Evaluating GNN Explainer Faithfulness in Molecular Property Prediction Using Comprehensiveness and Sufficiency

Heli Pajari[1]
Supervisors: Dr. Megha Khosla[1]**, Dr. Jana Weber**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Heli Pajari
Final project course: CSE3000 Research Project
Thesis committee: Dr. Megha Khosla, Dr. Jana Weber, Dr. Thomas Abeel

An electronic version of this thesis is available at `https://repository.tudelft.nl/`.

Abstract

Predicting properties, such as toxicity or water solubility of unknown molecules with Graph Neural Networks has applications in drug research. Because of the ethical concerns associated with using artificial intelligence techniques in the medical field, explainable artificial intelligence techniques are used to explain how GNNs make their decisions. To evaluate the performance of those techniques, different metrics are used. The BAGEL benchmark proposes four such metrics, designed to be useable with any GNN explainer. Of these, the applicability of *faithfulness* was investigated in molecular property prediction, measured by the submetrics of comprehensiveness and sufficiency. While comprehensiveness and sufficiency were designed to be task agnostic, several shortcomings were identified that make it unsuitable for molecular property prediction. Future recommendations are to investigate other pre-established faithfulness metrics or to develop ones that do not require splitting molecules.

1 Introduction

Predicting properties such as the toxicity or water solubility of a molecule based on its structure is a vital part of drug research. This process is called molecular property prediction (MPP), and it is used to identify drug candidates. Achieving high accuracy and efficiency in MPP is valuable because developing new drugs is both very expensive and time-consuming [1, 2], not taking into account the risks associated with clinical failure.

MPP generally correlates given structural information of an unknown molecule with the predicted property. For some properties there is a ground truth, such as if the molecule has a particular substructure. Some chemical properties, such as what causes liver damage in humans, are very difficult to predict because there are dozens of chemical fragments affecting the toxicity of a molecule[3]. There are also structurally near-identical molecules with drastically different behaviours[3]. Due to these challenges, even highly experienced medical chemists struggle to estimate the properties of unknown molecules[3].

Deep Learning has been applied to MPP due to its ability to model "complex nonlinear relationships" [4], and methods like Graph Neural Networks (GNN) have been used on graph representations of molecules. In experiments conducted in [3], GNNs were able to achieve higher predictive accuracy in identifying both human hepatotoxicity and related key substructures than human experts.

However, while GNNs perform well, their decision making is completely opaque[5], which leads to ethical and safety concerns. To solve the problem of having results from GNNs without insight into how the decisions were made, many different explainable artificial intelligence (XAI) methods have been developed. The current state of the art in this domain is the Substructure Mask Explanation (SME)[4]. SME can split molecules in ways that align with how medicinal chemists prefer to investigate them, such as BRICS or Murcko substructures, and functional groups. The explainer uses this information to identify the most important substructures responsible for the model's prediction.

Different metrics have been used to evaluate the quality of GNN explainers. The BAGEL benchmark[5] proposes four systematic ones of which faithfulness measured by comprehensiveness and sufficiency is investigated in this work. The BAGEL metrics have further been used in research that investigates how the choice of GNN architecture, dataset, and explainer affect the performance of a prediction[6], where a unique faithfulness measure from comprehensiveness and sufficiency was defined. The work did not investigate the what values the datasets obtained for the individual submetrics.

It is not yet known if the BAGEL metrics are suitable to evaluate GNN explainer performance in MPP. The medical research field can benefit from new ways to assess the fitness of GNN explanations, and it would increase expert trust in the assessed GNN models. It would also be valuable to know where the BAGEL metrics do not yet succeed. To investigate this, a research question is proposed:

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

To further define the research, the following subquestions are used:

1. How can comprehensiveness and sufficiency from the BAGEL benchmark be adapted to work with MPP?
2. How large are the differences in comprehensiveness and sufficiency between relevant explanations from Integrated Gradients and random explanations, using a Communicative Message Passing Neural Network model trained on a benzene ring dataset?

To answer the research questions, first the background information and methodology are established in section 2. The proposed modifications to comprehensiveness and sufficiency are explained in section 3. The results of the experiment are presented and discussed in section 4, and section 5 reflects on the aspects of responsible research in the context of this work. A summary of the findings and recommendations for further research on GNN explainers in MPP can be found in section 6.

2 Background and Experimental Work

This section provides information on graph neural networks, graph neural network explainers and an overview of the faithfulness metrics defined in BAGEL.

2.1 Graph Neural Networks

Graph Neural Networks (GNN) are a type of artificial neural network that process data represented as graphs $G(V, E)$, where V is the set of nodes and E is the set of edges. GNNs use both the feature information and the graph structure to make predictions.

