTAG             | Mutagen         | $0.987 \pm 0.100$   | -                 | -                 | -             | -              | 38.83    |
| (XGNN)            | Nonmutagen      | $0.999 \pm 0.002$   | -                 | -                 | -             | -              | 30.03    |
| MUTAG             | Mutagen         | $0.999 \pm 0.006$   | $1.000\pm0.000$   | $0.921{\pm}0.254$ | 1.00          | 0.00           | 0.79     |
| (GNNInterpreter)  | Nonmutagen      | $0.943 \pm 0.068$   | $1.000\pm0.000$   | $0.330 \pm 0.429$ | 0.87          | 0.00           | 0.19     |
| Cyclicity         | Red Cyclic      | $0.926 \pm 0.0677$  | $1.000\pm0.000$   | $0.000\pm0.000$   | 0.84          | 0.02           |          |
| (GNNInterpreter)  | Green Cyclic    | $0.665\pm0.372$     | $1.000\pm0.000$   | $0.101 \pm 0.290$ | 0.22          | 0              | 24.85    |
| (diviniterpreter) | Acyclic         | $0.525{\pm}0.120$   | $1.000\pm0.000$   | $0.000\pm0.000$   | 0.37          | 0.40           |          |
|                   | House           | $0.787 \pm 0.220$   | $0.991 \pm 0.006$ | $0.000\pm0.000$   | 0.41          | 0.08           |          |
| Motif             | House-X         | $0.276{\pm}0.085$   | $0.999 \pm 0.009$ | $0.000\pm0.000$   | 0.11          | 0.63           | 19.17    |
| (GNNInterpreter)  | Complete-4      | $0.077 {\pm} 0.202$ | $0.995\pm0.052$   | $0.000\pm0.000$   | 0.06          | 0.91           | 1.3.17   |
|                   | Complete-5      | $0.131 {\pm} 0.034$ | $0.997 \pm 0.053$ | $0.000\pm0.000$   | 0.07          | 0.82           |          |
|                   | Lollipop        | $0.222{\pm}0.294$   | $0.43{\pm}0.374$  | $0.096\pm0.199$   | 0.00          | 0.01           |          |
| Shape             | Wheel           | $0.84 \pm 0.279$    | $0.997\pm0.056$   | $0.058 \pm 0.231$ | 0.45          | 0.02           | 23.48    |
| (GNNInterpreter)  | Grid            | $0.782\pm0.327$     | $0.911 \pm 0.216$ | $0.612\pm0.408$   | 0.02          | 0.00           | 25.40    |
| ** 400 200        | Star            | $1.000\pm0.001$     | $1.000\pm0.000$   | $0.987{\pm}0.109$ | 1.00          | 0.00           |          |
| Reddit-Binary     | Question-Answer | $0.8454 \pm 0.019$  | 0.89199           | 0.72159           | -             | 1. <del></del> | 25.774   |
| (GNNInterpreter)  | Discussion      | $0.989\pm0.000$     | 0.9889            | 0.9889            | -             | -              | 20.114   |

