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

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

NEURAL INFORMATION
PROCESSING SYSTEMS

UNIVERSITEIT
VAN AMSTERDAM

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

- Why 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:

$$\max_{G} L(G) = \max_{A,Z,X} L(A,Z,X) = \max_{A,Z,X} \phi_c(A,Z,X) + \mu \cdot \mathrm{sim}_{\cos}(\psi(A,Z,X), \bar{\psi}_c)$$

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.

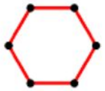
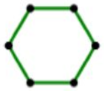
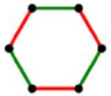
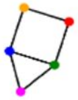
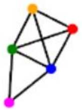
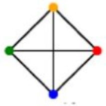
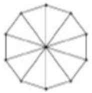
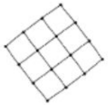
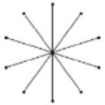
Scope of Reproducibility

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

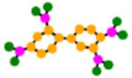
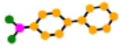
Methodology

- **Datasets**

Synthetic datasets

Dataset	Classes			
Cyclicity	Red Cyclic 	Green Cyclic 	Acyclic 	
Motif	House 	House-X 	Complete-4 	Complete-5 
Shape	Lollipop 	Wheel 	Grid 	Star 

Real-world datasets

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

- **GNN architectures** - GCN and NNConv

Results - Quantitative

		Average of all Models	Best Model	Worst Model	Percentage of good models	Percentage of bad models	Training time (s)
MUTAG (XGNN)	Mutagen	0.987 ± 0.100	-	-	-	-	38.83
	Nonmutagen	0.999 ± 0.002	-	-	-	-	
MUTAG (GNNInterpreter)	Mutagen	0.999 ± 0.006	1.000 ± 0.000	$\mathbf{0.921 \pm 0.254}$	1.00	0.00	0.79
	Nonmutagen	0.943 ± 0.068	1.000 ± 0.000	0.330 ± 0.429	0.87	0.00	
Cyclicity (GNNInterpreter)	Red Cyclic	0.926 ± 0.0677	1.000 ± 0.000	0.000 ± 0.000	0.84	0.02	24.85
	Green Cyclic	0.665 ± 0.372	1.000 ± 0.000	0.101 ± 0.290	**0.22**	0	
	Acyclic	$\mathbf{0.525 \pm 0.120}$	1.000 ± 0.000	0.000 ± 0.000	0.37	0.40	
Motif (GNNInterpreter)	House	0.787 ± 0.220	0.991 ± 0.006	0.000 ± 0.000	0.41	0.08	19.17
	House-X	$\mathbf{0.276 \pm 0.085}$	0.999 ± 0.009	0.000 ± 0.000	**0.11**	0.63	
	Complete-4	$\mathbf{0.077 \pm 0.202}$	0.995 ± 0.052	0.000 ± 0.000	**0.06**	**0.91**	
	Complete-5	$\mathbf{0.131 \pm 0.034}$	0.997 ± 0.053	0.000 ± 0.000	**0.07**	0.82	
Shape (GNNInterpreter)	Lollipop	$\mathbf{0.222 \pm 0.294}$	$\mathbf{0.43 \pm 0.374}$	0.096 ± 0.199	**0.00**	0.01	23.48
	Wheel	0.84 ± 0.279	0.997 ± 0.056	0.058 ± 0.231	0.45	0.02	
	Grid	0.782 ± 0.327	0.911 ± 0.216	0.612 ± 0.408	**0.02**	0.00	
	Star	1.000 ± 0.001	1.000 ± 0.000	$\mathbf{0.987 \pm 0.109}$	1.00	0.00	
Reddit-Binary (GNNInterpreter)	Question-Answer	0.8454 ± 0.019	0.89199	0.72159	-	-	25.774
	Discussion	0.989 ± 0.000	0.9889	0.9889	-	-	

Average class probabilities of 1000 explanation graphs, further averaged over 100 different seeds.
Good model: >0.9, Bad model: <0.1 correct class probability