Communicative Message Passing Neural Network (CMPNN) is a GNN specifically developed for MPP, and it can update both edge and node embeddings interactively. The notation used in its molecular embedding function is listed in Table 1, and the algorithm itself is shown in Algorithm 1. Lines 3-10 of the algorithm describe the update function for vertices and edges:

- For each vertex v, an intermediate message vector m_v^k is constructed by aggregating the hidden states of its incoming edges. The updated hidden state h_v^k for v is the output of the communicative function with the message vector and the previous hidden state of v as its inputs.
- For each directed edge e, construct an intermediate message vector m_e^k from the current hidden state of its starting vertex v subtracted by the previous hidden state of its inverse edge. The updated hidden state h_e^k is the sum of the initial hidden state of the edge $h^0(e_{v,w})$ and the weight matrix multiplied by the message vector m_e^k.

Table 1: Mathematical notation used in the CMPNN molecular embedding algorithm.

$G = (V, E)$	Input graph.
$u, v, \ldots$	Nodes in G.
$e_{u,v}$	An edge from node u to v.
$N(v)$	The set of neighbour nodes of node v.
$\mathbf{x}$	Raw feature.
$h^i(v)$	The hidden representation of node v in layer i.
$h^i(e_{v,w})$	The hidden representation of edge $e_{v,w}$ in layer i.
$\mathbf{W}$	Weight matrix.
σ	Activation function; ReLU.

Algorithm 1 CMPNN embedding generation algorithm, as defined in [7].

Input: Graph $G(V, E)$; depth K; input node and edge features $\{x_{e,v}, \forall e, v \in E, x_v, \forall v \in V\}$; aggregate function Aggregate, communicative function Communicate; weight matrix W
Output: Graph-wise vector representation $\mathbf{z}$

```
$h^0(e_{v,w}) \leftarrow x_{e,v}, \forall e_{v,w} \in E; h^0(v) \leftarrow x_v, \forall v \in V$
for $k = 1$ to $K$ do
    for $v \in V$ do
        $m_v^k \leftarrow$ Aggregate$(\{h^{k-1}(e_{v,u}), \forall u \in N(v)\})$
        $h_v^k \leftarrow$ Communicate$(m_v^k, h_v^{k-1})$
    end for
    for $e \in E$ do
        $m_e^k \leftarrow h^k(v) - h^{k-1}(e_{w,v})$
        $h_e^k \leftarrow h^0(e_{v,w}) + W \cdot m_e^k$
    end for
end for
for $v \in V$ do
    $m(v) \leftarrow$ Aggregate$(\{h^K(e_{v,u}), \forall u \in N(v)\})$
    $h(v) \leftarrow$ Communicate$(m(v), h^K(v), x(v))$
end for
$\mathbf{z} \leftarrow$ Readout$(\{h(v), \forall v \in V\})$
```

2.2 GNN Explainers

Explainers are an explainable AI technique for explaining how a GNN model came to its decision. An explanation is the output of an explanation function and an interpretable description of the model's behaviour. The exact form of the explanation depends on the model and data used, and who the intended user of the explanation is (Dr. Khosla, personal communication, May 28, 2024).

The explanation outputs in this work are two arrays; one with an importance value for every atom in the molecule, and the other for the importance values of each bond. Figure 1 shows an explanation for a model trained on benzene ring data, where atoms with importance values above a given threshold value and the bonds between such atoms are highlighted.

Experiments conducted in this paper use Integrated Gradients (IG), as defined in [8].

[1] 13 SMILES Cn1cncc1CCNC(=O)c1cc(cnc1)F

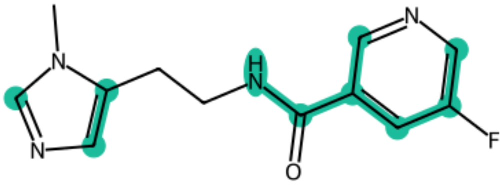


Figure 1: Visualisation of the outcome of the integrated gradients method for the prediction of benzene ring presence of molecule 5-fluoro-N-[2-(3-methylimidazol-4-yl)ethyl]pyridine-3-carboxamide (13^1). Atoms and bonds highlighted in green were identified as crucial for the model's prediction.

IG calculates the path integral between the input graph G and the baseline counterfactual G^b[9]. The integrated gradients for an input G can be approximated with a sum as follows:

$$IntegratedGrads_i^{approx}(G) ::= (G_i - G_i^b) \times \Sigma_{k=1}^{m} \frac{\vartheta f(G' + \frac{k}{m}(G_i - G_i'))}{\vartheta G_i} \times \frac{1}{m} \tag{1}$$

Where m is the number of steps for the Riemann sum approximation, and f is a function representing GNN output given an input. $\frac{\vartheta f(x)}{\vartheta x_i}$ represents the gradient of $f(x)$ along the i^{th} dimension.

In essence, the function calculates the gradients for a set of inputs, and averages them out before multiplying the subtraction of the input G and baseline G^b in the i^{th} dimension. The used implementation sets m to 200. The baseline G^b used for chemistry models in [8] is a zeroed out feature vector and is assumed to be the same for all chemistry applications.

2.3 GNN Explainer Evaluation

The BAGEL benchmark was developed as a general framework for GNN explainer evaluation[5]. It proposes the following metrics: 1) faithfulness, 2) sparsity, 3) correctness, and 4) plausibility, the first of which is described below.

Faithfulness is a measure of how well an explanation can describe model behaviour. It has two submeasures, the choice of which depends on what type of data or explanations are used.

