NExT++ Research Center

A Joint Research Collaboration Between NUS & Tsinghua University

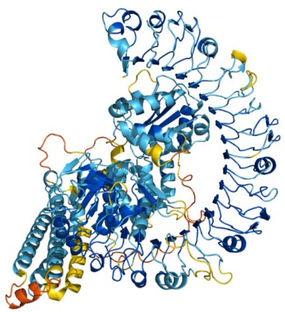
Towards Multi-Grained Explainability for Graph Neural Networks

Xiang Wang, Yingxin Wu, An Zhang, Xiangnan He, Tat-Seng Chua

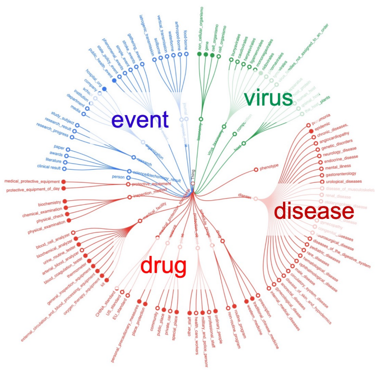
© Copyright NExT++. All Right Reserved.

https://github.com/Wuyxin/ReFine.

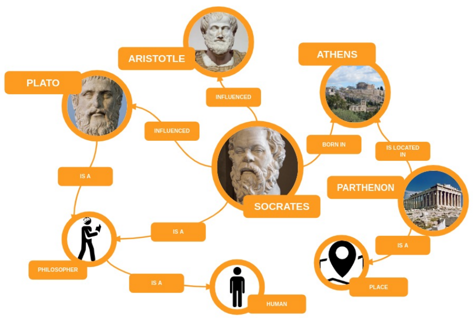
Graph Data is Ubiquitous



Protein Structure



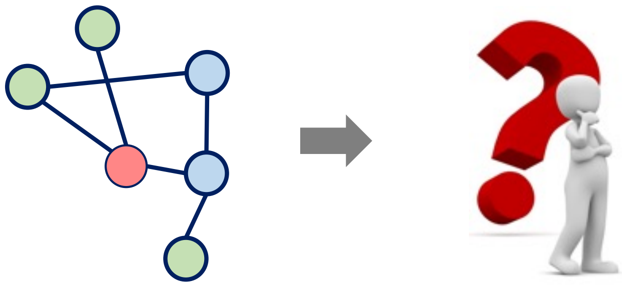
COVID Graph



Knowledge Graph

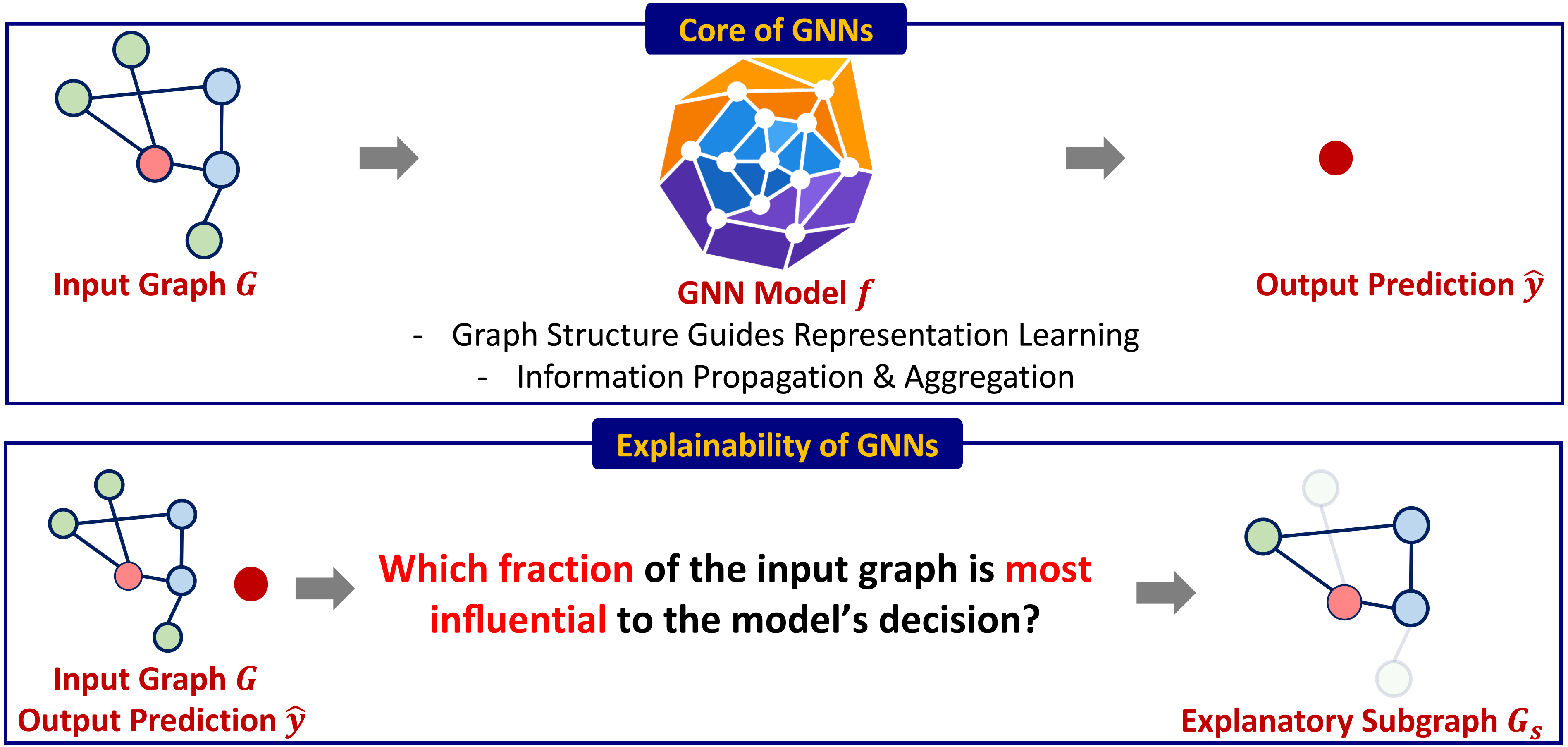
Social Network

Tasks of Graph Learning:
- Node Classification
- Graph Classification
- Link Prediction …



https://github.com/Wuyxin/ReFine.

Graph Neural Networks (GNNs)

