# Influence of molecular structures on graph neural network explainers' performance

**Tim Stols**[1]
**Supervisor(s): Dr. M. Khosla**[1], **Dr. J.M. 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: Tim Stols
Final project course: CSE3000 Research Project
Thesis committee: Dr. M. Khosla, Dr. J.M. Weber, Dr. T. Abeel

**Abstract**

This study evaluates how the explainer for a Graph Neural Network creates explanations for chemical property prediction tasks. Explanations are masks over input molecules that indicate the importance of atoms and bonds toward the model output. Although these explainers have been evaluated for accuracy, no information exists on how faithful they are to the model (faithfulness), or how closely they correspond to human rationale (plausibility). Using explainability metrics to measure this, the performance of the explainer is evaluated on different subsets based on the presence of benzene rings and halogens respectively, and on molecular weight. This study reveals that benzene rings influence the plausibility performance of the explainer, showing that performance is better at higher thresholds but worse at lower thresholds. Molecular weight and the presence of halogens do have no impact on plausibility. The ratio of positive samples in a set is shown to influence the metrics used for faithfulness. To accurately evaluate the faithfulness of different subsets, they should be changed to have equal positive rates or different metrics should be used. This research can be used as a starting point to research the influence of dataset properties on explainer performance. This is useful to create better explainers, leading to better acceptance of these models.

# 1 Introduction

## 1.1 Current situation

Research into molecule design and molecular properties has enabled many innovations that benefit our society. Advancements have been made in multiple fields, like drug discovery, electronics, food production, and construction.[1–4] In molecule design, Graph Neural Networks (GNNs) can be used for molecular property prediction. They show promising results compared to traditional QSAR models.[5] et al. in a repository called MolRep.[6]
In a GNN, molecules are represented as graphs, where atoms are nodes, and bonds are edges. GNNs can accurately take into account the different bonds and substructures of the molecule (as topological data), and thus, they are an excellent fit for molecular property prediction.[7]
To show their performance potential, different GNN models were compared and quantitatively evaluated by Rao

While these models perform well, they lack transparency and are difficult to interpret.[8] This can block the acceptance of the model. To counteract this issue, in a second paper published by Rao et al, several GNN explainers were introduced into the MolRep repository.[9] These GNN explainers produce masks for the molecules, where masks contain binary (hard mask) or numerical values (soft mask) for all atoms and bonds in the input molecule, where the values of a mask represent how large the contribution of that part of the molecule is toward the model output. This increases insight into the model's decision process towards end users. Figure 1 shows an example illustration of a hard mask on a molecule.

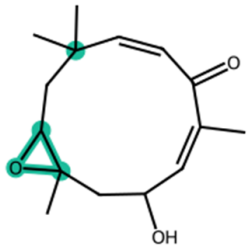


Figure 1: Illustration of a hard atom mask explanation generated for the molecule 3-Hydroxy-1,5,9,9-tetramethyl-12-oxabicyclo[9.1.0]dodeca-4,7-dien-6-one (index 92105 in the 3MR dataset).[6] Colored atoms and bonds are deemed by the explainer to be of great importance for the GNN's decision.

While the research in the MolRep repository contains a quantitative analysis using accuracy and f1-score, the research does not evaluate whether the explainer is faithful to the model. It could be that the explainer answers the correct ground truths, while the model does not use these ground truths for the classification task. This must not happen, since this mismatch between the explainer and the model can lead to incorrect interpretation of the results and errors in real-world uses. In research by Rathee et al., the BAGEL benchmark is proposed, which introduces four new eXplainable Artificial Intelligence (XAI) evaluation regimes for GNN explainers.[10] Each BAGEL regime evaluates a property of GNN explainers. Examples of these are how an explainer's output corresponds to human rationale, and to what degree the explainer selects compact but informative substructures of the molecules. Evaluation of these BAGEL regimes aids in the benchmarking and comparison of explainers, leading to better explainers and better interpretability of GNN models. The explainers for chemistry in the MolRep repository were not evaluated with these metrics. Doing so can lead to new insights about the explainers' behavior for chemical prediction models.
The four regimes in BAGEL are those of faithfulness, sparsity, correctness, and plausibility. This bachelor thesis mainly focuses on the faithfulness and plausibility metric. Faithfulness represents how closely the given explanation corresponds to the actual model's "thinking" process, and plausibility compares the explanation to a ground truth, where the ground truth is an explanation given by human beings.

Although the BAGEL paper evaluates nine explainers on four GNN models, there is no analysis present of the datasets that are used. The underlying properties of these datasets may influence the values of the metrics that are used to evaluate the explainers. When these influences are not acknowledged, there is a risk of incorrect conclusions and a false comparison between the explainers, models, and datasets. In chemistry, different groups of molecules can vary strongly in chemical properties, because of specific substructures or other properties they have. There is no knowledge of how the performance of explainers varies for subsets of datasets either containing or not containing specific subgroups. In this research, different splitting criteria are used to create subsets of datasets, and the performance of these subsets is compared. A first step is taken in the investigation of the influences of different molecular properties on the performance of GNN explainers.

### 1.2 Research Aim

This research evaluates whether the performance of the GNN explainers is the same for different types of molecules in a dataset when split based on specific substructures. The main question is:
**What is the impact of different dataset properties on GNN explainer performance?**

To assess this impact, a dataset is split up using splitting criteria corresponding to the following research sub-questions:

1. What is the impact of the presence of aromatics on the performance of GNN explainers?
2. What is the impact of the presence of halogens on the performance of GNN explainers?
3. What is the impact of molecule size on the performance of GNN explainers?

The different subsets are evaluated using metrics for the BAGEL regimes of faithfulness and plausibility.
The choice to focus on the presence of aromatics or halogens and the size of molecules has been made for the following reasons. Benzene rings have a delocalized electron cloud and strong bonding structure, making them react differently than other carbon groups.[11] Similar to aromatics, the presence of halogens in molecules strongly influences molecular properties. When halogens are present in organic molecules, they can react as electrophiles.[12]
Besides this division based on the presence of subgroups, this research also evaluates to what degree the performance of explainers varies for smaller or larger molecules. GNN models are known to perform well at handling large input sizes, as was demonstrated by Ying et al. on the Reddit-Binary dataset, which has 429.63 nodes per graph on average.[13] Understanding the impact of input size and the presence of the subgroups on the explainer's performance will aid in prescribing in which cases the use of explainers gives more benefits.

