D4Explainer:
In-Distribution GNN Explanations via Discrete Denoising Diffusion

Daan Brugmans
Marieke van Vreeswijk

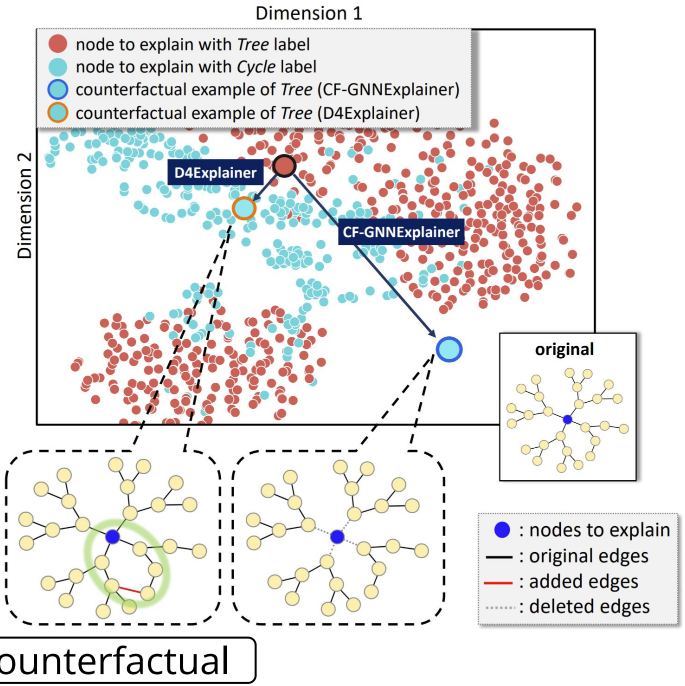
Radboud University

Understanding why GNNs make certain decisions is crucial for trusting their real-world use.

- **Motivation behind explaining GNNs for classification tasks**
  - High-stake applications such as healthcare
    - Graph with drug-protein interactions
      - Do we apply this drug as treatment?
  - Understandable → trust
  - Domain-specific rules
    - Molecule generation
    - Follow rules → remain in-distribution
- **Main objectives**
  - Class transitions
  - Class patterns

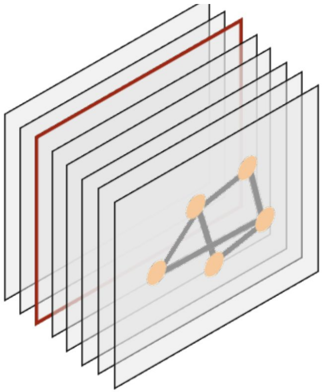
Counterfactual explanations can produce out-of-distribution graphs

- **Shoulder of Giants class transitions**
  - Challenging: Discrete structure of graph vs images
    - No smooth transitions
  - Class transition → counterfactual
  - Counterfactual explanation methods
    - Minimal modification to other class
      - Edge deletion (most)
      - Edge addition (CLEAR)
    - Out-of-distribution
      - Random guess
      - Reliability



