# Comparing GNN explainer faithfulness quantitatively in molecular property prediction

## Are GNN explanations for benzene rings comprehensive and sufficient?

Author: Heli Pajari
Contact: h.v.m.pajari@student.tudelft.nl
Supervisors: Dr. Megha Khosla and Dr. Jana Weber

TU Delft Delft University of Technology

## 1. Background

Drug development involves predicting molecule properties based on their structure
➔ e.g. toxicity affected by tens of molecule fragments
➔ Graph Neural Networks (GNN) perform well in identifying these fragments
➔ GNNs can reduce time and money spent on drug research

GNNs by nature have opaque decision making, their predictions cannot be used as-is
➔ If you don't know how a decision was made, is it safe or ethical to use it?

Explainable AI (XAI) techniques explain GNN decisions
➔ Performance evaluated with e.g. attribution accuracy/precision, fidelity

**BAGEL benchmark** [1]: offers four task agnostic metrics for evaluating GNN explainers, of which *faithfulness* is investigated:
- *Faithfulness*: does explanation replicate model behavior?
  - *Comprehensiveness* does explanation select all nodes/edges for a prediction?
  - *Sufficiency*: are selected nodes/edges enough to come up with model prediction?

## 2. Research question

*"How applicable are comprehensiveness and sufficiency as a way to measure GNN explainer faithfulness in molecular property prediction?"*

**RQ1** How can comprehensiveness and sufficiency from the Bagel benchmark be modified to work with MPP?

**RQ2** How large are the differences in comprehensiveness and sufficiency between explanations from Integrated Gradients and random explanations, using a CMPNN model trained on a benzene ring dataset?

## 3. Methodology

$G$ input graph, $G_E$ explaining graph, $G_N$ nonexplaining graph, $\mathbf{f}(G)$ GNN prediction with chemically valid input $G$

**Comprehensiveness** = $\mathbf{f}(G) - \mathbf{f}(G_N)$ , target value $\mathbf{f}(G)$
**Modified comprehensivenenss:** $\mathbf{f}(G_N)$ = sum of predictions for disjoint nonexplaining molecules

**Sufficiency** = $\mathbf{f}(G) - \mathbf{f}(G_E)$ , target value $\mathbf{0}$
**Modified sufficiency:** $\mathbf{f}(G_E)$ = mean of predictions for disjoint explaining molecules

**Average comprehensiveness** = $\sum_{g \in G} \frac{\mathbf{f}(g) - \mathbf{f}(g_N)}{f(g)} * \frac{1}{|G|}$

**Average sufficiency** = $\sum_{g \in G} \frac{|\mathbf{f}(g) - \mathbf{f}(g_E)|}{f(g)} * \frac{1}{|G|}$

**Dataset**: Benzene dataset from MolRep [2]
**GNN**: CMPNN
**Explainers**: Integrated Gradients, random splitting
**GNN explanations:** importance values for each atom in molecule, random selects from [0, 2 * 0.0001]
**Visualisation:** highlight atoms with importance ≥ 0.0001 and bonds between them with green

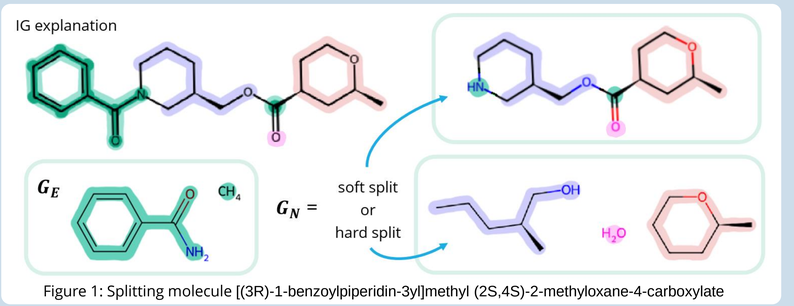


Figure 1: Splitting molecule [(3R)-1-benzoylpiperidin-3yl]methyl (2S,4S)-2-methyloxane-4-carboxylate

## 4. Results

### RQ1 Comprehensiveness and sufficiency of IG explanations

**Comprehensiveness**: values near 0, predictions for non-explanations between 0.4 - 0.5
- Modification has significantly lower values than original
- Original formula with hard split performs the best
- All scores far from target value