1. Rate Distortion Based (RDT) Fidelity is used with explanations with node and edge feature attributions but not necessarily attributions related to the graph structure. An explaining subgraph and its features are relevant to the predicted property, if the model prediction stays nearly the same when the rest of the node and edge features have been randomised.

2. Comprehensiveness and Sufficiency are used to measure faithfulness in explanations that have only structural information available. Comprehensiveness measures if the explanation has captured every node/edge that led to the model's prediction, and sufficiency if the nodes and edges in the explanation alone are enough to come up with the original prediction[5]. The submetrics are defined as follows:

$$comprehensiveness = f(G)_j - f(G/G_E)_j \tag{2}$$

$$sufficiency = f(G)_j - f(G_E)_j \tag{3}$$

Where G_E is the explaining graph of G with $G_E \subseteq G$; G/G_E is the non-explaining graph, expressed as the difference between G and G_E, and which will be referred to as G_N in this work; f is the trained GNN model; and $f(G)_j$ the GNN prediction for G, for the j^{th} class.

For comprehensiveness, a value near the original prediction $f(G)_j$ is preferred to indicate that non-explaining nodes and edges in G have low predictive power. For sufficiency, a value near zero shows that most of $f(G)_j$ is because the explanation graph G_E is part of G.

2.4 Molecular Representation

Simplified Molecular-Input Line-Entry System (SMILES) is a specification to create ASCII string representations of molecules, where a period (.) is used as a separator between disjoint molecules. For example, the string ”$O.C1CCCCC1$” represents a water molecule and a ring of 6 carbon atoms separate from it. In this report, a "valid" SMILES refers to a string that can be converted into a molecule graph.

The International Union of Pure and Applied Chemistry (IUPAC) names for molecules used in this work were obtained through the Python library *PubChemPy*.

2.5 Experimental Pipeline

This work adapts comprehensiveness and sufficiency for the purpose of molecular property prediction on two distinct explainers. In the following, the choice of dataset, GNN and GNN explainers for the experimental pipeline are motivated.

Dataset The synthetic benzene ring dataset used in this work is from MolRep, a Python package that provides datasets, deep learning models and explainers for MPP[9]. The dataset was used because it has a ground truth and the model can be trained to recognise if benzene rings are present. These features make it a suitable baseline to test how comprehensiveness and sufficiency perform in MPP. The numbers used for the molecules in this work refer to their indices in the dataset, which had 12 000 entries in total.

GNN CMPNN was chosen as the GNN model to use the evaluated GNN explainers on. The choice was based on the benchmarking done in [3], where CMPNN[7] performed the best overall and was able to achieve perfect accuracy with the benzene dataset. Accuracy metrics for the used model are shown in Table 2.

Table 2: Accuracy, Area-Under-Curve, F1, Precision and Recall for the CMPNN model trained for a single epoch on the benzene dataset, with claimed Area-Under-Curve value from [3] in parentheses, and the target values in brackets.

ACC [1]	AUC [1]	F1 [1]	Precision [1]	Recall [1]
0.748	0.845 (1.000)	0.773	0.708	0.851

Despite the lower performance compared to the claimed Area-Under-Curve value of 1.000 in [3], the trained model had a decent value for it. The precision score indicates that from all the positive samples the model identified, most were true positives. Recall shows that the model could identify more true positives than false negatives.

GNN Explainers The explainers chosen were Integrated Gradients (IG) and a MolRep provided baseline that outputs random values for each feature of the input molecule. The choice of IG was because it was able to achieve a high explanation accuracy with CMPNN and the Benzene dataset in [3]. The random baseline was chosen so IG could be compared against an explainer that should not be able to give a benzene ring as an explanation. The parameters were chosen so that the average random explanation had approximately 50% of the atoms in the input molecule. The accuracy metrics for the explainers used for the experiment are shown in Table 3.

Table 3: Explainer accuracy, Area-Under-Receiving-Operating-Curve, F1, Precision and Recall for IG and a random explainer for a CMPNN model trained on the benzene dataset, with estimated AUROC from Figure 2B in [3] in parentheses, and the target values in brackets.

Explainer	ACC [1]	AUROC [1]	F1 [1]	Precision [1]	Recall [1]
IG	0.826	0.948 (0.903)	0.271	0.410	0.202
Random	0.815	0.502 (0.513)	0.000	0.000	N/A

2.6 Extracting Subgraphs From Molecule Explanation

Both comprehensiveness and sufficiency require extracting subgraphs from the input. To obtain the subgraphs, the threshold value that the explanation visualisation method in MolRep uses, was used to divide atoms and bonds into explaining and nonexplaining ones. The explaining graph G_E was extracted from the molecule graph G by removing atoms with an importance value strictly lower than the threshold, as well as every bond between those atoms. Two methods to extract the non-explaining graph G_N were used, called "soft" and "hard" splitting. Soft splits preserve bonds between explaining and non-explaining atoms, while hard splits do not.