Results - Quantitative

		Average of all Models	Best Model	Worst Model	Percentage of good models	Percentage of bad models	Training time (s)
MUTAG (XGNN)	Mutagen	0.987 ± 0.100	-	-	-	-	38.83
	Nonmutagen	0.999 ± 0.002	-	-	-	-	
MUTAG (GNNInterpreter)	Mutagen	0.999 ± 0.006	1.000 ± 0.000	$\mathbf{0.921 \pm 0.254}$	1.00	0.00	0.79
	Nonmutagen	0.943 ± 0.068	1.000 ± 0.000	0.330 ± 0.429	0.87	0.00	
Cyclicity (GNNInterpreter)	Red Cyclic	0.926 ± 0.0677	1.000 ± 0.000	0.000 ± 0.000	0.84	0.02	24.85
	Green Cyclic	0.665 ± 0.372	1.000 ± 0.000	0.101 ± 0.290	**0.22**	0	
	Acyclic	$\mathbf{0.525 \pm 0.120}$	1.000 ± 0.000	0.000 ± 0.000	0.37	0.40	
Motif (GNNInterpreter)	House	0.787 ± 0.220	0.991 ± 0.006	0.000 ± 0.000	0.41	0.08	19.17
	House-X	$\mathbf{0.276 \pm 0.085}$	0.999 ± 0.009	0.000 ± 0.000	**0.11**	0.63	
	Complete-4	$\mathbf{0.077 \pm 0.202}$	0.995 ± 0.052	0.000 ± 0.000	**0.06**	**0.91**	
	Complete-5	$\mathbf{0.131 \pm 0.034}$	0.997 ± 0.053	0.000 ± 0.000	**0.07**	0.82	
Shape (GNNInterpreter)	Lollipop	$\mathbf{0.222 \pm 0.294}$	$\mathbf{0.43 \pm 0.374}$	0.096 ± 0.199	**0.00**	0.01	23.48
	Wheel	0.84 ± 0.279	0.997 ± 0.056	0.058 ± 0.231	0.45	0.02	
	Grid	0.782 ± 0.327	0.911 ± 0.216	0.612 ± 0.408	**0.02**	0.00	
	Star	1.000 ± 0.001	1.000 ± 0.000	$\mathbf{0.987 \pm 0.109}$	1.00	0.00	
Reddit-Binary (GNNInterpreter)	Question-Answer	0.8454 ± 0.019	0.89199	0.72159	-	-	25.774
	Discussion	0.989 ± 0.000	0.9889	0.9889	-	-	

Average class probabilities of 1000 explanation graphs, further averaged over 100 different seeds.
Good model: >0.9, Bad model: <0.1 correct class probability

Results - Quantitative

		Average of all Models	Best Model	Worst Model	Percentage of good models	Percentage of bad models	Training time (s)
MUTAG (XGNN)	Mutagen	0.987 ± 0.100	-	-	-	-	38.83
	Nonmutagen	0.999 ± 0.002	-	-	-	-	
MUTAG (GNNInterpreter)	Mutagen	0.999 ± 0.006	1.000 ± 0.000	**0.921±0.254**	1.00	0.00	0.79
	Nonmutagen	0.943 ± 0.068	1.000 ± 0.000	0.330 ± 0.429	0.87	0.00	
Cyclicity (GNNInterpreter)	Red Cyclic	0.926 ± 0.0677	1.000 ± 0.000	0.000 ± 0.000	0.84	0.02	24.85
	Green Cyclic	0.665 ± 0.372	1.000 ± 0.000	0.101 ± 0.290	**0.22**	0	
	Acyclic	**0.525±0.120**	1.000 ± 0.000	0.000 ± 0.000	0.37	0.40	
Motif (GNNInterpreter)	House	0.787 ± 0.220	0.991 ± 0.006	0.000 ± 0.000	0.41	0.08	19.17
	House-X	**0.276±0.085**	0.999 ± 0.009	0.000 ± 0.000	**0.11**	0.63	
	Complete-4	**0.077±0.202**	0.995 ± 0.052	0.000 ± 0.000	**0.06**	**0.91**	
	Complete-5	**0.131±0.034**	0.997 ± 0.053	0.000 ± 0.000	**0.07**	0.82	
Shape (GNNInterpreter)	Lollipop	**0.222±0.294**	**0.43±0.374**	0.096 ± 0.199	**0.00**	0.01	23.48
	Wheel	0.84 ± 0.279	0.997 ± 0.056	0.058 ± 0.231	0.45	0.02	
	Grid	0.782 ± 0.327	0.911 ± 0.216	0.612 ± 0.408	**0.02**	0.00	
	Star	1.000 ± 0.001	1.000 ± 0.000	**0.987±0.109**	1.00	0.00	
Reddit-Binary (GNNInterpreter)	Question-Answer	0.8454 ± 0.019	0.89199	0.72159	-	-	25.774
	Discussion	0.989 ± 0.000	0.9889	0.9889	-	-	