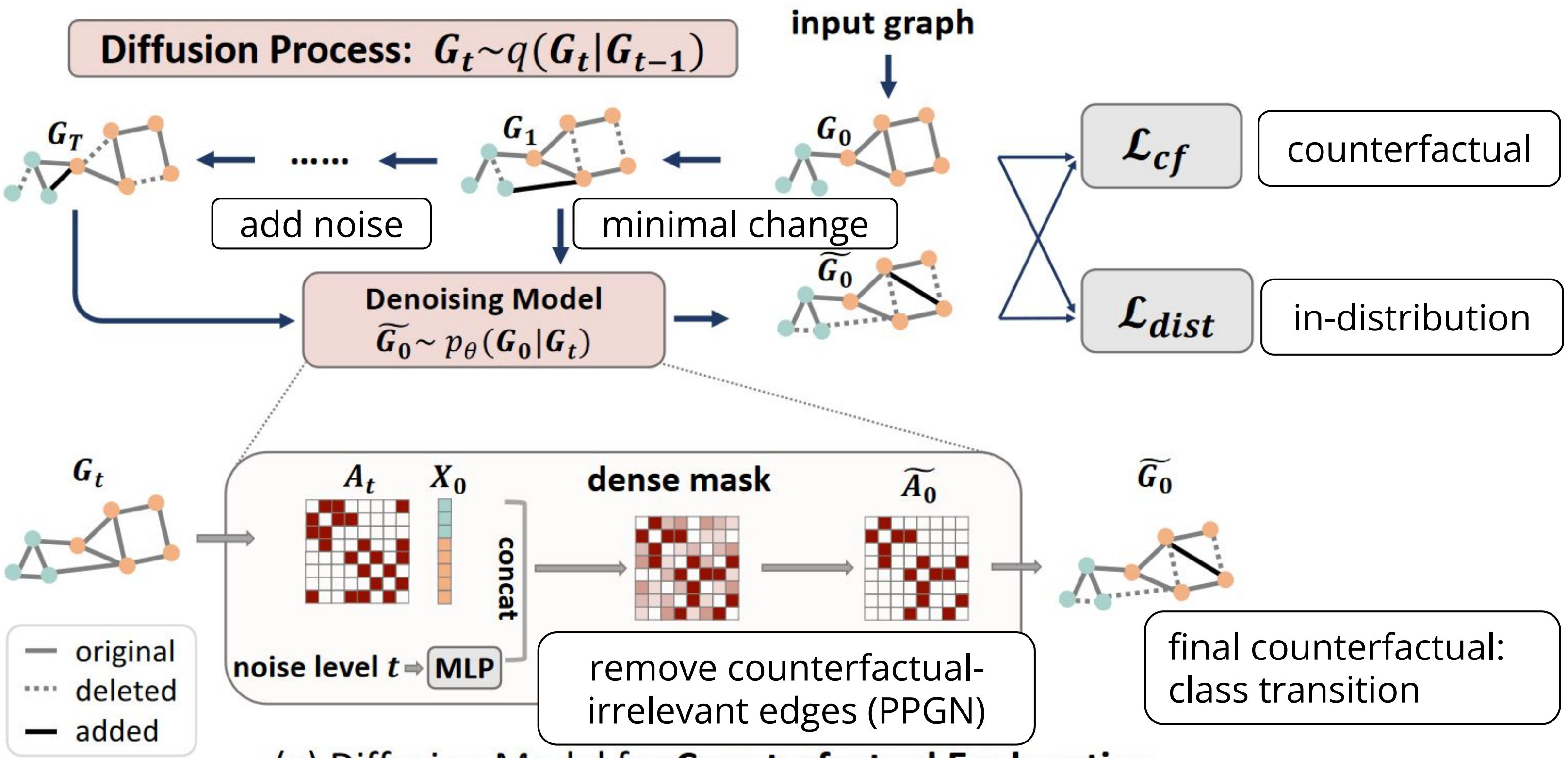
Model-level explanations provide class patterns

- **Shoulder of Giants class patterns**
  - Generation-based explanations
    - How is a specific prediction generated?
    - Instance-level
      - Generate one graph
      - Decision-making one graph
      - GNNExplainer
    - Model-level explanations
      - Generate multiple graphs
      - Capture class patterns
        - representative for class