The structure of this paper is as follows. Section 2 provides the background information and describes the methodology used to perform the experiment. Section 3 presents, interprets, and discusses the results. Additionally, Section 3.3 describes the responsible research practices that were used throughout the research process. Finally, in Section 4, a summary and recommendations for future work are given.

## 2 Background and Methodology

This paragraph discusses the research methodology. It first briefly describes the selection of datasets to be tested, and the choice of GNN, explainers, and BAGEL metrics to be applied, before sharing details about the working order.

## 2.1 Datasets

The dataset that will be reported on in this study is the 3MR dataset, provided by Rao et al..[9] It contains molecules gathered from the ZINC15 dataset by Sterling et al., which are labeled with binary target data (a column of 0s and 1s) representing whether or not they contain a 3-membered ring (3MR).[14] 3MRs are substructures of three molecules, two carbon and one non-carbon.[15] The dataset contains 2877 molecules in total, of which 1745 contain a ring (76%). The MolRep model uses a pre-labeled test set, which is a subset of the 3MR dataset. 244 out of 576 molecules in the test set contain 3MRs (42%).
Two other datasets from the MolRep repository were also used for the experiment: the Benzene dataset and the Hepatotoxicity dataset.[16]

## 2.2 GNN Models

The Communicative Message Passing Neural Network (CMPNN) is used as baseline GNN for the prediction of molecular properties since it performs best in the results provided by Rao et al..[9,17] An illustration of CMPNN message passing is given in Figure 2. CMPNN is chosen, because the intuition behind the model is a good fit for chemical purposes. The main difference from MPNN networks is that the CMPNN uses message passing which first computes hidden node values (Equation 1), and uses these hidden nodes to compute the hidden edge values (Equation 2).[17] This procedure combines both atom and bond information to pass to neighbor nodes. This is important in chemistry since reactivity is dependent on both the atoms and the bond structure through which they are connected.

$$h^k(v) = \textsc{Communicate}\left(\textsc{Aggregate}\left(\{h^{k-1}(e_{u,v}) \mid u \in N(v)\}\right), h^{k-1}(v)\right) \tag{1}$$

where $h^k(v)$ is the hidden state of node $v$ at layer $k$,
Aggregate collects information from the neighbors $N(v)$ of node $v$,
Communicate updates the node's hidden state based on the aggregated information and the previous hidden state $h^{k-1}(v)$.

$$h^k(e_{v,w}) = \sigma\left(h^0(e_{v,w}) + W \cdot \left(h^k(v) - h^{k-1}(e_{v,w})\right)\right) \tag{2}$$

where $h^k(e_{v,w})$ is the hidden state of edge $e_{v,w}$ at layer $k$,
$h^0(e_{v,w})$ is the initial edge feature,
$W$ is a weight matrix,
$h^k(v)$ is the hidden state of node $v$ at layer $k$,
$h^{k-1}(e_{v,w})$ is the hidden state of edge $e_{v,w}$ at the previous layer,
$\sigma$ is an activation function.

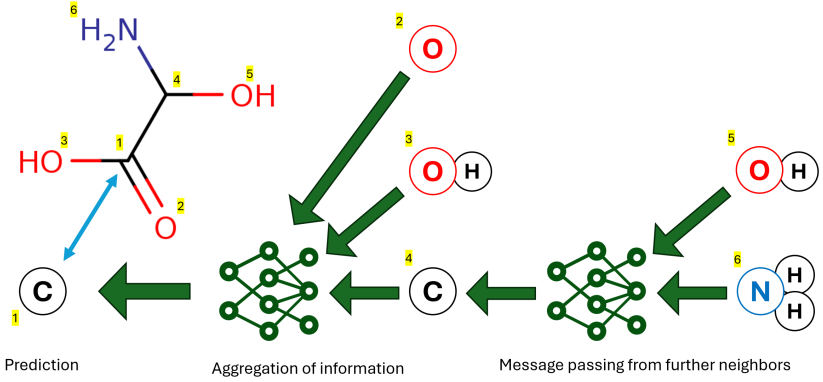


Figure 2: Example illustrating the message passing and feature aggregation of $\alpha$-hydroxyglycine. Adapted from Zhang et al..[18] The image shows how during GNN training, different neighboring atom features are passed, allowing the model to search for structural connections.

## 2.3 Explainers

To explain GNN models, different graph explainers can be used.[13,19–22]
Evaluation by Rao et al. has shown that the IG method performs best out of these graph explainers when evaluated using existing ground truths for chemical datasets. The IG explainer tries to approximate the gradients of the model's output along the path from baselines back to inputs[21] (Formula 3). This produces a soft mask over all edges and nodes of the input graph. It can be converted into a hard mask using a threshold as binary activation.

$$\mathrm{IG}_i(\mathbf{x}) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial x_i} \, d\alpha \tag{3}$$

where $\mathrm{IG}_i(\mathbf{x})$ is the Integrated Gradient for the $i$-th feature,
$\mathbf{x}$ is the input for which we want to compute the attributions,
$\mathbf{x}'$ is the baseline input (typically a zero vector or a mean input),
$F$ is the function representing the neural network,
$\alpha$ is a scalar that interpolates between the baseline and the input,
$\frac{\partial F}{\partial x_i}$ is the gradient of the function $F$ with respect to the $i$-th input feature.

## 2.4 BAGEL Metrics

To assess the performance of the IG explainer under different conditions, the metrics proposed in the BAGEL benchmark are used.[10] The four regimes proposed in BAGEL are faithfulness, sparsity, correctness, and plausibility. Faithfulness and plausibility are explained in this section, as sparsity and correctness are not used in this research.

Faithfulness concerns the capability of an explainer to correctly identify the GNN model's action. The explainer should correctly show which parts of the molecule the model uses to make its prediction. Three metrics exist for measuring faithfulness: Rate DisTortion-based

fidelity (RDT Fidelity), comprehensiveness, and sufficiency.[10]
RDT fidelity is only used for explanations that are not masks.[23] This is not applicable in this research.
Comprehensiveness measures whether all the nodes and edges that are needed for a prediction are selected. The model prediction for the non-explaining part of the input is subtracted from the original prediction, as shown in Equation 4.

$$Comprehensiveness = f(G) - f(G \setminus G_E) \tag{4}$$

where $f(G)$ is the original prediction for the full graph $G$,
$f(G \setminus G_E)$ is the model prediction for the graph without the explainer subgraph $G_E$.