Average class probabilities of 1000 explanation graphs, further averaged over 100 different seeds.
Good model: >0.9, Bad model: <0.1 correct class probability

Results - Quantitative

		Average of all Models	Best Model	Worst Model	Percentage of good models	Percentage of bad models	Training time (s)
MUTAG (XGNN)	Mutagen	0.987 ± 0.100	-	-	-	-	38.83
	Nonmutagen	0.999 ± 0.002	-	-	-	-	
MUTAG (GNNInterpreter)	Mutagen	0.999 ± 0.006	1.000 ± 0.000	$\mathbf{0.921 \pm 0.254}$	1.00	0.00	0.79
	Nonmutagen	0.943 ± 0.068	1.000 ± 0.000	0.330 ± 0.429	0.87	0.00	
Cyclicity (GNNInterpreter)	Red Cyclic	0.926 ± 0.0677	1.000 ± 0.000	0.000 ± 0.000	0.84	0.02	24.85
	Green Cyclic	0.665 ± 0.372	1.000 ± 0.000	0.101 ± 0.290	**0.22**	0	
	Acyclic	$\mathbf{0.525 \pm 0.120}$	1.000 ± 0.000	0.000 ± 0.000	0.37	0.40	
Motif (GNNInterpreter)	House	0.787 ± 0.220	0.991 ± 0.006	0.000 ± 0.000	0.41	0.08	19.17
	House-X	$\mathbf{0.276 \pm 0.085}$	0.999 ± 0.009	0.000 ± 0.000	**0.11**	0.63	
	Complete-4	$\mathbf{0.077 \pm 0.202}$	0.995 ± 0.052	0.000 ± 0.000	**0.06**	**0.91**	
	Complete-5	$\mathbf{0.131 \pm 0.034}$	0.997 ± 0.053	0.000 ± 0.000	**0.07**	0.82	
Shape (GNNInterpreter)	Lollipop	$\mathbf{0.222 \pm 0.294}$	$\mathbf{0.43 \pm 0.374}$	0.096 ± 0.199	**0.00**	0.01	23.48
	Wheel	0.84 ± 0.279	0.997 ± 0.056	0.058 ± 0.231	0.45	0.02	
	Grid	0.782 ± 0.327	0.911 ± 0.216	0.612 ± 0.408	**0.02**	0.00	
	Star	1.000 ± 0.001	1.000 ± 0.000	$\mathbf{0.987 \pm 0.109}$	1.00	0.00	
Reddit-Binary (GNNInterpreter)	Question-Answer	0.8454 ± 0.019	0.89199	0.72159	-	-	25.774
	Discussion	0.989 ± 0.000	0.9889	0.9889	-	-	

Average class probabilities of 1000 explanation graphs, further averaged over 100 different seeds.
Good model: >0.9, Bad model: <0.1 correct class probability