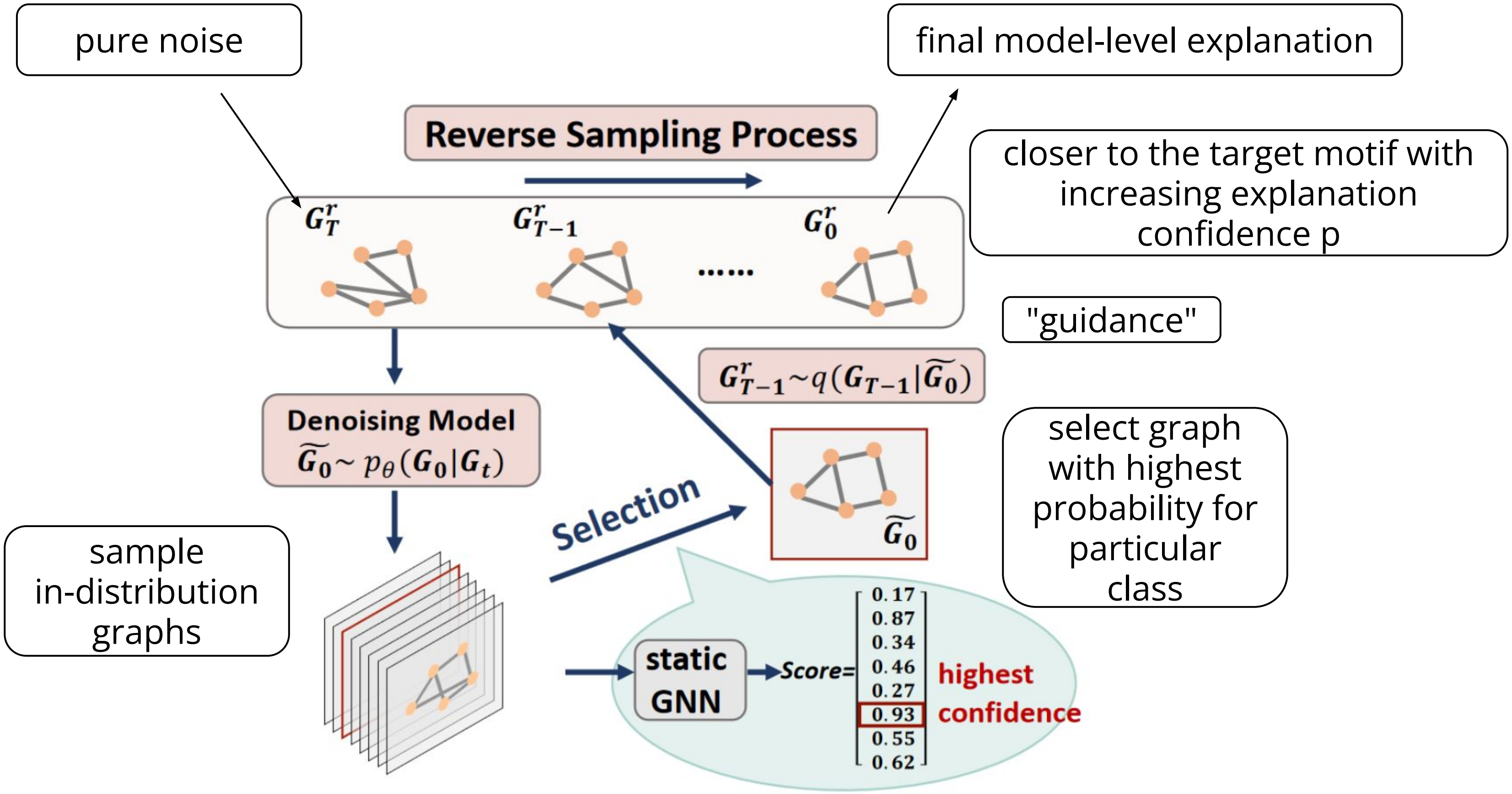


D4Explainer:
In-Distribution GNN Explanations via Discrete Denoising Diffusion

- In-distribution counterfactual explanations representing class transitions
- Model-level explanations capturing class patterns
- Generation-based approach utilizing a diffusion model