#### Results - Qualitative (1)

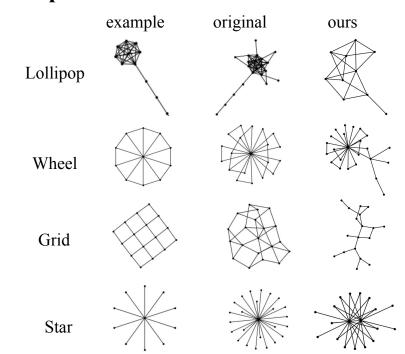


**Reddit-Binary** 

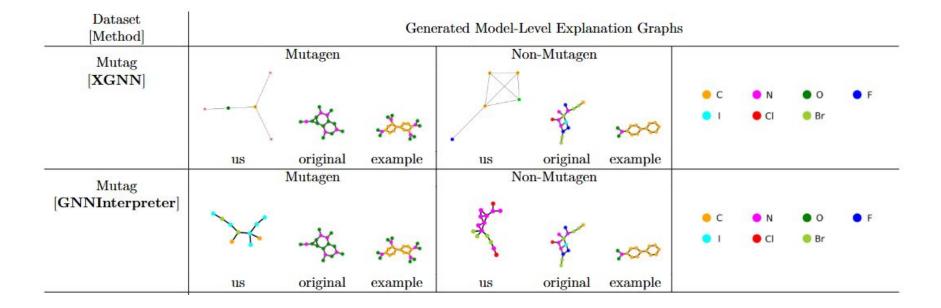
#### Results - Qualitative (1)

Motif example original ours House House-X Complete-4 Complete-5

Shape

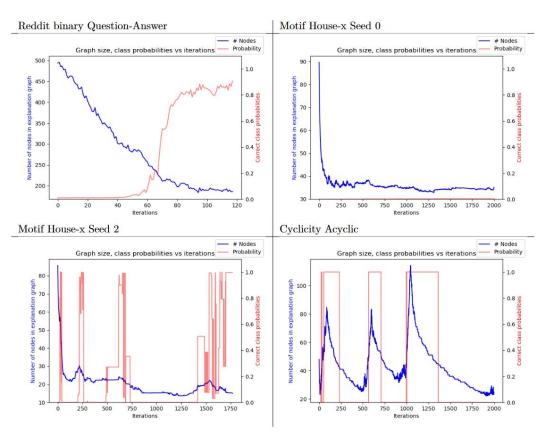


### Results - Qualitative (3)



Qualitative comparison on the Mutag dataset between XGNN and GNNInterpreter.

## Analysis of Training Instability (1)

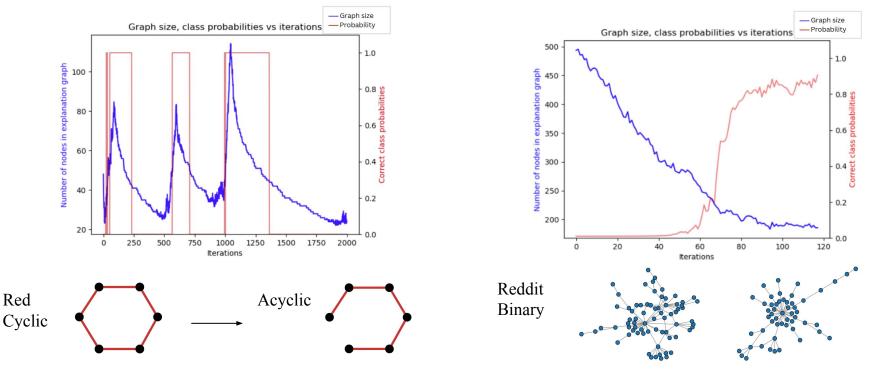


#### 4 scenarios (Top-Left to Bottom-Right):

- Expected behaviour (decreasing graph size and increasing correct class probability)
- Never converging
- Convergence
- Non-convergence

#### Analysis of Training Instability (2)

Main Reason - Discrete Behavior in Loss Despite Continuous Graph Relaxation



#### **Discussion and Main Takeaways**

- **Performance**: GNNInterpreter works with different types of node and edge features and can produce realistic explanations. However, its performance is inconsistent across datasets and highly sensitive to seed initializations and hyperparameters.
- Faithfulness and Reliability: Good quantitative results don't always translate to faithful or realistic explanation graphs.
- **Comparison to XGNN**: Explanation graphs are generally on-par, but GNNInterpreter has a lower time complexity. However, the time required for hyperparameter tuning and initialization can offset this advantage in practice.
- Graph size and complexity: GNNInterpreter performs best on large graphs, but experiences training instability on small graphs and highly specific structures.





# Thank you!

#### Questions?

Email us at: ana-maria.vasilcoiu@student.uva.nl