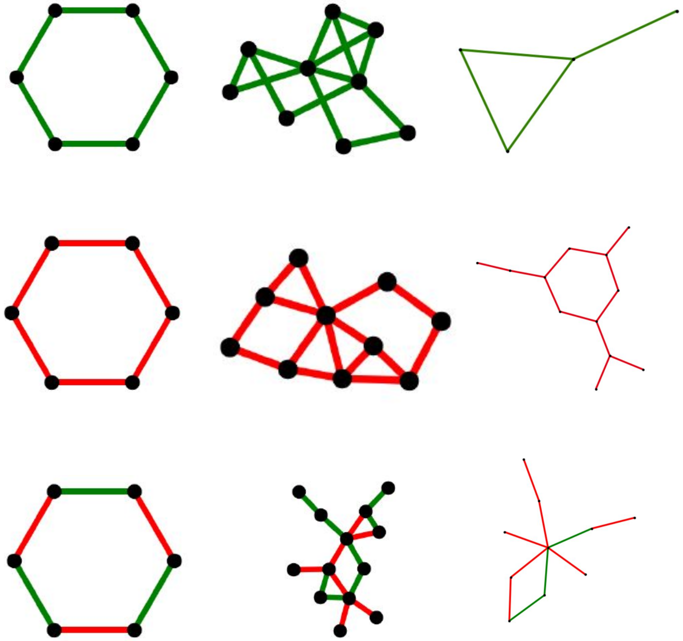
Results - Quantitative

		Average of all Models	Best Model	Worst Model	Percentage of good models	Percentage of bad models	Training time (s)
MUTAG (XGNN)	Mutagen	0.987 ± 0.100	-	-	-	-	38.83
	Nonmutagen	0.999 ± 0.002	-	-	-	-	
MUTAG (GNNInterpreter)	Mutagen	0.999 ± 0.006	1.000 ± 0.000	$\mathbf{0.921 \pm 0.254}$	1.00	0.00	0.79
	Nonmutagen	0.943 ± 0.068	1.000 ± 0.000	0.330 ± 0.429	0.87	0.00	
Cyclicity (GNNInterpreter)	Red Cyclic	0.926 ± 0.0677	1.000 ± 0.000	0.000 ± 0.000	0.84	0.02	24.85
	Green Cyclic	0.665 ± 0.372	1.000 ± 0.000	0.101 ± 0.290	**0.22**	0	
	Acyclic	$\mathbf{0.525 \pm 0.120}$	1.000 ± 0.000	0.000 ± 0.000	0.37	0.40	
Motif (GNNInterpreter)	House	0.787 ± 0.220	0.991 ± 0.006	0.000 ± 0.000	0.41	0.08	19.17
	House-X	$\mathbf{0.276 \pm 0.085}$	0.999 ± 0.009	0.000 ± 0.000	**0.11**	0.63	
	Complete-4	$\mathbf{0.077 \pm 0.202}$	0.995 ± 0.052	0.000 ± 0.000	**0.06**	**0.91**	
	Complete-5	$\mathbf{0.131 \pm 0.034}$	0.997 ± 0.053	0.000 ± 0.000	**0.07**	0.82	
Shape (GNNInterpreter)	Lollipop	$\mathbf{0.222 \pm 0.294}$	$\mathbf{0.43 \pm 0.374}$	0.096 ± 0.199	**0.00**	0.01	23.48
	Wheel	0.84 ± 0.279	0.997 ± 0.056	0.058 ± 0.231	0.45	0.02	
	Grid	0.782 ± 0.327	0.911 ± 0.216	0.612 ± 0.408	**0.02**	0.00	
	Star	1.000 ± 0.001	1.000 ± 0.000	$\mathbf{0.987 \pm 0.109}$	1.00	0.00	
Reddit-Binary (GNNInterpreter)	Question-Answer	0.8454 ± 0.019	0.89199	0.72159	-	-	25.774
	Discussion	0.989 ± 0.000	0.9889	0.9889	-	-	

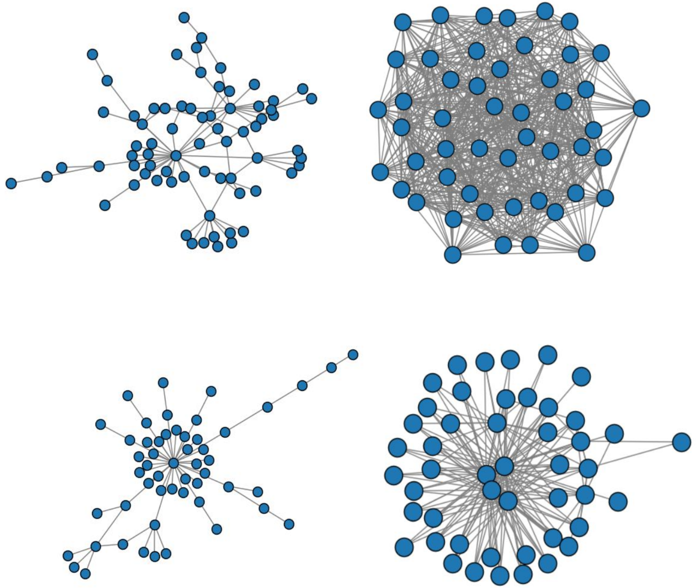
Average class probabilities of 1000 explanation graphs, further averaged over 100 different seeds.
Good model: >0.9, Bad model: <0.1 correct class probability