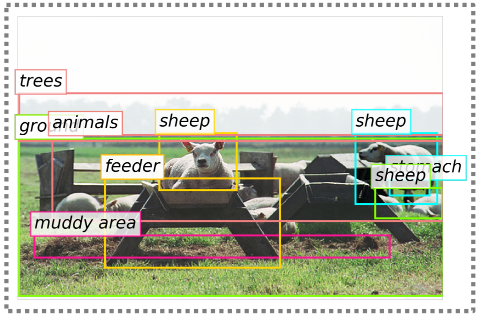


https://github.com/Wuyxin/ReFine.

Local & Global Explainability

Case: Scene graph classification

Question: Which fraction of the input graph is most influential to the model's decision?



Input Scene Graph

Output Prediction
$\hat{y}$: Farm

Explanation 1

Explanation 2

	Related works	Potential drawback
Local Explainability: Interprets each instance independently.	GNNExplainer [Ying et al. 2019] PGM-Explainer [Vu et al. 2020]	They hardly exhibit the class-wise patterns
Global Explainability: Systematizes the globally important patterns.	PGExplainer [Luo et al. 2020] XGNN [Yuan et al. 2020]	They might be trivial in the local context.

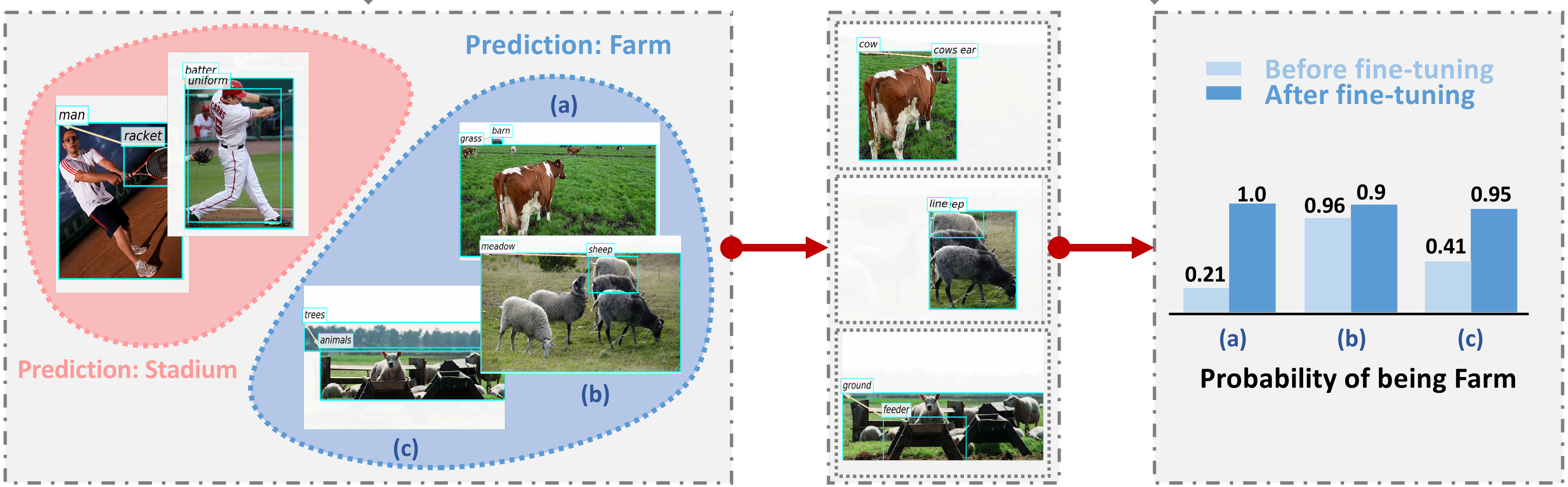
https://github.com/Wuyxin/ReFine.