Sufficiency measures whether the explaining nodes and edges are sufficient to reach the original prediction. It is calculated by subtracting the model prediction for just the explaining subgraph from the original prediction, shown in Equation 5.

$$Sufficiency = f(G) - f(G_E) \tag{5}$$

To evaluate comprehensiveness and sufficiency, a method developed by research project peer Heli Pajari was used in this project.[24] As explained in Pajari's research paper, the method to calculate these values is not always applicable in chemical cases. Sufficiency and comprehensiveness assume that any subgraph of the input is also valid, which is not true in chemistry. As a result of this issue, the comprehensiveness or sufficiency cannot be calculated for all molecules.
To evaluate plausibility, ground truths consist of the 3-membered rings that are present in the datasets. By comparing these ground truths to the generated explanations, the overlap between the human explanation and the explainer is evaluated. As prescribed in the BAGEL paper, plausibility is measured through the Area Under the Precision-Recall Curve (AUPRC) for soft masks.

The two other regimes of sparsity and correctness are not evaluated in this research. Since the size of ground truths is pre-determined, sparsity, which evaluates the size of the explanation, is not relevant.[25] Correctness, which concerns how well the explainer detects externally injected correlations, is also not relevant, because a chemical dataset is used with only objective measurements of properties.[10]

## 2.5 Implementation Methodology

To obtain results to answer the research question, new code is implemented on top of the MolRep repository.

We choose to implement the BAGEL metrics in separate new Python classes inside the MolRep repository rather than adding and executing the entire BAGEL codebase. Only the *metrics.py* file in the BAGEL repository is relevant for this project.

The notebook steps are outlined in Figure 3. The first three blocks consist of functionalities that were adapted and modified from the MolRep repository, to train the chosen GNN models and explainers on the correct datasets.
The splitting and the evaluation using BAGEL metrics were implemented in separate classes, and new components were added to the notebook to use this code and run it with the trained model and the obtained explanations. This is shown in the last two blocks of Figure 3.

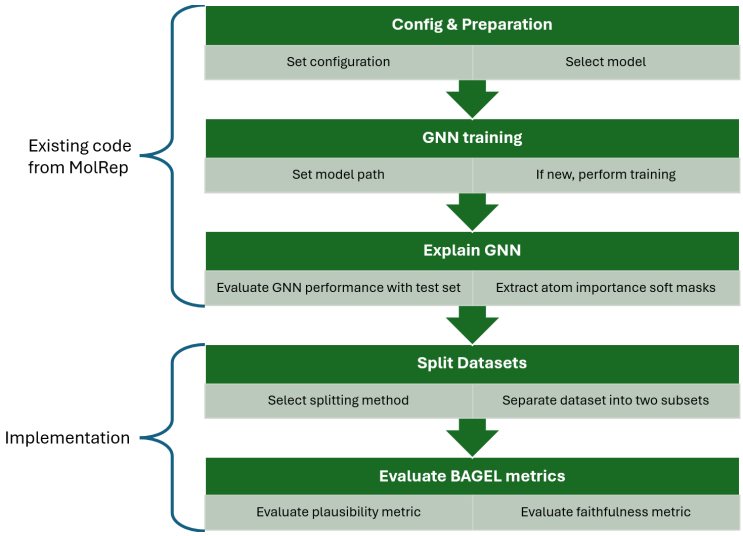


Figure 3: Flow chart of the functionality in the implemented Jupyter Notebook Explainer_Experiments_.py

The currently selected splitting methods are for aromatics, halogens, and molecular weight. By providing the median value for molecular weight, the set can be split into two equal-sized subsets. For plausibility, the ROC curve charts of both subsets are plotted both in separate figures and together in one figure. The ROC curve for the total dataset is also shown. For faithfulness, the comprehensiveness and sparsity values are gathered for all molecules in the set. The values are then split into subsets, and the values for the subsets are used to generate boxplots. An example of the charts generated can be found in Appendix A.

# 3 Results and Discussion

## 3.1 Observed GNN performance

The CMPNN model did not perform as expected. In Table 1, the observed performance is documented. Whereas Rao et al.'s research claims that the GNN has above-average classification performance, the observed model performed about equal to a random classification (as 0.5 is the expected AUC value for a random classifier).[9]
Because these results were lower than expected, the experimental procedure was adjusted. There is no point in assessing the explainer of a model that has no significant performance. Therefore, the rest of the experiment was conducted on the 3MR dataset, the only dataset for which the GNN seems to converge.

Table 1: Performance metrics for different datasets, after training the CMPNN model for 100 epochs. Cells above 0.8 are colored green, and cells below are colored red. The values in brackets are claimed performance values by Rao et al. [9].

| Dataset | Epochs | ACC | AUC | F1 | Precision |
|---|---|---|---|---|---|
| 3MR | 100 | 0.98 | 0.99 (1.000) | 0.98 | 0.96 |
| Benzene | 100 | 0.5 | 0.5 (1.000) | 0 | 0 |
| Liver | 100 | 0.44 (0.681) | 0.53 | 0.2 | 0.15 |

**Discussion**

There are a few possible reasons for this underperformance. There can be problems in the learning process, incorrect hyperparameters, or the use of deprecated packages in the original repository. Verifying this falls outside of the scope of this research project.

These results lead to questions about the validity of the MolRep research. However, before conclusions are drawn, further research should be done by other researchers to see whether this model can be made to converge.

Overall, the current reproducibility of the results using the MolRep repository is deemed insufficient. There is little documentation on how to run the code, and many of the packages used are deprecated, meaning that many changes are needed to run the experiment or add new functionalities.

## 3.2 Explainer Performance

Following the initial evaluation described in Section 3.1, the 3MR dataset was split using the developed notebook, using the benzene, halogen, and molecular weight filters. The sizes of the split datasets are given in Table 3 in Appendix A.1.

Some subsets are larger than others. Especially in the case of halogens, the set not containing halogens is almost five times as large. Although this can impact the results, the decision was made not to trim the larger subsets to the same size as the smaller ones. This would only decrease the accuracy of measurements on the larger subsets. The GNN model is trained on the complete dataset, after which the explanations of the two subsets are evaluated.

### 3.2.1 Faithfulness metric

Figures 4-6 in Appendix A show the sufficiency scores measured for the dataset when split on the benzene, halogen, and molecular weight splitting criteria. For sufficiency, a score near 0 indicates better performance. The p-values in the figures represent the probability of observing this difference purely by chance, assuming a t-distribution. A threshold of 0.01 will be used for the p-value.