Results - Qualitative (1)

Cyclicity

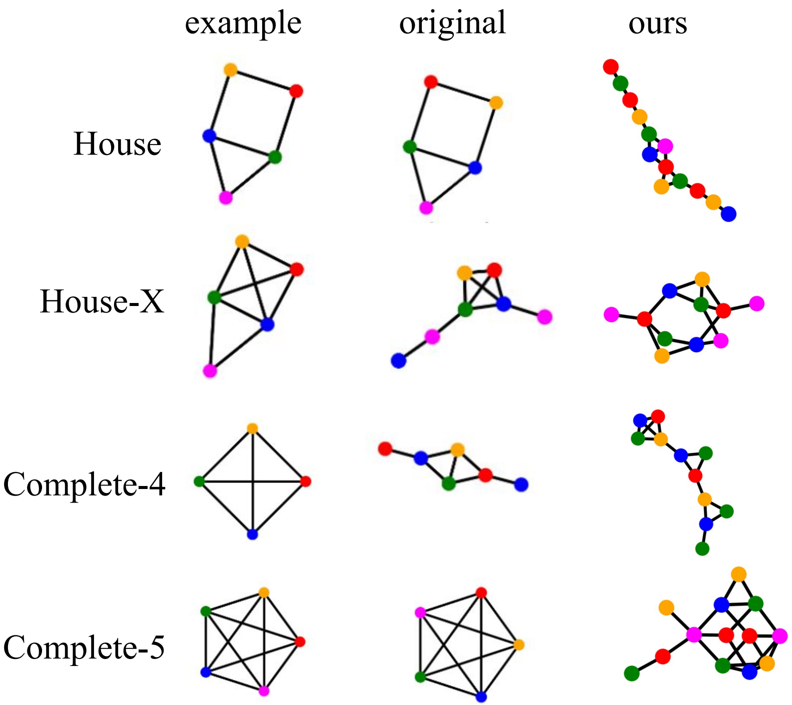


Reddit-Binary

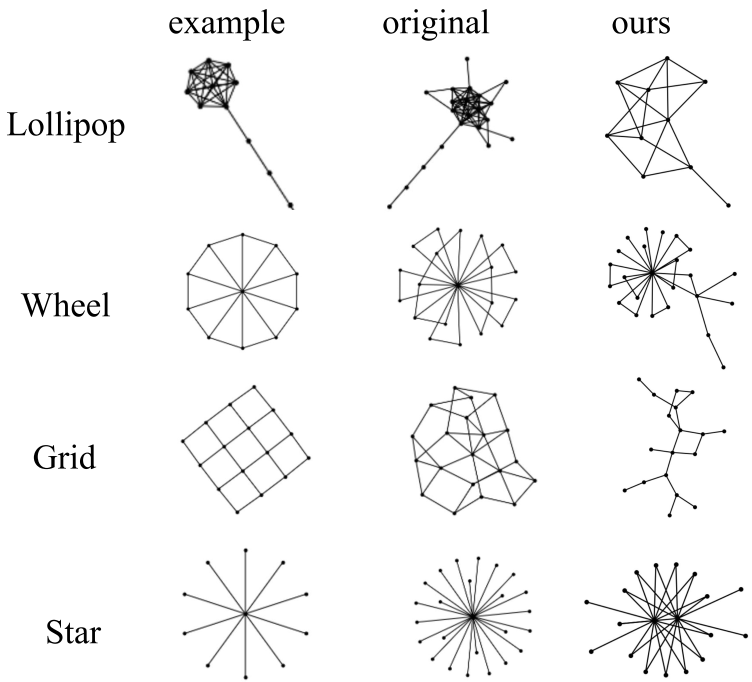


Results - Qualitative (1)

