We thank the reviewers for their time, valuable feedback, and recommendations for improvements. Overall, the reviewers found our methodology interesting, novel and technically sound, and our contributions to be very timely. However, a key point of clarification was raised regarding the selection of hyperparameters and the effects of subgraph constraints on generated explanations. We address this key point of clarification in detail below. The four reviewers also raised important clarification points on the motivation for the use of mutual information (R5), synthetic datasets (R2, R3, R5) and quantitative experiments (R3, R5, R6), and we provide clarification on these issues below as well. These clarification issues arise—in large part—because certain details were omitted from the paper due to space constraints. However, an extra page and lengthened appendix will allow us to address these clarification points in the revised version.

Parameters & effects of constraints on explanations. R2, R3 and R5 rightly point out the need for more investigation into hyperparameters and regularization constraints. In new experiments, we observe that varying regularization strength of constraints can affect explanations (**Fig. S1**). However, this gives GNNEXPLAINER flexibility to encode domain-specific priors into constraints and, crucially, allows the explainer to balance verbosity and completeness of the explanation. In particular, without regularization, the explainer will include many edges in the explanation even though those edges only negligibly increase confidence of a GNN's prediction. Conversely, imposing a very small size constraint K_M will produce meaningless explanations, *e.g.,* single edges. In practice, we observe that prior knowledge about a task (*e.g.,* select K_M to be the size of a chemical functional group) or commonly used heuristics (*e.g.,* select K_M that results in largest increase of GNN's confidence score) determine an appropriate value for K_M. Furthermore, the range of acceptable explanation sizes is also important in determining the threshold for cutting off low-importance edges (*e.g.* how much time a user has to examine an explanation). In all experiments, we ensure that the setup is fair for all methods and use the same value for explanation size.

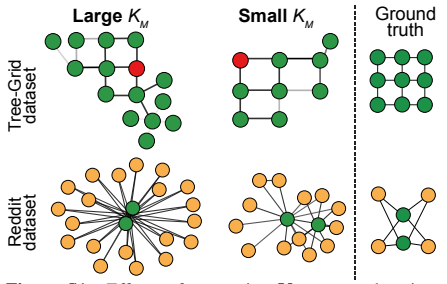

Figure S1: Effects of constraint K_M on explanations. On Reddit, we find that modifying K_M doesn't have strong effects, one can easily see the explanatory structure. Effects are stronger when these structures are of fixed size (*e.g.,* 3x3 grid) rather than general shapes (*e.g.,* star or cycle of any size). Further, on Tree-Grid dataset, GNNEXPLAINER is robust to selection of hyperparameter K_M; AUCs for $K_M = 6, 8, 10, 12, 14, 16$ are $0.71, 0.83, 0.88, 0.89, 0.86, 0.81$, respectively.

Mutual information. R5 raises a key point about motivation behind using mutual information. This is indeed fundamental, but there is intuition for using mutual information: The goal is to identify a small subgraph together with a small subset of node features that maximize confidence of a GNN's prediction. Mutual information finds such a subgraph as a connected subgraph of GNN's computation graph. Here, the restriction to be "connected" is on computation graph rather than the original graph, which is intuitive. Even GNNs that capture information from far-away nodes through message passing (*e.g.,* Deep Graph Infomax, Jumping Knowledge Networks) have computation graphs that are different (*i.e.,* multi-hop neighborhoods) but are always connected even though corresponding nodes in the original graph are not necessarily connected. Further, unlike counterfactual reasoning, mutual information-based objective allows users to understand what graph structure is critical for a particular prediction in a way that gives a concise, semantically meaningful explanation. For instance, in Question-Answer graphs of Reddit threads (**Fig. S1**) we see that explanations have *"2-3 high degree nodes that simultaneously connect to many low degree nodes,"* which translates to the explanation of *"2-3 experts who all answer many different questions in a QA thread on Reddit."*

Synthetic datasets. R3 and R5 raise an important clarification point regarding the description of synthetic datasets. This is important as these datasets allow us to quantify the quality of explanations without necessitating manual curation, and thus they represent an advance over the prior art on explaining predictions, which often only provide hand-picked real-world examples. We acknowledge that the data generation procedure was not adequately explained in the draft and we will include more information on datasets/groundtruth in the main text with details in the Appendix. Further, we will open-source the code for data generation together with GNNEXPLAINER's code. Briefly, we first generate a base graph (*e.g.,* Barabási-Albert graph). For a random set of nodes, we then attach a particular structure (*e.g.,* a house- or grid-shaped motif) to each of the nodes. These nodes will have labels that are different others, which gives us a dataset in which absence/presence of the structure indicates a label. The groundtruth thus corresponds to these structures, and we calculate precision/recall of edges in an explanation relative to edges in the groundtruth.

New experiments. In response to the constructive feedback we conducted further experiments whose results confirm our findings and increase our confidence that GNNEXPLAINER is a general approach for explaining GNNs. Among others, for R5, we used GNNEXPLAINER to explain GNN models that use simple attention mechanisms and we observed similarly good performance as when explaining GCNs. Second, in addition to explaining node classification and graph classification, we also tested GNNEXPLAINER on link prediction (for R3). Experiments on a 2D-Grid dataset, where each grid graph has 20% of random edges removed, show that explanation for a predicted edge are typically edges forming a 4-cycle with the predicted edge, consistent with groundtruth. Third, R6 alludes to a hypothetical situation in which explanation is predictive of a wrong label with high probability. This is very interesting and related to adversarial attacks, however, we note that this situation was never realized in any of our experiments. That is because, at GNNEXPLAINER's initialization, the probability of GNN's predicted label is already the highest and thus it is unlikely that explainer would optimize the subgraph for false prediction. We will include all new analyses in Appendix.

Finally, reviewers raise minor points, such as a typo "K_M nodes" by R5 (it should be K_M edges) and a suggestion by R6 to investigate how explanations look if their size is constrained by the number of nodes instead of edges (our experiments show such constraints result in explanations being fully induced subgraphs). We will add formal discussions to the Appendix, including explicit discussion on future directions and adversarial attacks.