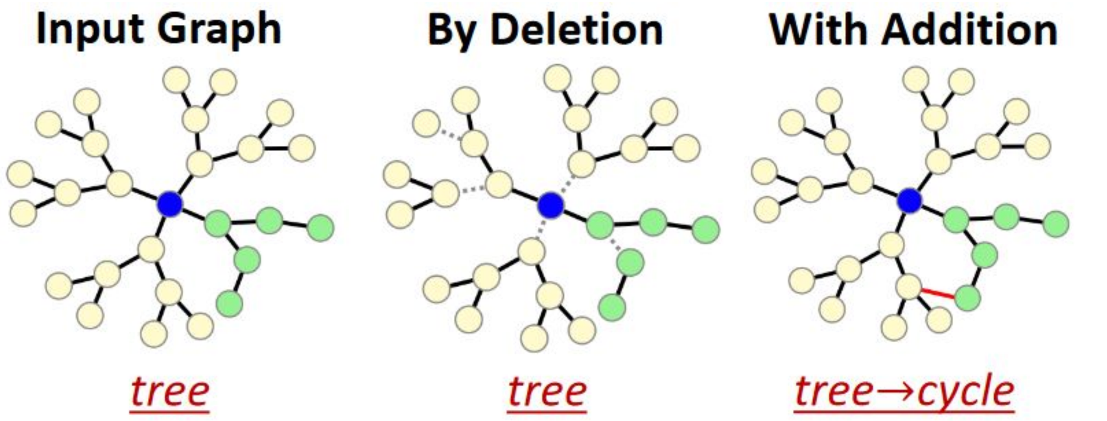
(a) Diffusion Model for **Counterfactual Explanation**

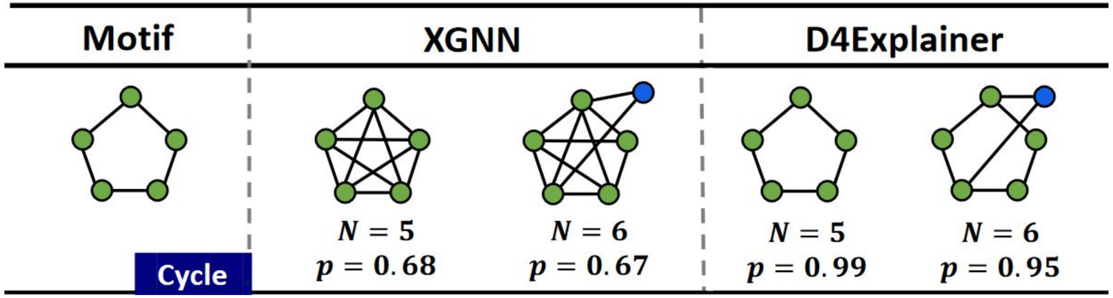


(b) Reverse Sampling for **Model-level Explanation**

Assessing D4Explainer

- **Experiments**
  - 8 datasets
  - Baselines
  - Counterfactual & Model-level
  - Diverse & Robust
- **Quantitative**
  - Consistently best
  - Counterfactual
    - Accuracy
    - Fidelity
    - MMD
  - Model-level
    - Prediction confidence
    - Density
- **Qualitative** →





Wrapping Up

- **Reproducibility**
  - Code is public
  - Hyperparameters in paper
- **Impact**
  - GNN decision-making transparency
    - Incomplete
- **Takeaway**
  - Counterfactual & model-level GNN explanations
  - In-distribution, robust & diverse
  - Based on the diffusion process