Multi-Grained Explainability

NEURAL INFORMATION PROCESSING SYSTEMS

What class-wise knowledge does the GNN leverage to make predictions in general?

Why the GNN model made the certain prediction for the instance at hand?

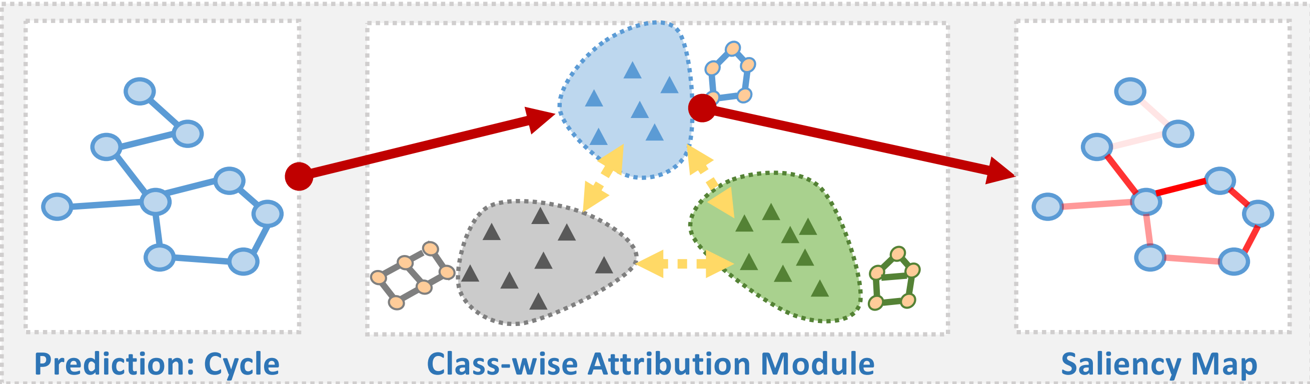


Pre-training towards Global Explainability

Fine-Tuning towards Local Explainability

https://github.com/Wuyxin/ReFine.

Pre-training & Fine-tuning (ReFine)



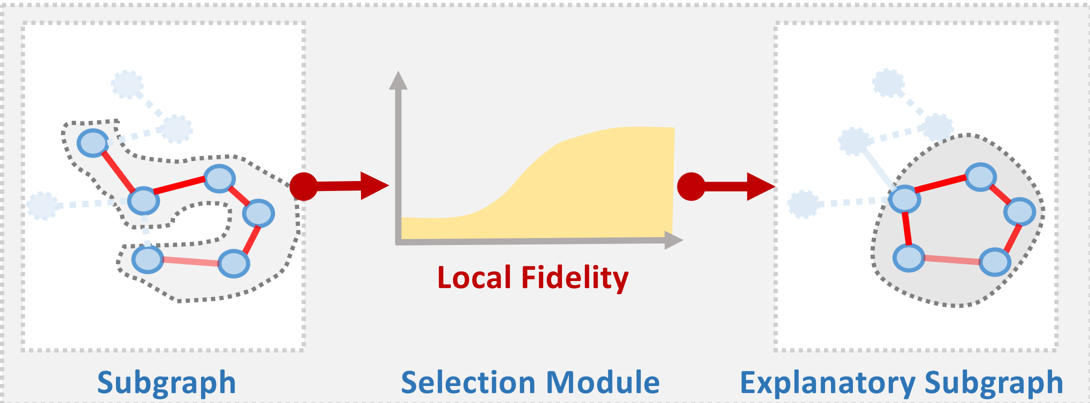
Pre-training

$\mathcal{L}_1 = MI(Y, \boldsymbol{M} \odot G_{att})$

Negative mutual information

$$\min_{\theta} \mathcal{L}_1 + \gamma \mathcal{L}_{\text{cts}}$$

Contrastive Loss



Fine-tuning

$\theta_0' = \theta$

$$\min_{\theta'} \mathcal{L}_2 = MI(Y, G_{exp})$$

where $G_{exp} = \textbf{Top}_{\boldsymbol{\rho}}(G_{att})$

User-defined ratio

https://github.com/Wuyxin/ReFine.