An example of splitting a molecule into its explaining and nonexplaining parts is shown in Figure 2. Molecule 168[2] had two molecules[3] in G_E. With the soft split, G_N was a single molecule[4], while the hard split produced three disjoint molecules[5].

Because of the implementation of the splitting method, invalid molecules were often generated for both G_E and G_N. Because the GNN requires chemically valid input, the molecules were sanitised by splitting their SMILES into a list using a period (.) as the delimiter, and constructing a new period-delimited SMILES of the valid fragments. If no valid parts remained, the explanation could not be evaluated. These cases were considered separately from the ones where the (non)explaining graph consisted of the entire input molecule.

2.7 Code Repositories

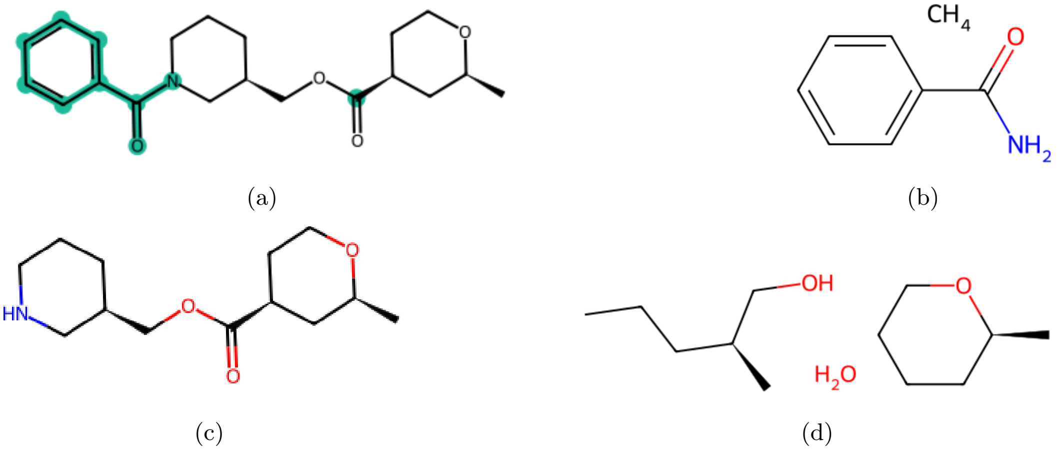
MolRep MolRep provides datasets, deep learning models and explainers for MPP[9]. It can train GNN models on MPP datasets, generate explanations, and visualisations of those

[2] 168 SMILES C[C@H]1C[C@H](CCO1)C(=O)OC[C@@H]1CCCN(C1)C(=O)c1ccccc1

[3] Explaining SMILES C.NC(=O)c1ccccc1

[4] Non-explaining (soft) SMILES C[C@H]1C[C@@H](C(=O)OC[C@@H]2CCCNC2)CCO1

[5] Non-explaining (hard) SMILES CCC[C@H](C)CO.C[C@H]1CCCCO1.O



(a)

(b)

(c)

(d)

Figure 2: Demonstration of splitting molecule [(3R)-1-benzoylpiperidin-3-yl]methyl (2S,4S)-2-methyloxane-4-carboxylate (168[2]) into explaining and nonexplaining parts: (a) IG explanation for prediction of benzene ring presence for molecule 168, with atoms and bonds crucial to model prediction highlighted in green, (b) Explaining fragments: benzamide and methane[3], (c) Soft split into one non-explaining molecule [(3R)-piperidin-3-yl]methyl oxane-4-carboxylate[4], (d) Hard split into three non-explaining molecules: (2S)-2-methylpentan-1-ol, (2S)-2-methyloxane, and oxidane[5]. Colours in (b), (c) and (d) represent non-carbon atoms and molecules.

explanations. It is available at `https://github.com/biomed-AI/MolRep`.

This work The metrics and experiment for this project were implemented in Python. Pre-existing MolRep code was used to train a model and generate explanations. The implementation is publicly available at `https://github.com/helipajari/mpp-comp-suff`.

3 Applying Comprehensiveness and Sufficiency to MPP Explainers

3.1 Proposed Modifications to Formulae

Splitting the molecule often results in disjoint fragments in both G_E and G_N. To investigate how predicting the fragments separately affects comprehensiveness and sufficiency in binary prediction tasks, changes were made to both Equation 2 and Equation 3, presented below:

$$comprehensiveness' = f(G) - \sum_{g \in G_N} f(g) \tag{4}$$

$$sufficiency' = f(G) - \frac{\sum_{g' \in G_E} f(g')}{|G_E|} \tag{5}$$

Where g is a non-explaining fragment in G_N, g' is an explaining fragment in G_E and

$f(x)$ is the GNN prediction of the given input x.

The optimal value for unmodified comprehensiveness is as close to f(G) as possible. Assuming that f(G_N) is the sum of all nonexplaining fragments, predicting non-explaining fragments separately should yield low values for each of them. Values for modified comprehensiveness are expected to be similar to those produced by the original one.

The changes to sufficiency were made because the set of explaining fragments can contain multiple target structures. For binary classification such as benzene ring detection, any one of them should be sufficient to come up with the original prediction. Because the target value for unmodified sufficiency is close to 0, the predictions for explaining fragments for modified sufficiency are averaged before subtracting from the prediction of the original molecule.