For the set separated on aromatics, the subset not containing aromatics has a mean closer to 0. For the set separated on halogens, the subset not containing halogen also has a mean closer to 0. Since the p-values for the first two tests were below the threshold, the null hypothesis is rejected, showing that there is a correlation between the splits and sufficiency. For the molecular weight, the p-value is above the threshold (0.164), so the null hypothesis is accepted and there is no significant difference.

Figures 7-9 in Appendix A show the measured comprehensiveness scores for the dataset when split on different criteria. For comprehensiveness, a score near 1 indicates better per-

formance. The sets containing aromatics, containing halogens, and with high molecular weight all perform worse than their counterparts. For the first two, the p-value is below the threshold, meaning the null hypothesis is rejected and there is a correlation between the splits and comprehensiveness. For the molecular weight, the p-value is again above the threshold (0.04), so the null hypothesis is accepted.

Additional analysis was conducted on the generated subsets, to evaluate the positive rate (ratio of molecules with target value 1) in the sets. This positive rate seems to closely correlate to both sufficiency and comprehensiveness values observed, as shown in Table 2. The difference from the mean positive rate nearly corresponds with the difference from the mean sufficiency for all subsets, as well as the difference from the mean comprehensiveness multiplied by two.

Table 2: Table showing the positive rate, sufficiency, and comprehensiveness measured on the complete dataset and experiment subsets. Difference columns (colored yellow) show the difference between the subset values and the value for the complete dataset. The yellow values in these rows match closely.

| Subset | Positive rate | Difference from complete | Sufficiency score | Difference from complete | Comprehensiveness score | Difference from complete ×2 |
|---|---|---|---|---|---|---|
| Complete | 0.4236 | - | -0.08 | - | 0.26 | - |
| Containing benzene | 0.3062 | 0.1174 | -0.19 | 0.11 | 0.2 | 0.12 |
| Not containing benzene | 0.6329 | -0.2093 | 0.08 | -0.16 | 0.36 | -0.2 |
| Containing halogens | 0.1111 | 0.3125 | -0.33 | 0.25 | 0.07 | 0.38 |
| Not containing halogens | 0.4885 | -0.0649 | -0.03 | -0.05 | 0.29 | -0.06 |
| High MW | 0.3175 | 0.1061 | -0.1 | 0.02 | 0.23 | 0.06 |
| Low MW | 0.4757 | -0.0521 | -0.05 | -0.03 | 0.3 | -0.08 |

**Discussion**

The reason why this positive rate influences sufficiency is that the model is not familiar with the substructures that are passed as model input when evaluating $f(G_E)$. These substructures are small parts of the molecules, with which the model is not familiar. As a consequence, the model often outputs values around 0.5, indicating that it does not know how to classify these molecules. The sufficiency will then be around $0 - 0.5 = -0.5$. Because negative samples give negative sufficiency values, the positive rate influences the average sufficiency scores of the subsets.

While this explains the statistically significant difference in sufficiency scores, this correlation between the positive rate and the sufficiency means that no relevant influences of the actual splitting criteria can be determined based on the results.

A similar issue occurs in the case of comprehensiveness (Table 2). The comprehensiveness for negative samples is 0 subtracted by $f(G \setminus G_E)$. The molecule(s) of $G \setminus G_E$ are, just like for sufficiency, often much smaller than the GNN model is trained for. The prediction for these molecules is around 0.5, and the comprehensiveness score is around -0.5 for negative samples. Therefore, the positive rate also influences the comprehensiveness score.

This again explains the significant difference in comprehensiveness scores, but it means that no other relevant influences of the subset data properties can be determined from these results.

### 3.2.2 Plausibility metric

Figures 10-13 show the observed AUROC curves for the explainers, measured on the subsets containing or not containing aromatics, respectively halogens. The area under the curve is shown in the legend.
The difference in area under the curve is not statistically significant for the two aromatics curves. However, the shape of the curves does differ. The subset containing aromatics performs better with a lower false positive rate than the other. Conversely, the subset not containing aromatics performs better with a high false positive rate.
Similar to the aromatics, the difference in area under the curve for the halogen subsets is not statistically significant. The curve for the subset containing halogens is jumpy and contains many straight-line parts, probably because the size of this dataset is quite small. Although this is not a continuous line, the rough shape of the two lines for halogens is nearly the same.
The results for molecule size do not show any significant difference. The lines have roughly the same shape, and the area under the curve is nearly equal.

**Discussion**
The different shape of the ROC curves for the aromatic and non-aromatic subsets occurs because the explainer either highlights the entire benzene ring or does not highlight it at all. Since the ring contains six atoms, it is usually a substantial part of the molecule. In the FPR range from 0.0 to 0.2, the atoms that are highlighted as explanations are not in the benzene rings and have a higher chance of being in the 3MR. However, as the threshold lowers, a tipping point can be reached, where the explainer will highlight the benzene ring shapes as relevant. This would explain why the ROC curve for the set containing benzene is lower than the other curve in the interval from 0.6 to 0.9 FPR. In this interval, the benzene rings are incorrectly given high explanation values, which causes the FPR to increase faster than the true positive range (TPR).
Halogens could be a factor that the model takes into account since 3MRs do not contain halogens, but this does not appear to be the case. The ROC curves with and without halogens exhibit similar behavior. It is expected that halogens have little influence because of the low number of halogens in the dataset.
Since the ROC curves for heavier and lighter molecules are very similar and the difference in AUROC is not significant, there is no influence of molecular weight on explainer plausibility in this case. However, as the dataset only contains molecules within the range of 148.2 to 349.8, not all sizes of molecules could be tested.

## 3.3 Responsible Research

### 3.3.1 Trust and safety considerations

This project aims to evaluate explainers of GNN models for chemical property prediction purposes. In itself, it does not seem to introduce any direct human risks, however, the effects of the use of this technology need to be evaluated to discuss possible risks that can emerge in future use cases. An important aspect of the development of these GNN models and explainers is that they are designed to be used by an expert, not to replace the expert. Black-box neural networks are inherently unpredictable and while they can converge to reach incredible predictive powers, they will always keep some degree of error rate. The predictions made by the model cannot be assumed to be true, as they are non-justifiable

guesses for new molecules, based on previous knowledge of other molecules.

