We thank the reviewers for their time and valuable feedback. Overall, we are glad that the reviewers found GIB as the first work that applies the IB principle to the GNN literature with clear motivation and novelty especially on compressing the structural information. Below, we address the points raised by the reviewers and resolve possible misunderstandings.

(1) Novelty with RGCN (R3) and comparison with prior methods (R1). R3 questions our novelty compared with RGCN. The reviewer argues that "while only AIB, which is to model attention between nodes, is novel" and also "Its connection to information bottleneck is just a theory analysis, not is essential." We kindly disagree. First, there is no previous work that applies the IB principle to GNNs and we are the **first** to develop such a theoretically sound framework (as is also recognized by R2 and R4). We would like to remind the reviewers that applying the IB principle to GNNs is by no means trivial since we need to model both the node feature X and the adjacency matrix A, and the two are not independent and thus cannot be modeled separately. Second, the AIB itself is already a significant contribution. As demonstrated in Nettack (Zügner et al.), structural attack is much more effective than feature attack, as is also demonstrated in our experiments that GIB which includes structural bottleneck outperforms XIB (Table 2). Thirdly, through theoretical analysis we show that in order to constrain the information flow such that we obtain an upper bound for $I(\mathcal{D}; Z_X^{(L)})$ (Eq. 3), both AIB and XIB are required (proposition 3.2). They together constitute the source of robustness, which is a concern by R2. Taken together, the application of IB principle to GNN for learning *minimal sufficient* representation, the principled theoretical framework that derives XIB and AIB as indispensable components, and the importance of AIB to improving the robustness of GNNs constitute a novel and significant contribution.

We thank R1's suggestion on including more technical and experimental comparison with information-related graph representation works.

Accuracy (%)	pert-1	pert-2	pert-3	pert-4
DGI-evasive (v.s. GIB)	54.5 (-5.5)	41.5 (-8.5)	35.5 (-6.5)	31.0 (-4.5)
DGI-poisoning (v.s. GIB)	53.5 (-6.0)	38.5 (-7.0)	33.0 (-6.5)	29.0 (-1.0)

Difference from deep graph infomax (DGI). This line of work has several differences. (1) **Task:** Their task mainly focuses on unsupervised representation learning while we target at improving the robustness of GNNs in supervised learning settings with IB. (2) **"Intuition":** The objective of DGI is to **maximize** mutual information between patch representations and high-level summaries of graphs, however, the objective of GIB is to **minimize** the mutual information between the **node feature/graph structure and their latent representation**, as well as **maximize** the mutual information between the **representation and the prediction**. (3) **Implementation.** Since DGI explicitly optimizes for the model to make wrong/opposite predictions on the corrupted graphs, (Eq. 1 in DGI), the robustness of the learned representation decreases and the learned model is very sensitive to perturbations and corruptions. We also conducted experiments where we trained DGI against adversarial attack on Cora, as shown in the table. DGI's accuracy after evasive and poisoning attacks are on average 5.7% lower than GIB, whose formulation naturally increases the robustness of the representation. **Difference from pretraining.** This line of work uses a large amount of data for pretraining while we focus on a novel formulation. Hence, the directions are completely orthogonal and pretraining can be combined if we have access to large amount of data. We will add the new results and the detailed discussions to the paper.

(2) Citeseer performance (R2, R4). R2 and R4 ask why GIB in the Citeseer poisoning attack setting is not so good. In fact, as discussed in lines 267-275, and in Appendix H (line 269), we analyzed why GIB underperforms GCNJaccard in Citeseer poisoning setting. As shown in Table 9 in Appendix H, most Nettack attacks end up adding edges between different classes, and Citeseer has much more nodes with very few degrees (1-3) than Cora and Pubmed. Note that this setting exactly matches the assumption of GCNJaccard, which behaves worse than GIB in all the other setups. We emphasize that GIB doesn't have *any* assumption on the attack model and is most general across many setups: another evidence is shown in Fig. 4 in Appendix H where ours outperforms GCNJaccard in deleting-edge attacks.

(3) Source of robustness (R2). R2 questions where the robustness of GIB comes from. Based on Proposition 3.2 we think that IB on both the node representation (XIB) as well as the graph structure (AIB) contribute to it. And the two mutually affect each other as shown in Figure 2 (a), where we draw the Markov chain and you can see that $Z_X^{(l)}$ is derived from $Z_X^{(l-1)}$ and $Z_A^{(l)}$, $Z_A^{(l)}$ is derived from $Z_X^{(l-1)}$ and A. Namely, through AIB we guarantee that the graph structure at each step contains the minimum necessary structure for the information flow, while through XIB we guarantee that the updated node representation also preserves the minimum sufficient information for downstream tasks. As listed in Table 2 in the experiment section, we also conduct extensive ablation studies to investigate the contribution of AIB and XIB. As R2 suggested, the XIB is exactly "the IB method which fixes graph structures".

(4) Applying GIB to other GNNs (R2, R4). R2 and R4 raise the question why GIB is only applied to GAT. A main contribution of our paper is to introduce a theoretically sound framework with inspiration from the IB principle and the first work to apply the IB to GNNs. We agree that the GIB framework can be further applied to many existing GNNs. In fact many GNN models can be viewed as a special case to the GIB. For example, the original GCN can be viewed as we do not optimize the KL term and use a Dirac delta for the posterior; and GraphSAGE can be viewed as we use the same uniform neighbor sampling scheme without any regularization. We also implemented GIB-GCN and it yields better results than GCN in both robustness (27% boost) and deep architectures (similar trend as GIB-GAT). Applying the GIB framework to other architectures is a promising future direction. For node embedding methods, it is out of the scope of this paper since we focus on applying IB to GNNs. We will revise the paper to include this discussion.