3.2 Comparing Explainers Under Comprehensiveness and Sufficiency

To compare explainer performance, average comprehensiveness and sufficiency scores were calculated for each explainer. Average comprehensiveness for an explainer was defined as the mean of $(comprehensiveness(G)/f(G))$ for every molecule G in the dataset where comprehensiveness was defined. Values near 1 indicated good performance. Average sufficiency was calculated as the mean of the absolute value for sufficiency, for every molecule where sufficiency was defined. Values near 0 indicated good performance.

4 Results and Discussion

4.1 RQ1: Adapting Comprehensiveness and Sufficiency for MPP

To demonstrate the effects of the different formulae and splitting methods on comprehensiveness and sufficiency, IG explanations for molecules 168[2], 238[6], 847[7], 1018[8], and 1637[9] were used, shown in Figure 3.

Comparisons for comprehensiveness between soft and hard splits, and differences between the original and modified formulae are demonstrated with molecules 168, 238, 847, 1018 and 1637 in Table 4.

Table 4: Comparisons between soft and hard split methods and formulae for comprehensiveness on IG benzene explanations for molecules 168, 238, 847, 1018 and 1637, with the target value in brackets.

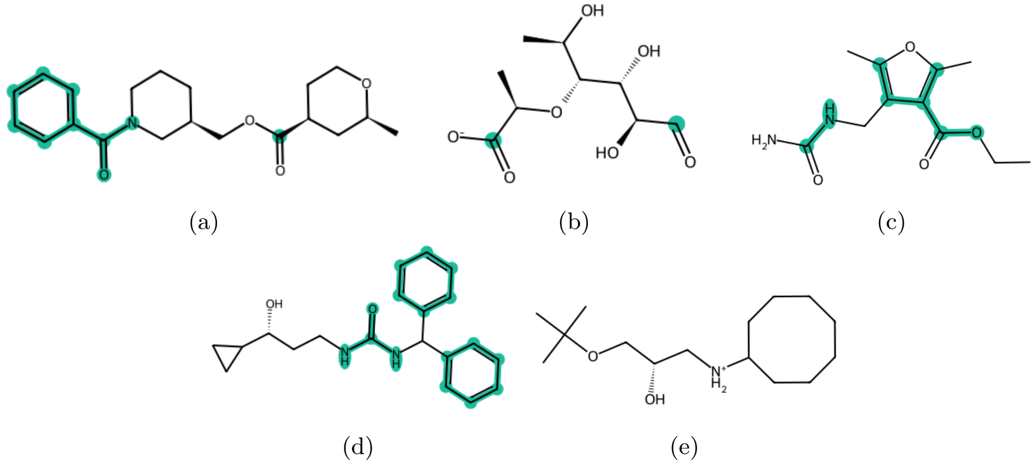
mol	f(G)	original	modified	fragments
		soft / hard [f(G)]	soft / hard [f(G)]	soft / hard
168	0.482	0.050 / 0.052	0.050 / -0.856	1 / 3
238	0.448	0.001 / 0.009	0.001 / -1.402	1 / 4
847	0.513	0.045 / 0.058	-0.905 / -2.778	3 / 7
1018	0.537	0.105 / 0.103	0.105 / -0.368	1 / 2
1637	0.427	-0.001	-0.001	1 / 1

[6] 238 SMILES C[C@H]([C@H]([C@@H]([C@@H](C=O)O)O)O[C@H](C)C(=O)[O-])O

[7] 847 SMILES CCOC(=O)c1c(oc(c1CNC(=O)N)C)C

[8] 1018 SMILES c1ccc(cc1)C(c1ccccc1)NC(=O)NCC[C@H](C1CC1)O

[9] 1637 SMILES c1cc(nc(c1)OCC(F)F)C(=O)Nc1ccncc1O



(a)

(b)

(c)

(d)

(e)

Figure 3: Visualisation of the outcome of the integrated gradients methods for the prediction of benzene ring presence with atoms and bonds crucial to model prediction highlighted in green. Representations for molecules (a) [(3R)-1-benzoylpiperidin-3-yl]methyl (2S,4S)-2-methyloxane-4-carboxylate (168[2]), (b) *no IUPAC name*, SMILES: *C[C@H]([C@H]([C@@H]([C@@H](C=O)O)O)O[C@H](C)C(=O)[O-])O* (238[6]), (c) ethyl 4-[(carbamoylamino)methyl]-2,5-dimethylfuran-3-carboxylate (847[7]), (d) 1-benzhydryl-3-[(3R)-3-cyclopropyl-3-hydroxypropyl]urea (1018[8]), and (e) cyclooctyl-[(2S)-2-hydroxy-3-[(2-methylpropan-2-yl)oxy]propyl]azanium (1637[9]).

Comprehensiveness scores for the original formula in Table 4 are positive values close to zero for every molecule except for 1637, for which G_N was equal to G. For the rest, the values show that G_N had a lower prediction that G for both split types. f(G) and comprehensiveness appear to be positively correlated, except for molecules 168 and 847 when using the soft split. The meaning of this is unclear.