Although the implementation of GNN explainers aids in showing the decision process of the models, the explanations provided cannot be assumed to be valid reasons, neither should they be assumed to show the correct decision process of the model. Users of GNN explainers need to be aware of how these explainers work, and how the explainers can also be prone to mistakes. When molecular property predictions made are interpreted as being the truth, this can lead to unwanted consequences, depending on the type of property that is predicted and for what purpose this is. When results are used in practice, they need to be verified if possible, and otherwise, they need to be labeled as what they are, statistical predictions.

In addition to the ethical risk considerations for the explainers, the results proposed by this research could contain errors due to insufficient testing of the implementation. As the research concerns a research project conducted over a relatively short period of 10 weeks by a single student, it was not feasible to thoroughly test all the code implemented. To generalize this research and further reinforce the conclusions, more testing needs to occur, using a testing pyramid containing unit, integration, system, and user tests.

### 3.3.2 Privacy

Since the data used for this research is publicly available and concerns molecular data, which is in no way related to human individuals, there is no privacy risk. The data can not be traced back to any nationalities, races, or genders of human individuals.

### 3.3.3 Reproducibility and repeatability

The experiments conducted are both repeatable and reproducible. The repository containing all the developed products is publicly available. It contains clear documentation on how to run the code. The notebooks have code documentation, detailing all the steps that are executed. By executing these steps, the results for the experiment can be achieved, using the pre-trained model that is present in the repository. This model is the same that was used to achieve the results. This makes the experiment repeatable.

To achieve reproducibility, the research paper as well as the codebase contains explanations and documentation on the steps that are taken to perform the experiment. For the existing parts of the functionality that were adapted from the MolRep repository, there is a description for how this was done, and where the code was taken from. Although the exact results can vary slightly over iterations of the training process, the experiment was conducted multiple times leading to highly corresponding results.

In addition, efforts to follow the FAIR principles for research software were made. Public availability and linking to the open-source repository from the paper aid in the findability and accessibility of the research. The selection of Python as a programming language makes the code interoperable across operating systems. Since the developed product is in the form of separated Python classes and Jupyter notebook blocks, it is inherently reusable for similar or different purposes.

# 4 Conclusions and Future Work

## 4.1 Main Conclusions

This research investigates whether the performance of GNN explainers is influenced by data subsets with different molecular properties. There is a specific focus on how the values for faithfulness and plausibility metrics are influenced by these molecular properties. This was done by splitting the dataset using three splitting criteria: the presence of benzene rings or halogens respectively, and the molecular weight being above or below the median for a dataset.
By learning how explainers perform on data subsets with different properties, more can be learned about in which cases explainers perform great and for which circumstances they should be improved. Through this process, better-performing explainers can be used to make GNN models more interpretable and transparent. This aids in their acceptance and makes them more reliable models to be used in field research.

For the BAGEL regime of faithfulness, a clear difference was observed in both sufficiency and comprehensiveness for all dataset splits. However, an investigation into the fraction of positive samples for each subset showed that this difference is likely caused mainly by the impact of the number of positives and negatives for each subset. Because of this, the actual influence of the subset's molecular properties on faithfulness was not determined. However, the conclusion that the positive rate is an influence factor for the faithfulness metrics is also very relevant. When different datasets are explained and evaluated with BAGEL metrics, the difference in faithfulness values can incorrectly be interpreted as a difference in explainer performance.

For the plausibility regime, no significant differences were observed in the Area Under ROC curves. This shows that the chosen subsets do not perform better or worse than their counterparts or the complete dataset. However, for the benzene splitting, it was observed that the set containing benzene performs better for more reluctant explainers with a high threshold, and worse for eager explainers with a low threshold. This is because the explainer has some turnover point where it either gives the entire benzene ring as an explanation or none of it. Although the presence of benzene rings does not impact the overall accuracy of the explainer, there is an impact on the accuracy when a set threshold is chosen and the explanations are converted to hard masks.

By taking these structural differences into account in future research, efforts can be made to limit the impact of the use of varying datasets on the BAGEL metrics evaluation.

## 4.2 Limitations and Challenges

These results give a great start in the evaluation of chemical GNN explainers using BAGEL metrics. However, more evaluation will be needed to better understand the behavior of these explainers and their strengths and flaws.

Firstly, the experiment procedure should be done on different chemical datasets. The 3MR dataset is relatively small, and the molecules are all of a comparable size. Performing this experiment on larger datasets with a wider range of molecules could show new results or confirm observations from this research. An example of this is the ROC curve for the

halogen-containing subsets. Due to a low amount of molecules with halogens in the dataset, it was not possible to plot a smooth ROC curve and properly compare the shapes.
The other MolRep datasets could yield more information. The Liver dataset is interesting in particular since liver toxicity classification is a complicated task that cannot be performed perfectly by humans. Previous results show that GNNs can outperform humans in this.

Changes in the method of BAGEL evaluation could also lead to new knowledge about GNN performance. The influence of the positive rate on the faithfulness metrics should be minimized. This could be achieved by removing samples from the subsets to give them a similar positive rate, or by only evaluating faithfulness for positive samples. However, this would give a lower sample size, decreasing the statistical significance of the results. Only investigating positive samples also limits the generalizability of the results. When a larger dataset is available, these changes could be viable.

# References


(1) Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **2004**, *1*, 337–341.

(2) Sun, L.; Diaz-Fernandez, Y. A.; Gschneidtner, T. A.; Westerlund, F.; Lara-Avila, S.; Moth-Poulsen, K. Single-molecule electronics: from chemical design to functional devices. *Chem. Soc. Rev.* **2014**, *43*, 7378–7411.

(3) Kauffmann, A. C.; Castro, V. S. Phenolic Compounds in Bacterial Inactivation: A Perspective from Brazil. *Antibiotics* **2023**, *12*, 645.

(4) Huang, J.; Chen, R.; Zhou, Y.; Ming, J.; Liu, J. Molecular design and experiment of ion transport inhibitors towards concrete sustainability. *Cement and Concrete Composites* **2022**, *133*, 104710.

(5) Song, Y.; Zheng, S.; Niu, Z.; Fu, Z.-h.; Lu, Y.; Yang, Y. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, ed. by Bessiere, C.; Main track, International Joint Conferences on Artificial Intelligence Organization: 2020, pp 2831–2838.

(6) Rao, J.; Zheng, S.; Song, Y.; Chen, J.; Li, C.; Xie, J.; Yang, H.; Chen, H.; Yang, Y. MolRep: A Deep Representation Learning Library for Molecular Property Prediction. *bioRxiv* **2021**.

(7) Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57–81.

