We thank the reviewers for their detailed feedback and fair reviews. We are glad to hear reviewers found our work "novel and convincing" (R2), will "attract broad interests in GNN and explainability research community" (R3), clear (all), correct (all), and reproducible (all). Nonetheless, the reviewers' critiques have highlighted areas of improvement we thoroughly address below, and will include in our final manuscript.

R1 mentioned that we do not focus on identifying relevant subgraphs that together inform the prediction generated for a given example. We respectfully disagree for two reasons. First, the ranked node-level importance values can be easily thresholded to extract relevant subgraphs. Second, our attribution performance evaluations directly evaluate (via Kendall's tau or AUROC) whether each attribution method places correct weights on relevant subgraphs (L160).

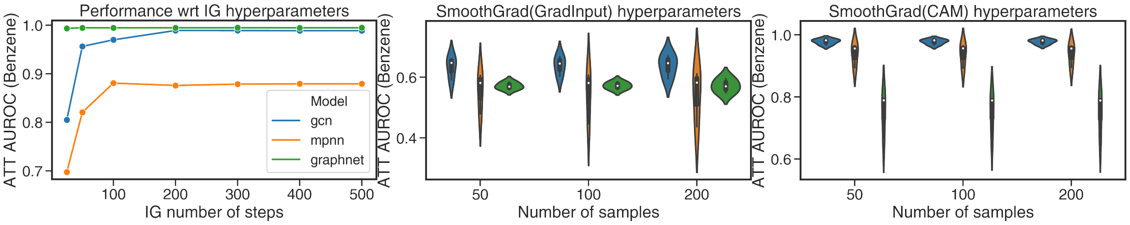
R1 notes we do not evaluate attention as a mechanism of attribution. In the present work we focused on post-hoc methods due to the many open issues around the interpretability of attention weights (Jain and Wallace, 2019; Brunner et al, 2019). However, the community will be able to readily extend our these experiments with our open-source benchmarking code.

R3 and R4 asked for results on graph types beyond molecules. We thank the reviewers for their desire to expand our approach's impact to a wider range of graph types. Our tasks are quite general because they are grounded in the theoretical task of classifying the presence or absence of logical combinations of subgraphs; we thus expect them to still be useful in other contexts. Additionally, we featurize input graphs with minimal chemical information, using a single categorical variable for nodes and edges. However, additional tasks would improve our work, so we will include the synthetic node-level tasks from GNNExplainer, and an edge-level task, as suggested by R1 and R2.

We were as surprised as R3 about the high performance of older methods like IG and CAM adapted to graphs. CAM's success might also encourage research in the direction of novel 'interpretable' architecture components (e.g. GAP layers), instead of more black box-approaches like GradCAM.

We agree with R4 that statistical tests would bolster our claims. We identified all such claims (e.g. L202, 203, 205, etc.) and performed t-tests with Holm-Bonferroni correction when comparing means and Bartlett's test when comparing variances. Our main findings are significant under these tests, and we have updated our manuscript accordingly.

R4 noted that attribution methods can be sensitive to hyperparameters. In our experiments, only IG and SmoothGrad rely on hyperparameters. We computed attribution scores on several configurations to demonstrate the variability in performance of these methods (figures below). IG is not reliable when the number of steps is too low, and SmoothGrad's performance can degrade when not tuned correctly. When trying new hyperparameters, neither method's performance improved over what we report in our paper. More detailed analysis will be included in the appendix.



R4 highlights relevant citations, which we will include. These references measure the relationship between an attribution method's weightings for image segments to a model's predictions, via selective image ablations. Unfortunately, there is no analogous ablation operation in graphs. Removing a node in an example graph will create a different graph. We may instead delete the node representation, but the adjacency matrix still contains predictive information. Our experiments and PARC score (Figs 4 and 5) address this by tracking how attribution performance tracks model performance, without ablations. For our stability experiment (analogous to Alvarez-Mills and Jaakkola, 2018), we only include perturbations that do not affect the graph's predicted label. We will add this clarification to the final version.

R4 asks about the other interpretability criteria in [3]. In the space available, we focused on four qualities from [3] that can be rigorously and quantitatively evaluated given ground-truth attributions, which are cheaply available for graph tasks. Other qualities require human evaluation, are already covered implicitly by attribution, or apply to interpretability broadly, instead of attribution specifically.

Lastly, several reviewers noted minor typos, which we have corrected. We thank the reviewers for their careful and insightful critique of our work, as our work has already improved as a result of their feedback. We have addressed each of the reviewer's critiques of our work, and hope the reviewers and AC will consider this "novel and convincing" (R2) work for publication in NeurIPS.