For the modified formula, the soft split has the same values to the hard one when G_N has one fragment, else the values are more negative the more fragments there are. It can be reasoned that evaluating G_N fragments separately results in similar values to G for each of them, leading to a low overall value. The formula modification does not improve explanation comprehensiveness values, and the assumption the modification was based on appears to have been incorrect.

Overall, comprehensiveness does not perform as expected. Using the target value f(G), high values for f(G_N) indicate that the explanations did not capture every node and edge relevant to the model decision. If the target value for comprehensiveness is changed to a positive value to simply indicate G_N has lower prediction than G, it could be argued that the explanations are comprehensive.

To investigate the effect of different sufficiency formulae on molecules with multiple explaining benzene rings, molecule 1018 was used. Its explanation was manipulated to only include benzene rings. The results in Table 5 show that the formula modification has a 0.008 improvement over the original one when compared to the target value.

The table also shows that the model prediction for a single benzene ring is 0.581, and multiple benzene rings have a higher prediction when evaluated together. This is a useful

Table 5: Model predictions and different sufficiency formulae of molecule 1018 when the explaining graph G_E contains only two benzene rings, with the target value in brackets.

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

feature for MPP tasks like toxicity prediction, where the overall toxicity of a molecule is due to the presence of multiple less toxic substructures, such as shown in the hepatotoxicity study in [3]. By finding a suitable threshold value for splitting molecules, sufficiency could be used as implemented in this work.

Comparisons for using the original and modified sufficiency formulae on unmanipulated explanations for molecules 168, 238, 847, 1018 and 1637 are in Table 6.

Table 6: Comparisons between formulae for sufficiency on IG benzene explanations for molecules 168, 238, 847, 1018 and 1637, with the target value in brackets.

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

Sufficiency scores are near zero for every molecule other than 1637. 1637 has an empty explanation, so its sufficiency is equal to f(G). For the other molecules, f(G_E) was slightly higher than f(G). The exception to this is molecule 847, for which f(G_E) was lower. Both original and modified f(G_E) are on the same side of 0.5 as f(G), except for molecules 168 and 847. This appears to indicate that the explanations are not faithful to the model prediction. This is supported by the molecule graphs Figure 3a and Figure 3c: 168 has a benzene ring while 847 does not.

The modified sufficiency score is closer to 0 for molecules 168 and 1018 which have both benzene rings and other molecules in G_E. This shows that either 0 should not be used as the target value, or the formula modification should be reconsidered, as irrelevant molecules in the explanations bring the average value closer to 0. It does not seem to be a good target for the original formula either, because the more benzene rings a molecule has, the higher f(G_E) becomes, leading to lower sufficiency. This suggests that sufficiency should be as low as possible, but the nuance of contradicting explanation and model predictions is lost, as the explanation for molecule 168 would be more sufficient than 238, even if the f(G_E) is closer to f(G) for 238.

Based on these results, sufficiency can show if the model and explanation are in agreement, which is derived of the definition in [5]: is the explanation alone enough to come up with the model prediction. This can be useful for MPP, if the target values are reconsidered.

4.2 RQ2: Comparing Explainers

To compare how different explainers perform under comprehensiveness and sufficiency, the average scores were obtained from IG and a random explainer, using 600 random samples without replacement, both splitting methods and the original formulae. The results are shown in Table 7 and Table 8.

Table 7: Comparing average comprehensiveness of IG and a random explainer for (a) soft and (b) hard split, using original formulae on 600 samples, with the target values in brackets.

(a) Soft split

explainer	average comp % [1]	% of samples [1]
IG	0.095	0.603
Random	0.049	0.695

(b) Hard split

explainer	average comp % [1]	% of samples [1]
IG	0.125	0.995
Random	0.088	0.957

Average comprehensiveness scores with the soft split are 0.095 for IG and 0.048 for random, and for the hard split 0.125 and 0.88, respectively. IG has scores closer to the target in both cases. Like in Table 4, using the hard split resulted in higher values for both explainers. The hard split also resulted in almost every molecule having at least one valid nonexplaining fragment for both explainer, compared to the soft split averaging at about 65% valid samples. The higher number of samples is not a clear improvement, because at its extreme a single atom of the molecule was used to calculate the comprehensiveness of an explanation. A metric that might only use such a small fraction of a molecule can hardly be considered useful, but one that can not be used for 30-40% of all data is not very useful either.

Average sufficiency scores in Table 8 are about 0.05 for both explainers, which shows that their explanations were close to the molecule predictions.

Table 8: Comparing average sufficiency of IG and a random explainer, using original formulae on 600 samples, with the target values in brackets.

explainer	average suff % [0]	% of samples [1]
IG	0.045	0.802
Random	0.050	0.957

Sufficiency is 0.005 lower for IG, indicating it performs better. Sufficiency could be calculated for about 80% of samples for IG and 96% for the random explanations, meaning that most explanations had at least one chemically meaningful fragment. As with comprehensiveness, a higher number does not indicate a better explainer. Random atom and bond explanations being only 5/1000th worse on average than explanations that could contain benzene rings seems to indicate that any explanation would be highly sufficient. This is consistent with comprehensiveness scores in subsection 4.1, showing that non-target structures attain high prediction values. The score for random explanations is also very close to 0.052, the average sufficiency which the perfect explanation for molecule 1018 in Table 5 would get. This shows that the GNN model does not create large differences between random fragments and the target structures, which makes it difficult to compare explainers with sufficiency.