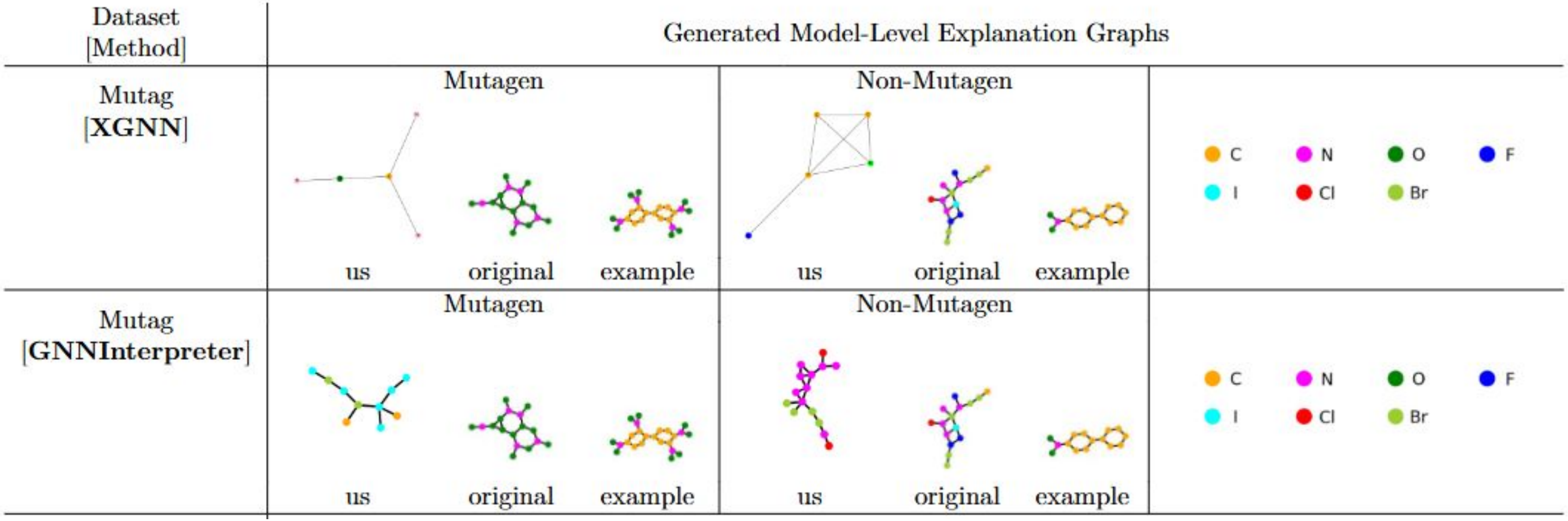
Motif



Shape

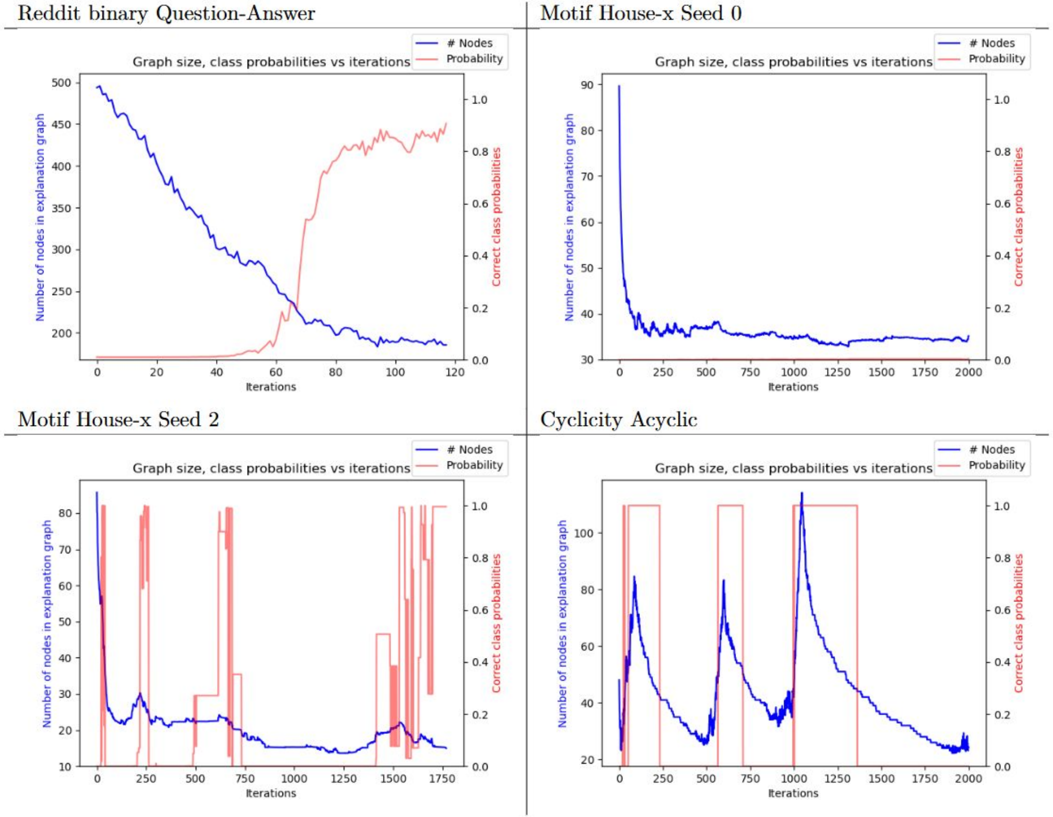


Results - Qualitative (3)



Qualitative comparison on the Mutag dataset between XGNN and GNNInterpreter.

Analysis of Training Instability (1)

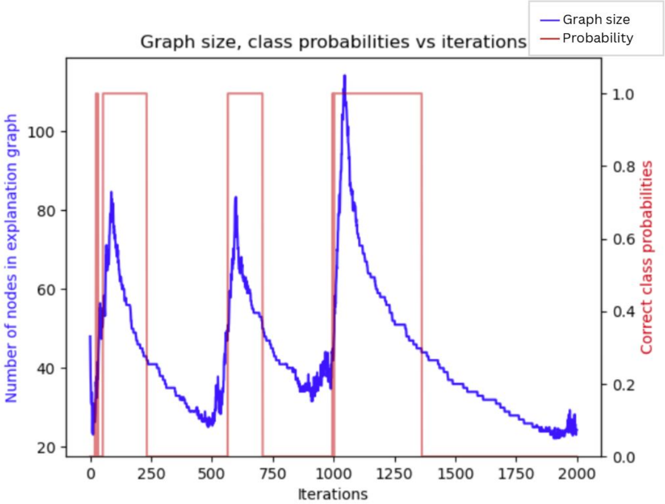


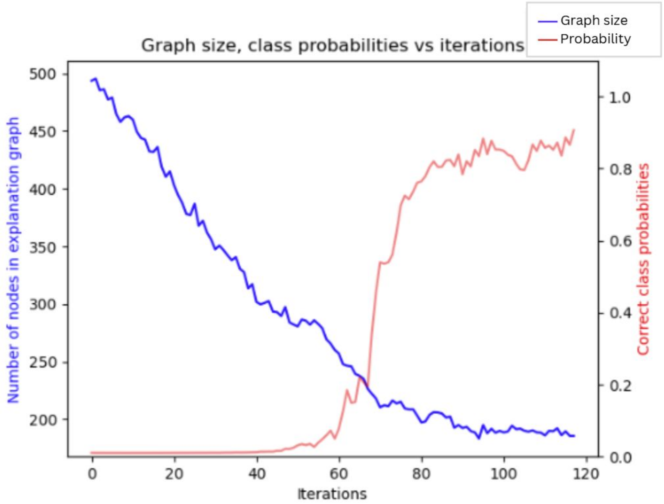
4 scenarios (Top-Left to Bottom-Right):

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

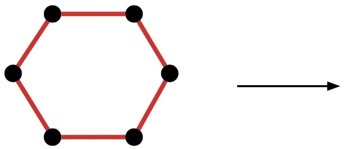
Analysis of Training Instability (2)

Main Reason - Discrete Behavior in Loss Despite Continuous Graph Relaxation

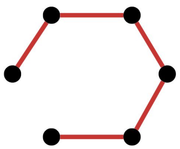




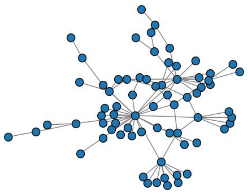
Red Cyclic

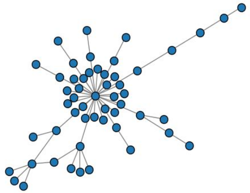


Acyclic



Reddit Binary





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.

NEURAL INFORMATION PROCESSING SYSTEMS

UNIVERSITEIT VAN AMSTERDAM

Thank you!

Questions?

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