Empirical Results of ReFine

Table 1: Comparison of our ReFine and other baseline explainers

	Mutagenicity	VG-5	MNIST	BA-3motif	
	ACC-AUC	ACC-AUC	ACC-AUC	ACC-AUC	Recall@5
SA[9]	0.769	0.769	0.559	0.518	0.243
GNNExplainer[6]	0.895±0.010	0.895±0.003	0.535±0.013	0.528±0.005	0.157±0.002
PG-Explainer[7]	0.631±0.008	0.790±0.004	0.504±0.010	0.586±0.004	0.293±0.001
PGM-Explainer[19]	0.714±0.007	0.792±0.001	0.615±0.003	0.575±0.002	0.250±0.000
ReFine-CT	0.888±0.008	0.891±0.002	0.526±0.007	0.610±0.004	0.248±0.001
ReFine-FT	0.945±0.011	0.906±0.002	0.587±0.008	0.616±0.003	0.299±0.002
ReFine	**0.955**±0.005	**0.914**±0.001	**0.636**±0.003	**0.630**±0.006	**0.304**±0.000
Improvement	6.7%	2.1%	3.4%	7.5%	3.8%

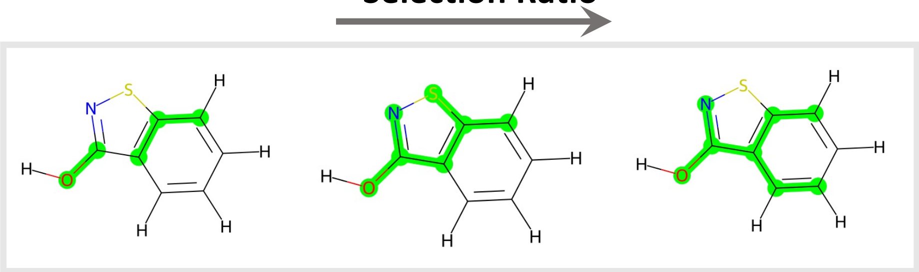
Table 2: Performance under different selection ratios before and after fine-tuning.

	Mutagenicity		VG-5		MNIST		BA-3motif	
ACC@ρ	0.4	0.6	0.4	0.6	0.4	0.6	0.4	0.6
ReFine-FT	96.8%	94.0%	91.3%	91.4%	41.4%	61.4%	36.0%	65.7%
ReFine	97.8%	96.2%	92.2%	93.4%	71.4%	82.0%	39.0%	72.8%
Improvement	+1.0%	+2.2%	+0.9%	+2.0%	+30.0%	+20.6%	+3.0%	+7.1%

https://github.com/Wuyxin/ReFine.

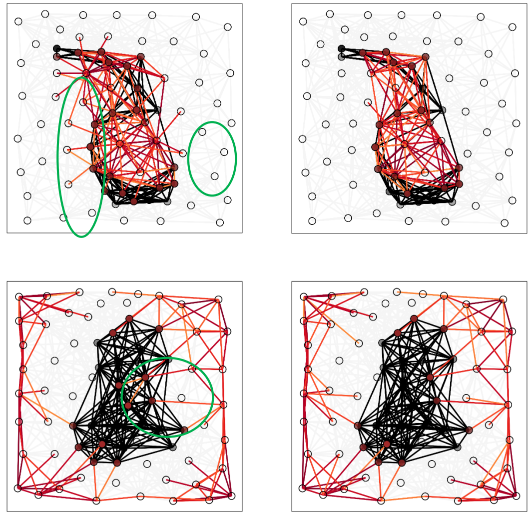
Empirical Results of ReFine

Selection Ratio



ACC: 0.543 0.888 0.976

Pre-trained Fine-tuned



https://github.com/Wuyxin/ReFine.

Summary

- Local explainability & Global explainability present different views of GNN models.
- Multi-grained explainability can offer more reliable & faithful explanations.

Check out our code and models at

- https://github.com/Wuyxin/ReFine.

THANK YOU!