github.com/Graph-and-Geometric-Learning/D4Explainer

Supplemental Slides

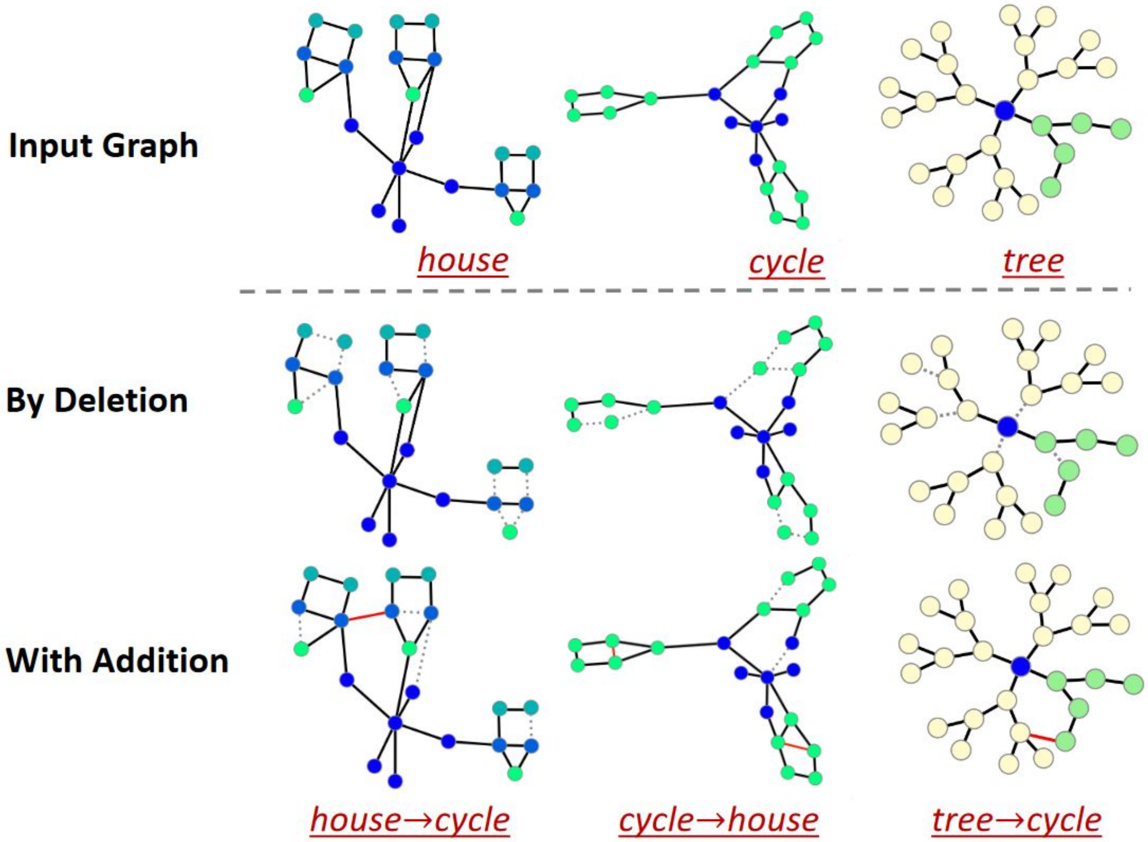
Counterfactual Evaluation (1)

Models	BA-Shapes CF-ACC	BA-Shapes FID	Tree-Cycle CF-ACC	Tree-Cycle FID	Tree-Grids CF-ACC	Tree-Grids FID	Cornell CF-ACC	Cornell FID	BA-3Motif CF-ACC	BA-3Motif FID	Mutag CF-ACC	Mutag FID	BBBP CF-ACC	BBBP FID	NCI1 CF-ACC	NCI1 FID
Random	0.251	0.261	0.260	0.281	0.337	0.375	0.138	0.172	0.404	0.452	0.192	0.256	0.073	0.113	0.288	0.352
GNNExplainer	0.473	0.444	0.652	0.580	0.672	0.622	0.075	0.120	0.250	0.253	0.450	0.449	0.212	0.241	0.375	0.443
SAExplainer	0.773	0.773	0.405	0.408	0.547	0.542	0.199	0.241	0.474	0.500	0.300	0.338	0.110	0.133	0.421	0.446
GradCam	0.552	0.570	0.637	0.613	0.590	0.578	0.138	0.189	0.459	0.495	0.202	0.250	0.274	0.301	0.467	0.488
IGExplainer	0.208	0.240	0.198	0.226	0.308	0.372	0.233	0.281	0.440	0.474	0.231	0.280	0.159	0.183	0.347	0.389
PGExplainer	0.361	0.357	0.353	0.322	0.293	0.340	0.128	0.204	0.320	0.323	0.208	0.313	0.233	0.282	0.338	0.366
PGMExplainer	0.208	0.210	0.242	0.214	0.128	0.237	0.206	0.274	0.212	0.213	0.128	0.251	0.105	0.154	0.348	0.390
CXPlain	0.125	0.168	0.245	0.220	0.222	0.274	0.132	0.180	0.235	0.239	0.187	0.305	0.067	0.131	0.489	0.484
CF-GNNExplainer	0.773	0.728	0.812	0.718	0.537	0.527	0.328	0.297	0.302	0.304	**0.797**	**0.751**	0.623	0.632	0.715	0.674
D4Explainer	**0.838**	**0.828**	**0.917**	**0.862**	**0.905**	**0.832**	**0.623**	**0.559**	**0.912**	**0.922**	0.765	0.675	**0.781**	**0.739**	**0.737**	**0.690**