4.3 Reflection

The initially assessed suitability of comprehensiveness was based on the assumption that the non-explaining graph of a molecule would be a valid molecule in isolation. With a better understanding of chemistry, comprehensiveness could have been dismissed altogether.

The usefulness of comprehensiveness and sufficiency for MPP is also severely limited by G_E and G_N having to be chemically valid, so that both subgraphs can be evaluated by the GNN model. Assuming a well performing explainer and model, sufficiency is more applicable of the two, because the target structures in G_E should be chemically valid to have the predicted molecular properties.

Without modifications to the metrics, assigning the entire molecule to be the explanation is a trivial way to define the absolute most comprehensive and sufficient explanations.

4.4 Limitations

GNN Model and Explainers

The trained CMPNN model could not reproduce the cited AUROC scores of [3]. It could get a value of 0.845, but it was still lower than 1.000 reported in the literature. Training the model for more epochs made it worse: model AUC was 0.5 with ten epochs, and explainer precision fell to zero when two or more epochs were used. The tables for model and explainer accuracies are included in Appendix A. Additionally, the MolRep CMPNN model ignored generating bond importances for explanations by default, raising questions about the reliability of the produced models and explanations. Further concerns were raised by the bizarrely high accuracy score (0.815) of the random explanations.

Despite the peculiarities, the explanations could show that the model predictions were very similar regardless of input. It is possible that a more accurate model could have led to IG predicting fewer false positives and negatives, but it would not have created larger differences between molecules with benzene rings and those without them.

An explainer with high precision and recall would have been interesting to compare against the high accuracy but low F1 IG. This could have given more insight into how comprehensiveness and sufficiency worked with the achieved implementation.

Implementation

The molecule splitting method could have been made to be more faithful to the model prediction. This could have been achieved by assigning atoms with importance values under the threshold to the explaining graph G_E when the model prediction was less than 0.5. For molecules with empty explanations such as 1637 in Figure 3e, this would result in perfect sufficiency and comprehensiveness scores because the whole molecule would have been used as the explaining fragment. This idea was not implemented due to a lack of time.

The splitting was not rigorously tested or appraised by a domain expert, so it is possible that extracted subgraphs did not exactly correspond with the input molecules. An example of this is in Figure 2d, where it can be seen that the thicker bond does not connect to the OH-group like it does in Figure 2a. It is thus clear that some molecules used for calculating comprehensiveness and sufficiency were incorrect. For the conclusions this does not matter, but the exact results may be wrong. The molecules used for the results in subsection 4.1 were verified to be correct except for molecule 168, but checking the molecules for results in subsection 4.2 was infeasible.

5 Responsible Research

5.1 Repeatability and Reproducibility

The methods used to obtain comprehensiveness and sufficiency for a molecule are both repeatable and reproducible when using the same model and explanations between repetitions. The trained CMPNN model, explanations for each used explainer, the random split for comprehensiveness and sufficiency, and code implemented for the project have been uploaded to a public repository, so the same results for comprehensiveness and sufficiency can be obtained by using the provided files.

Additionally, the MolRep repository and the BAGEL benchmark are open access. With the description of the implementation in section 3, it is possible to reproduce the entire experiment. However, the results will vary, unless the same model and explanations are used.

5.2 Benefit to Drug Research

The goal of this project was to establish if comprehensiveness and sufficiency are applicable to MPP as a measure of faithfulness. If so, they could be used to evaluate the goodness of GNN explanations and be indirectly beneficial to society by potentially reducing the time and money costs associated with drug research. Even if the metrics performed well with MPP, it is important for experts to remain sceptical and not to trust them blindly. New GNN explanation evaluation methods can be used as a tool, but the responsibility of assessing if the model predictions are trustworthy lies on the humans.

5.3 FAIR Principle

Efforts to follow the FAIR principles for scientific data management[10] have been made: the code and datasets are publicly available and open access; the code is written in a platform independent programming language; the code has been documented to allow others to use it, and the data used for the experiments has been provided.

6 Conclusions and Future Work

Drug research can benefit from new ways to evaluate the performance of explainability techniques for black-box Graph Neural Network (GNN) predictions. The BAGEL benchmark has proposed four task agnostic metrics, of which faithfulness measured as comprehensiveness and sufficiency of a molecule was adapted for molecular property prediction (MPP) in this work. The following research question was posed:

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

This question was answered by two subquestions:

1. How can comprehensiveness and sufficiency from the BAGEL benchmark be adapted to work with MPP?
2. How large are the differences in comprehensiveness and sufficiency between relevant explanations from Integrated Gradients and random explanations, using a Communicative Message Passing Neural Network model trained on a benzene ring dataset?