(8) Yuan, H.; Yu, H.; Gui, S.; Ji, S. Explainability in Graph Neural Networks: A Taxonomic Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, 1–19.

(9) Rao, J.; Zheng, S.; Lu, Y.; Yang, Y. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *Patterns* **2022**, *3*, 100628.

(10) Rathee, M.; Funke, T.; Anand, A.; Khosla, M. BAGEL: A Benchmark for Assessing Graph Neural Network Explanations. *arXiv preprint arXiv:2206.13983* **2022**.

(11) Moran, D.; Simmonett, A.; Leach, F.; Allen, W.; Schleyer, P.; Schaefer, H. Popular Theoretical Methods Predict Benzene and Arenes To Be Nonplanar. *Journal of the American Chemical Society* **2006**, *128*, 9342–9343.

(12) Gribble, G. W., *Naturally Occurring Organohalogen Compounds - A Comprehensive Update*, Retrieved April 23, 2022; Springer: 2009.

(13) Ying, R.; Bourgeois, D.; You, J.; Zitnik, M.; Leskovec, J. GNNExplainer: Generating Explanations for Graph Neural Networks, 2019.

(14) Sterling, T.; Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.

(15) Smith, M.; March, J., *March's Advanced Organic Chemistry: Reactions, Mechanisms, and Structure*; Wiley: 2007.

(16) Liu, R.; Yu, X.; Wallqvist, A. Data-driven identification of structural alerts for mitigating the risk of drug-induced human liver injuries. *Journal of Cheminformatics* **2015**, *7*, 1–8.

(17) Song, Y.; Zheng, S.; Niu, Z.; Fu, Z.-H.; Lu, Y.; Yang, Y. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 2020, pp 2831–2838.

(18) Zhang, Y.; Luo, M.; Wu, P.; Wu, S.; Lee, T.-Y.; Bai, C. Application of Computational Biology and Artificial Intelligence in Drug Design. *International Journal of Molecular Sciences* **2022**, *23*, 13568.

(19) Luo, D.; Cheng, W.; Xu, D.; Yu, W.; Zong, B.; Chen, H.; Zhang, X. Parameterized explainer for graph neural network. *Advances in neural information processing systems* **2020**, *33*, 19620–19631.

(20) Akkas, S.; Azad, A. In *Proceedings of the ACM on Web Conference 2024*, ACM: 2024.

(21) Jiménez-Luna, J.; Skalic, M.; Weskamp, N.; Schneider, G. Coloring Molecules with Explainable Artificial Intelligence for Preclinical Relevance Assessment. *Journal of Chemical Information and Modeling* **2021**, *61*, PMID: 33629843, 1083–1094.

(22) Jin, W.; Barzilay, R.; Jaakkola, T. In Cited by: 49, 2020; Vol. PartF168147-7, pp 4799–4809.

(23) Ribeiro, M. T.; Singh, S.; Guestrin, C. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp 1135–1144.

(24) Pajari, H. Why Know When You Can Guess? *TU Delft Repositories* **2024**.

(25) Funke, T.; Khosla, M.; Rathee, M.; Anand, A. Zorro: Valid, Sparse, and Stable Explanations in Graph Neural Networks, 2022.

# A Experimental Results

## A.1 Measured Subset Sizes

Table 3: Sizes of the split datasets containing or not containing the specified group. Obtained by splitting the 3MR dataset using the developed notebook.

| Group | Contains | Does not contain |
|---|---|---|
| Aromatics | 369 | 207 |
| Halogens | 99 | 477 |
| Molecular Weight | 288 | 288 |

## A.2 Boxplots for Sufficiency Metric

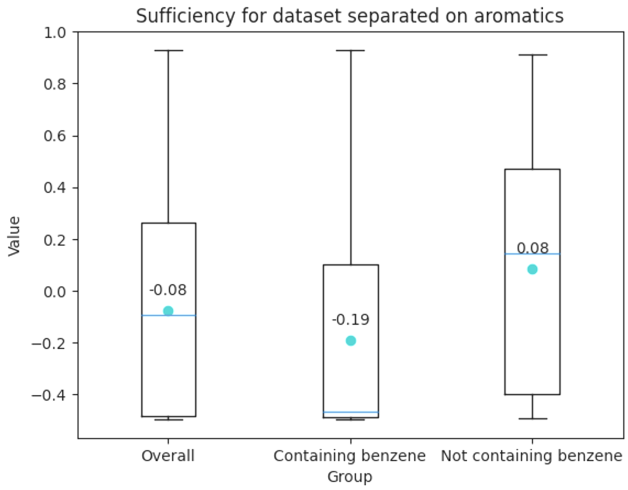


Figure 4: Box plot for measured sufficiency for the explanations for subsets split on benzene. Generated by the IG explainer on the CMPNN model. The blue dot and label indicate the mean. The p-value for the difference between the second and third group is $8.68 \times 10^{-13}$

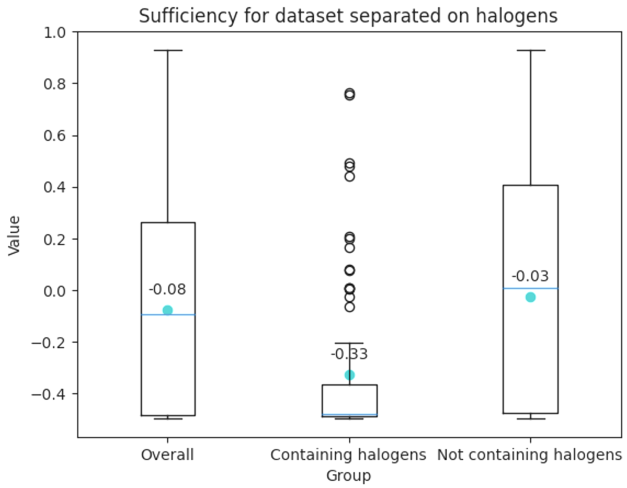


Figure 5: Box plot for measured sufficiency for the explanations for subsets split on halogens. Generated by the IG explainer on the CMPNN model. The blue dot and label indicate the mean. The p-value for the difference between the second and third group is $4.32 \times 10^{-9}$