Counterfactual Evaluation (2)

Models	Mutag				BBBP				NCI1			
	Deg.	Clus.	Spec.	Sum.	Deg.	Clus.	Spec.	Sum.	Deg.	Clus.	Spec.	Sum.
RamdomCaster	0.1593	0.0247	0.0417	0.2257	0.1693	0.0072	0.0397	0.2162	0.1847	1.9769	0.0404	2.2020
GNNExplainer	0.1614	0.0002	0.0409	0.2025	0.1615	0.0002	0.0395	0.2012	0.1577	0.0005	0.0405	0.1987
SAExplainer	**0.0940**	0.0032	0.0412	**0.1384**	0.1594	0.0032	0.0402	0.2028	0.189	0.0002	0.0408	0.2300
GradCam	0.1122	0.0083	0.0416	0.1621	0.0699	0.0026	0.0384	0.1109	0.1638	0.0003	0.0404	0.2045
IGExplainer	0.1292	**0.0000**	0.0411	0.1703	0.0908	**0.0000**	0.0394	0.1302	0.4288	0.0002	0.0398	0.4688
PGExplainer	0.1475	0.0002	0.0418	0.1895	0.2014	0.0018	0.0403	0.2435	0.1937	**0.0000**	0.0396	0.2333
PGMExplainer	0.1800	0.0002	0.0419	0.2221	0.1916	0.0003	0.0403	0.2322	0.2199	**0.0000**	0.0404	0.2603
CXPlain	0.1734	1.2706	0.0417	1.4857	0.1768	0.0001	0.0394	0.2163	0.1629	0.0001	0.0404	0.2034
CF-GNNExplainer	0.1172	**0.0000**	0.0380	0.1552	0.0870	0.0001	0.0393	0.1264	0.1224	0.0001	0.0404	0.1629
D4Explainer	0.1172	**0.0000**	**0.0244**	0.1416	**0.0530**	**0.0000**	**0.0331**	**0.0861**	**0.1006**	**0.0000**	**0.0353**	**0.1359**