To adapt comprehensiveness and sufficiency for MPP, an algorithm was introduced to extract chemically valid explaining and non-explaining subgraphs of a given molecule if they exist, and provided a method to calculate the comprehensiveness and sufficiency of a GNN explanation. It was shown that the model predictions of subgraphs were similar to one another, regardless of if they contained a benzene ring or not.

To compare explainers, the average comprehensiveness and sufficiency scores were obtained for explanations produced by an Integrated Gradient (IG) explainer, and random guesses. The comparison established that comprehensiveness or sufficiency are not useful metrics for MPP, because 1) the attained scores between IG and guessing were very similar to each other, meaning that the quality of explainers is difficult to establish, 2) comprehensiveness ignores much of the input data due to the GNN requiring chemically meaningful input, and 3) the average sufficiency of random guesses is closer to a perfect explanation than that of IG.

As they have been implemented in this work, neither comprehensiveness nor sufficiency can be recommended for evaluating GNN explainer faithfulness. Comprehensiveness is limited because both split methods ignore much of the data, and both metrics are unable to create meaningful differences between the evaluated explainers.

Future Work

The usefulness of sufficiency seems to require having a ground truth of the predicted property on which to base a target value. This excludes all prediction tasks which don't have a ground truth. It is recommended to investigate the applicability of faithfulness metrics that do not require one, such as RDT-Fidelity from the BAGEL benchmark.

Generally, it appears that explainer evaluation metrics for molecular property prediction should not split molecules, as the fragments can not be guaranteed to be chemically valid.

References

[1] C. Sarkar, B. Das, V. S. Rawat, J. B. Wahlang, A. Nongpiur, I. Tiewsoh, N. M. Lyngdoh, D. Das, M. Bidarolli, and H. T. Sony, "Artificial intelligence and machine learning technology driven modern drug discovery and development," *International Journal of Molecular Sciences*, vol. 24, p. 2026, Jan. 2023.

[2] D. Sun, W. Gao, H. Hu, and S. Zhou, "Why 90% of clinical drug development fails and how to improve it?," *Acta Pharmaceutica Sinica B*, vol. 12, pp. 3049–3062, July 2022.

[3] J. Rao, S. Zheng, and Y. Yang, "Quantitative evaluation of explainable graph neural networks for molecular property prediction," 2021.

[4] Z. Wu, J. Wang, H. Du, D. Jiang, Y. Kang, D. Li, P. Pan, Y. Deng, D. Cao, C.-Y. Hsieh, and T. Hou, "Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking," *Nature Communications*, vol. 14, May 2023.

[5] M. Rathee, T. Funke, A. Anand, and M. Khosla, "Bagel: A benchmark for assessing graph neural network explanations," 2022.

[6] A. Longa, S. Azzolin, G. Santin, G. Cencetti, P. Liò, B. Lepri, and A. Passerini, "Explaining the explainers in graph neural networks: a comparative study," 2023.

[7] Y. Song, S. Zheng, Z. Niu, Z.-H. Fu, Y. Lu, and Y. Yang, “Communicative representation learning on attributed molecular graphs,” in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI’20, 2021.

[8] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” 2017.

[9] J. Rao, S. Zheng, Y. Song, J. Chen, C. Li, J. Xie, H. Yang, H. Chen, and Y. Yang, “Molrep: A deep representation learning library for molecular property prediction,” *bioRxiv*, 2021.

[10] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, and et al., “The fair guiding principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, Mar. 2016.

A CMPNN Model Training and IG Explanation Metrics

Table 9: Training accuracy of MolRep CMPNN implementation

(a) Without bond importances generated

Epochs	ACC	AUC	F1	Precision	Recall
1	0.756	0.843	0.765	0.743	0.789
2	0.752	0.848	0.759	0.746	0.772
3	0.495	0.820	0.000	0.000	0.000
4	0.505	0.885	0.671	0.505	1.000
5	0.495	0.821	0.000	0.000	0.000
10	0.495	0.500	0.000	0.000	0.000

(b) With bond importances generated

Epochs	ACC	AUC	F1	Precision	Recall
1	0.739	0.823	0.753	0.721	0.788
5	0.495	0.773	0.000	0.000	0.000

Table 10: Accuracy of MolRep IG implementation, corresponding to Table 6a and 6b

(a) Without bond importances

Epochs	AUROC	F1	ACC	Precision	AUROC Mean	ACC Mean
1	0.952	0.370	0.825	0.416	0.892	0.825
2	0.954	0.000	0.815	0.000	0.893	0.805
3	0.907	0.000	0.815	0.000	0.854	0.805
4	0.851	0.000	0.815	0.000	0.809	0.805
5	0.672	0.000	0.815	0.000	0.654	0.805
10	0.526	0.000	0.815	0.000	0.514	0.805

(b) With bond importances

Epochs	AUROC	F1	ACC	Precision	AUROC Mean	ACC Mean
1	0.768	0.000	0.815	0.000	0.764	0.805
5	0.501	0.000	0.815	0.000	0.520	0.805