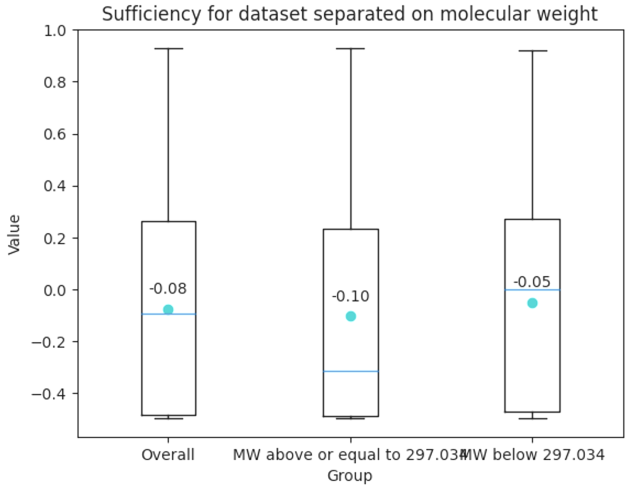


Figure 6: Box plot for measured sufficiency for the explanations for subsets split on molecular weight. Generated by the IG explainer on the CMPNN model. The blue dot and label indicate the mean. The p-value for the difference between the second and third groups is 0.164

## A.3 Boxplots for Comprehensiveness Metric

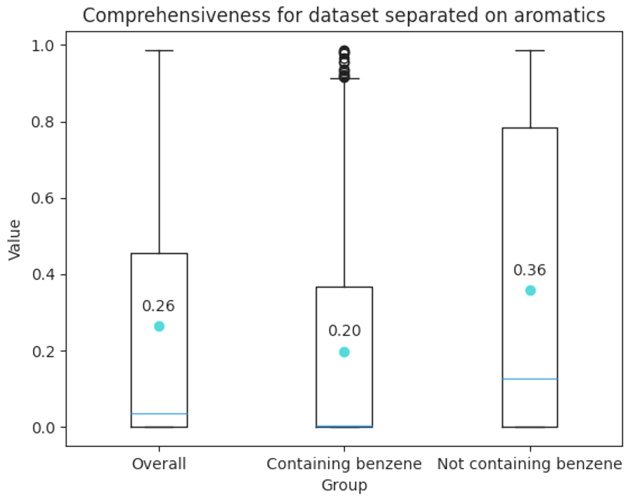


Figure 7: Box plot for measured comprehensiveness for the explanations for subsets split on benzene. Generated by the IG explainer on the CMPNN model. The blue dot and label indicate the mean. The p-value for the difference between the second and third group is $4.03 \times 10^{-7}$

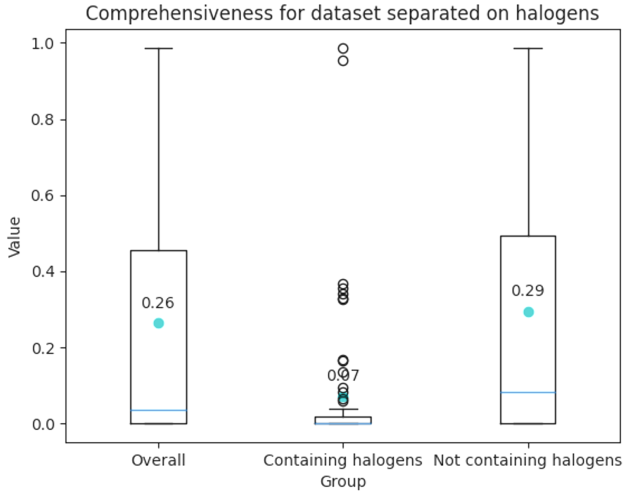


Figure 8: Box plot for measured comprehensiveness for the explanations for subsets split on halogens. Generated by the IG explainer on the CMPNN model. The blue dot and label indicate the mean. The p-value for the difference between the second and third group is $9.26 \times 10^{-7}$

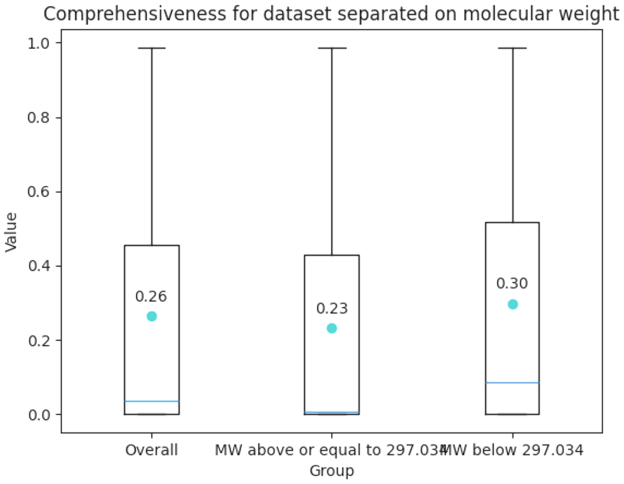


Figure 9: Box plot for measured comprehensiveness for the explanations for subsets split on molecular weight. Generated by the IG explainer on the CMPNN model. The blue dot and label indicate the mean. The p-value for the difference between the second and third groups is $3.62 \times 10^{-2}$

## A.4 ROC Curves for Plausibility Metric

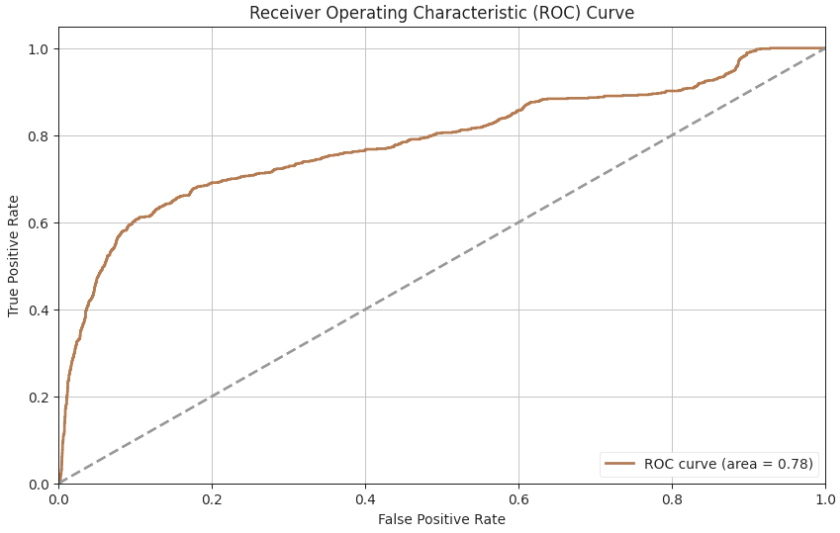


Figure 10: ROC curve for the explanations generated for the entire dataset, when compared to the ground truths for the 3MR dataset. Generated by the IG explainer on the CMPNN model. The dotted line represents a random classifier as a baseline. AUROC = 0.78