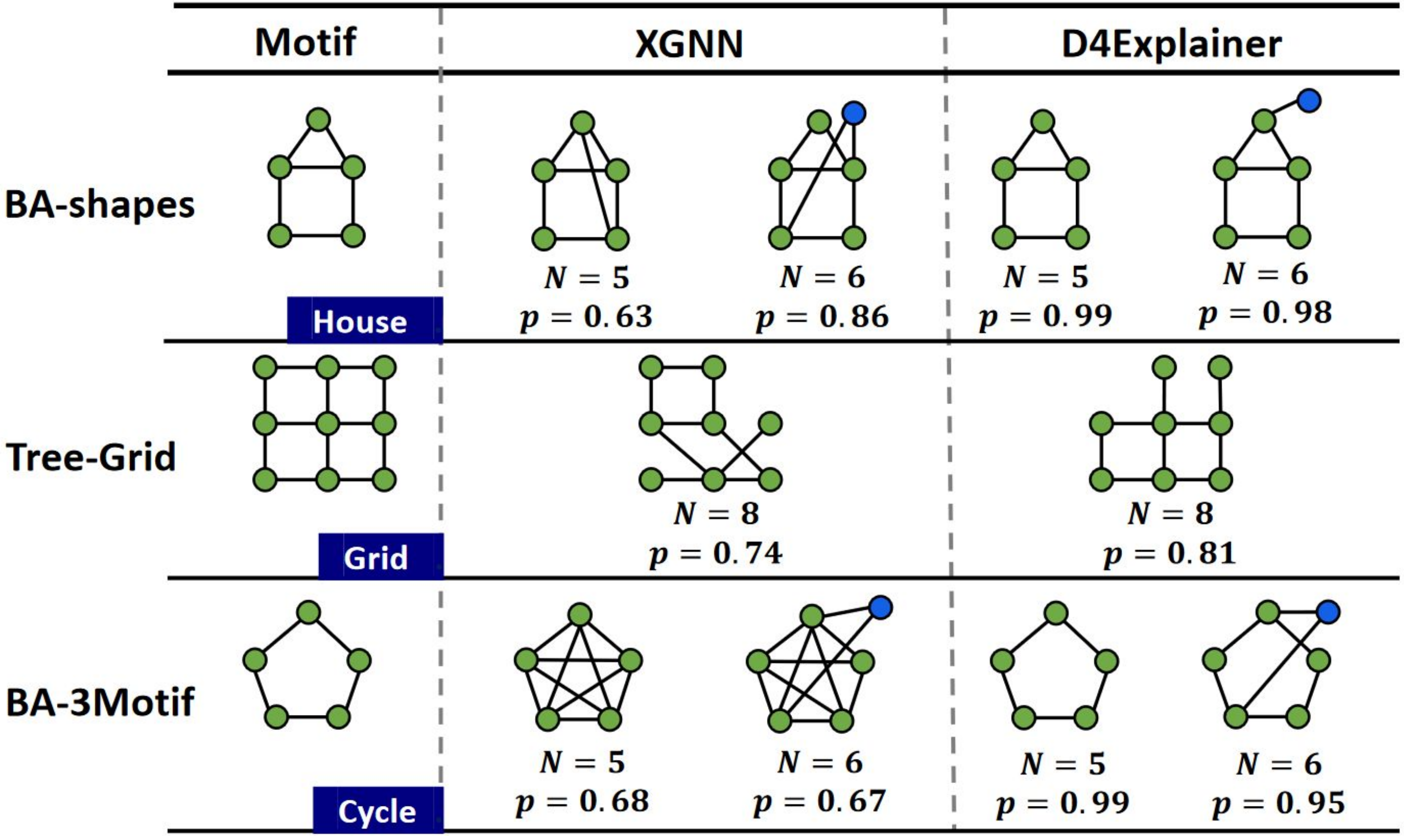
Counterfactual Evaluation (3)



Model-level Evaluation (1)

	# nodes	Mutag 6	Mutag 7	Mutag 8	Tree-Cycle 5	Tree-Cycle 6	Tree-Cycle 7
Ours	Prob.	**0.832**	**0.856**	**0.920**	**0.991**	**0.995**	0.989
	Density	**0.278**	**0.327**	**0.315**	0.400	**0.381**	**0.343**
XGNN	Prob.	0.523	0.824	0.875	0.968	0.989	**0.992**
	Density	0.537	0.479	0.437	0.400	0.390	0.367

Model-level Evaluation (2)



Limitations

- Scalability on large graphs
- Dependent on GNN architecture (limited generalizability)
- Model-level explanations: node count must be specified, domain-specific

Future Work

- **Address scalability**
  - Parallelization
  - Efficient training algorithms
  - Graph-specific optimizations
- **Explanation generation: include Node & Edge attributes**
  - Diffusion over continuous features
- **Address potential risks**