Table 1. Comprehensiveness: original vs. modified, soft vs. hard split

| mol | f(G) | original soft / hard \|f(G)\| | modified soft / hard \|f(G)\| | fragments soft / hard |
|---|---|---|---|---|
| 168 | 0.482 | 0.050 / 0.052 | 0.050 / -0.856 | 1 / 3 |
| 238 | 0.448 | 0.001 / 0.009 | 0.001 / -1.402 | 1 / 4 |
| 847 | 0.513 | 0.045 / 0.058 | -0.905 / -2.778 | 3 / 7 |
| 1018 | 0.537 | 0.105 / 0.103 | 0.105 / -0.368 | 1 / 2 |
| 1637 | 0.427 | -0.001 | -0.001 | 1 / 1 |

**Sufficiency**: values near 0 and usually negative, predictions for explanations > f(G)
- With a *perfect* explanation, ground truth can be found
- Modification improves the result
- With *imperfect* explanations, modified formula incorrectly better: non-benzene rings decrease average
  - 0 not a good target value
- Predictions for molecule and explanation can disagree, shows when explanation not faithful to model

Table 2. Sufficiency: *perfect* explanation for molecule 1018 original vs. modified

| f(G) | original f($G_E$) | modified f($G_E$) | original sufficiency \|0\| | modified sufficiency \|0\| |
|---|---|---|---|---|
| 0.537 | 0.589 | 0.581 | -0.052 | -0.044 |

Table 3. Sufficiency: *imperfect* explanations, original vs. modified

| mol | f(G) | original f($G_E$) | modified f($G_E$) | original vs. modified \|0\| | fragments |
|---|---|---|---|---|---|
| 168 | 0.482 | 0.565 | 0.521 | -0.083 / -0.039 | 2 |
| 238 | 0.448 | 0.466 | 0.472 | -0.018 / - 0.024 | 2 |
| 847 | 0.513 | 0.462 | | 0.051 | 1 |
| 1018 | 0.537 | 0.576 | 0.551 | -0.039 / -0.014 | 3 |
| 1637 | 0.427 | 0 | | 0.427 | 0 |

### RQ2 Comparing explainers using the original formulae

**Comprehensiveness:** IG 4.6% (soft split) and 3.7% better (hard split)

Table 4. Average comprehensiveness: soft split vs. hard split, n = 600

Table 4.1. Soft split

| explainer | average comp % \|1\| | % of samples \|1\| |
|---|---|---|
| IG | 0.095 | 0.603 |
| Random | 0.049 | 0.695 |

Table 4.2. Hard split

| explainer | average comp % \|1\| | % of samples \|1\| |
|---|---|---|
| IG | 0.125 | 0.995 |
| Random | 0.088 | 0.957 |

- Soft split discards 30 - 40% of input molecules
- Random has more chemically valid samples
- Hard split applicable to almost everything, might use very little of an input molecule
- Better results than soft split

**Sufficiency**: IG *0.5%* better = random fragment predictions indistinguishable from IG

Table 6. Average sufficiency, n = 600

| explainer | average suff % \|0\| | % of samples \|1\| |
|---|---|---|
| IG | 0.045 | 0.802 |
| Random | 0.050 | 0.957 |

## 5. Conclusions

**RQ 1** Comprehensiveness always low because $f(G_N)$ high

Sufficiency low because $\mathbf{f}(G_E)$ usually > $\mathbf{f}(G)$
- 0 not a good target value, requires a ground truth
- predictions higher with multiple target molecules
- shows when model and explanation disagree

**RQ 2** Comprehensiveness discards much of the data either on dataset or molecule level
➔ Comprehensiveness not good for MPP

Average sufficiency scores show that any explanation would have good sufficiency due to high model predictions
➔ Sufficiency not good for MPP

Comprehensiveness and sufficiency are ***not*** applicable for evaluating GNN explainer faithfulness in molecular property prediction

## 6. Limitations and future work

**Limitations**
- Model and explainer accuracy: could not reproduce claimed accuracy, random explainer had accuracy of 0.8
- Implementation: Hard splits not always 1-to-1 with input, results may not be exactly correct

**Future work:**
- investigate faithfulness metrics that don't require a ground truth or splitting molecules, such as RDT-fidelity from BAGEL

**References and acknowledgements**
[1] Rathee, M., Funke, T., Anand, A. & Khosla, M. BAGEL: A benchmark for assessing graph neural network explanations, 2022.

[2] Jiahua, R., Shuangjia, Z., Ying, S., Jianwen, C., Chengtao, L., Jiancong, X., Hui, Y., Hongming, C., & Yuedong, Y. Molrep: A deep representation learning library for molecular property prediction. bioRxiv, 2021.

Poster template adapted from PosterNerd