Quantitative Evaluation of Explainable Graph Neural Networks for Molecular Property Prediction

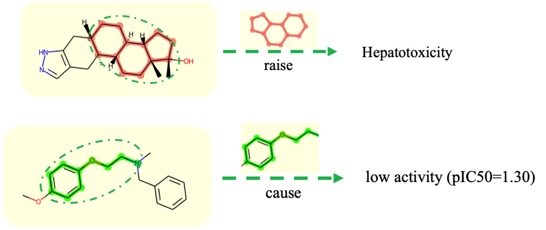
Jiahua Rao[1,2]*, Shuangjia Zheng[1,2]*, Yuedong Yang[1]

[1]Sun Yat-sen University [2]Galixir

Background

Motivating question:

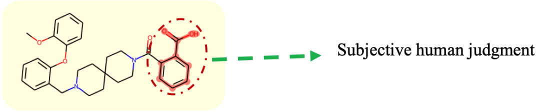
- Why do GNN models make the predictions they do?
- Whether the models learn the key substructures relevant for predicting labels correctly?



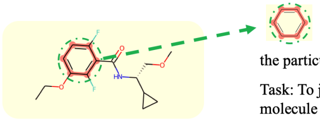
Challenge:
to quantitatively assess the interpretability of GNN

Currently Explainable datasets:

Without subgraph ground-truth



With subgraph ground-truth

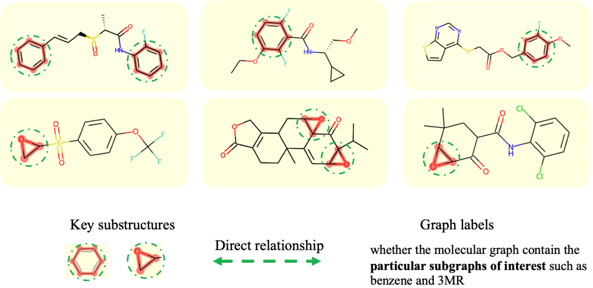


the particular subgraph as ground-truth

Task: To judge whether the corresponding molecule contains the particular subgraphs.

Construction of Benchmark

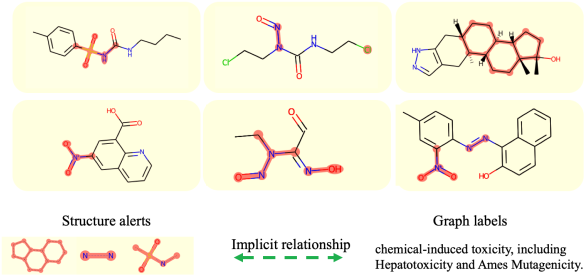
Single rationale: Regular Subgraph



Graph labels

whether the molecular graph contain the **particular subgraphs of interest** such as benzene and 3MR

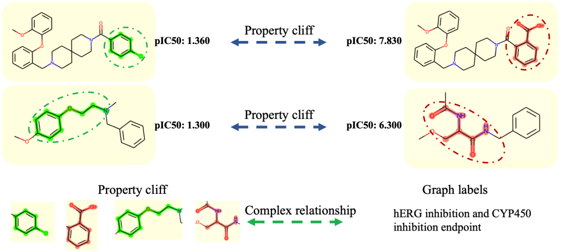
Multiple rationales: Diverse subgraph



Graph labels

chemical-induced toxicity, including Hepatotoxicity and Ames Mutagenicity.

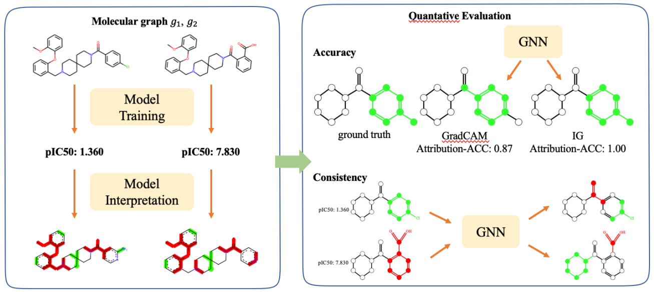
Property cliff: Uncertain subgraph



Graph labels

hERG inhibition and CYP450 inhibition endpoint

A uniform and rigorous Framework



Experimental Results

Single-rationale Dataset: 3MR

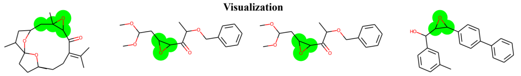
Attribution-AUC

XAI Method	CMPNN	GraphSAGE	GraphNet	GAT
Random Baseline	0.498	0.486	0.51	0.467
CAM	0.858	0.771	0.832	0.734
GradCAM (last)	0.794	0.783	0.754	0.745
GradCAM (all)	0.943	0.837	0.724	0.738
GradInput	0.951	0.844	0.967	0.887
IG	0.966	0.932	0.942	0.905
Attention				0.785

Attribution-ACC

XAI Method	CMPNN	GraphSAGE	GraphNet	GAT
Random Baseline	0.531	0.545	0.563	0.541
CAM	0.825	0.793	0.828	0.693
GradCAM (last)	0.605	0.576	0.714	0.626
GradCAM (all)	0.788	0.753	0.795	0.593
GradInput	0.925	0.876	0.931	0.888
IG	0.914	0.841	0.905	0.845
MCTS	0.809	0.769	0.865	0.745
Attention				0.789

Visualization



Single-rationale Dataset: Benzene

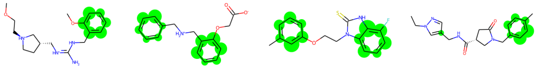
Attribution-AUC

XAI Method	CMPNN	GraphSAGE	GraphNet	GAT
Random Baseline	0.549	0.554	0.561	0.588
CAM	0.831	0.798	0.859	0.776
GradCAM (last)	0.566	0.582	0.657	0.604
GradCAM (all)	0.674	0.704	0.695	0.579
GradInput	0.894	0.859	0.934	0.867
IG	0.906	0.801	0.865	0.798
Attention				0.705

Attribution-ACC

XAI Method	CMPNN	GraphSAGE	GraphNet	GAT
Random Baseline	0.191	0.213	0.241	0.185
CAM	0.805	0.69	0.823	0.666
GradCAM (last)	0.767	0.738	0.746	0.687
GradCAM (all)	0.89	0.804	0.749	0.66
GradInput	0.94	0.907	0.946	0.876
IG	0.975	0.945	0.921	0.894
MCTS	0.846	0.523	0.877	0.838
Attention				0.775

Visualization



For more details, please refer to our paper in arXiv.