We would like to thank each of the reviewers for the constructive and insightful comments on our manuscript.

R1, R3: Comparison with more baselines. Table 1 shows more comparison results, where the suffixes "-I" and "-P", respectively, indicate that the identity matrix and the pretrained node2vec embeddings are used as the input features. We observe that the pretrained structural embeddings can indeed bring performance improvement. However, our MetaTNE still outperforms GCN-P and Meta-GNN-P by a significant margin. In addition, we see that Meta-GNN-P underperforms GCN-P and the reason is discussed in lines 311-315 in our paper. Due to limited space here, we will give the results of GCN-P and Meta-GNN-P on other datasets as well as with different $K_{*,+}$ and $K_{*,-}$, in the final version.

Table 1: Results with $K_{*,+} = 10$ and $K_{*,-} = 20$.

Method	BlogCatalog			PPI		
	AUC	F_1	Recall	AUC	F_1	Recall
GCN-I	0.6102	0.2730	0.2194	0.6544	0.3379	0.2721
GCN-P	0.6643	0.3892	0.3379	0.6596	0.4176	0.3729
Meta-GNN-I	0.4805	0.2375	0.2141	0.5466	0.3289	0.3081
Meta-GNN-P	0.6533	0.3567	0.2962	0.6537	0.3964	0.3373
MetaTNE	**0.6986**	**0.5380**	**0.6203**	**0.6865**	**0.5188**	**0.5621**

R1: (1) More ablation studies. Table 2 gives more ablation study results, where V1 denotes that the node embeddings are learned at the beginning and then left fixed and V2 denotes that each node is represented by a one-hot vector. Our method significantly outperforms V1 and V2 and the performance of these two variants is worse than that of the variants in Table 3 in our paper. In addition, V1 underperforms V2 even if the node embeddings of V1 are first learned from the graph structure. We speculate that the reason is that the latent space of node embeddings somewhat overfits to the metric of graph structure learning, making it harder to adapt to the metric of subsequent meta-learning or few-shot learning tasks. **(2) Explanation on Figure 3 in the supplement.** When there are more negative samples and the number of positive samples is fixed, the data becomes more skewed. A large degree of imbalance leads the classifier to bias towards the negative samples, which has two impacts: very few samples are predicted as positive samples, and the true positive samples are more difficult to identify. In general, the recall scores will drop significantly while the precision scores will not change too much. Consequently, both our method and the baseline show performance degradation in the F_1 scores when more negative samples are given and the number of positive samples keeps unchanged.

Table 2: Results of ablation study in terms of F_1.

Method	$K_{*,+} = 10, K_{*,-} = 20$		$K_{*,+} = 10, K_{*,-} = 40$	
	BlogCatalog	PPI	BlogCatalog	PPI
MetaTNE	**0.5380**	**0.5188**	**0.4398**	**0.4298**
V1	0.4748	0.4614	0.3549	0.3389
V2	0.4892	0.4819	0.3699	0.3777

R2: (1) About the meta-learning formulation. Due to limited space, we place the detailed meta-learning formulation of how to use the support and query sets in the supplement. We will clarify it in the final version and ensure that the paper is self-contained. Also, we will further polish our paper based on your suggestions to address other writing issues. **(2) Empirical justification for the optimization part.** We refer the reviewer to Table 3 in our paper where the results of V3 empirically justify the effectiveness of the optimization part. **(3) About tasks where node features exist.** Since we focus more on the featureless scenarios, MetaTNE currently cannot handle node features, and further research is needed to incorporate node features into the structural and meta-learning modules of our method.

R3: (1) Comparison against Meta-GNN. It is a standard paradigm, that both Meta-GNN and our MetaTNE follow, to conduct adaptation on the support set and then do evaluation on the query set in the meta-learning literature, however, it is non-trivial to effectively apply meta-learning to the considered *content-less graph data* under the *multi-label setting*. Our main technical contributions are the *specially designed transformation function and training scheduler*, which enable MetaTNE to achieve strong experimental results. In contrast, Meta-GNN simply uses MAML to train GCN models and does not show satisfactory performance in the scenario of interest even if using the node2vec embeddings as input as shown in Table 1. The reasons are discussed in lines 308-315 in our paper. **(2) Regarding the fine-tuning approaches.** As mentioned in Sec. 2.2 of the supplement, the baseline GCN is actually evaluated in a fine-tuning manner. Specifically, we first train a GCN model on the training data and then fine-tune the parameters of the last layer on the novel labels. During the rebuttal period, we further try fine-tuning all layers on the novel labels and find that the performance of fine-tuning all layers is slightly worse than that of only fine-tuning the last layer. For example, on BlogCatalog dataset with $K_{*,+} = 10$ and $K_{*,-} = 20$, the F_1 of the former is 0.3746 and the F_1 of the latter is 0.3892 (note that these numbers are obtained by using the node2vec embeddings as input). Complete results will be available in the final version and omitted here due to limited space. Overall, our proposed MetaTNE significantly outperforms the fine-tuning approaches. **(3)** We will resummarize the first contribution in lines 66-69 to make it more appropriate.

R3, R5: Explanation on why to use self-attention. For the embedding transformation, our goal is to find how a query node correlates with positive or negative support nodes. The self-attention has shown its power to effectively capture relationships between a set of elements in a wide range of applications and naturally meets our needs. We agree that it is insightful to explore different architectures to implement the transformation. We have already started working on this and will report our findings in the final version.

R5: By using 1-hop neighbors, our method already outperforms the baselines by a significant margin, and thus we did not try other ways to construct the neighbor set. In addition, we agree that it is more realistic to model label uncertainty. We leave these explorations as important future work.