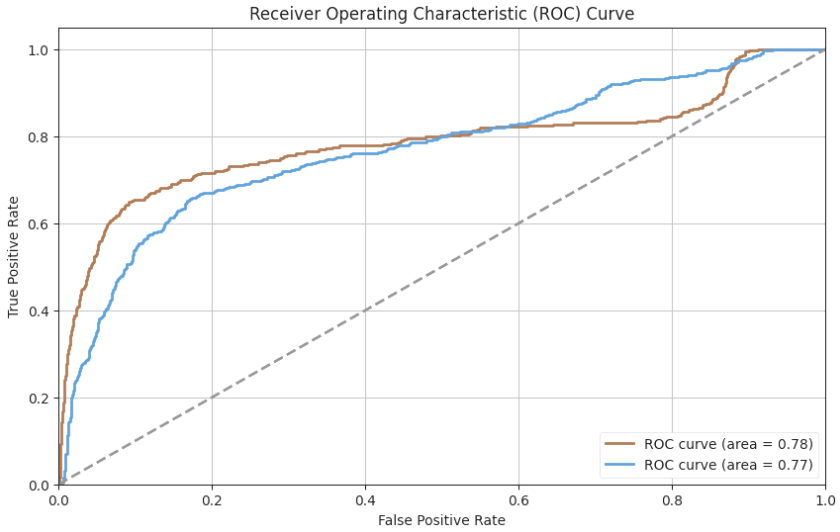


Figure 11: ROC curves for the explanations generated for molecules containing (orange) or not containing (blue) aromatics, when compared to the ground truths for the 3MR dataset. Generated by the IG explainer on the CMPNN model. The dotted line represents a random classifier as a baseline. AUROC = 0.78, 0.77 resp.

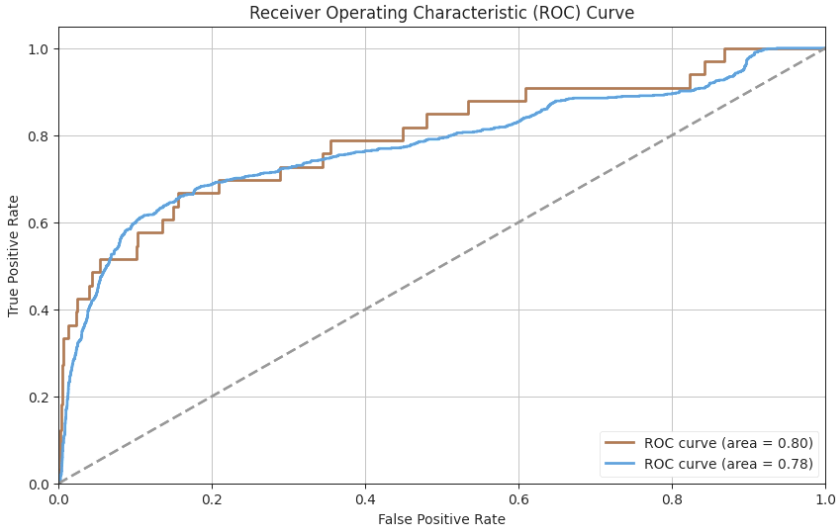


Figure 12: ROC curves for the explanations generated for molecules containing (orange) or not containing (blue) halogens, when compared to the ground truths for the 3MR dataset. Generated by the IG explainer on the CMPNN model. The dotted line represents a random classifier as a baseline. AUROC = 0.80, 0.78 resp.

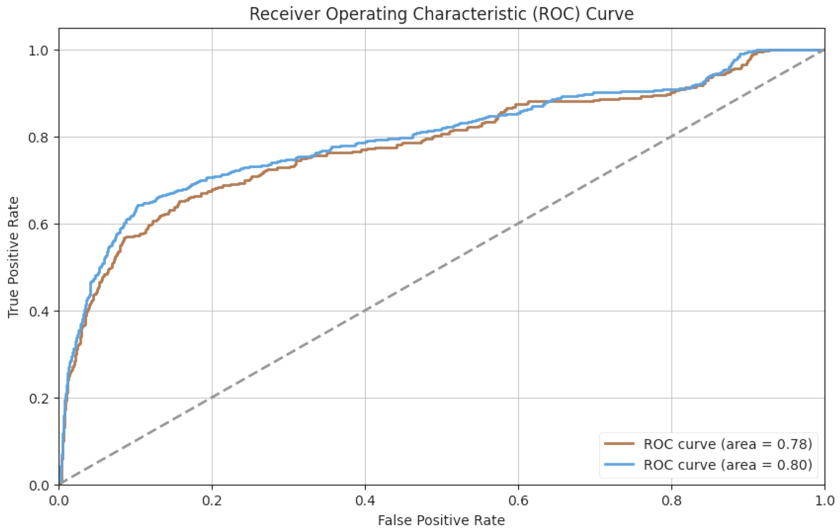


Figure 13: ROC curves for the explanations generated for molecules separated on higher and lower molecular weight (MW), when compared to the ground truths for the 3MR dataset. Generated by the IG explainer on the CMPNN model. The orange line is for molecules with MW greater or equal to 297.034, and the blue line is for molecules below this value. The dotted line represents a random classifier as a baseline. AUROC = 0.78, 0.80 resp.

# B Use of Large Language Models

Throughout this research project, the large language model (LLM) of ChatGPT was consulted multiple times for some conversational queries. The following links show the dialogues that were held with ChatGPT during the project.

1. https://chatgpt.com/share/40499fb1-7b89-4621-9cf4-2af6c71dd8dd
2. https://chatgpt.com/share/c9111c1f-5e0a-4d34-8519-fa9606ac506d
3. https://chatgpt.com/share/56df2d19-c921-4446-a674-bfbf6947cc08
4. https://chatgpt.com/share/b24b6f12-314f-481a-93ee-f2f3c4277ffa
5. https://chatgpt.com/share/1257e949-fd1f-4d24-bf41-4cfb23ea67e6
6. https://chatgpt.com/share/dbb000ce-03b7-4d2d-8248-0e9121c5a431
7. https://chatgpt.com/share/17d50248-a48c-4441-93d1-2be34a096213

While using the LLM, the responses were verified by consulting other sources. The model was not used directly in the writing parts of the thesis, and thus there are no direct citations from the LLM.