# Stumbling Blocks in Graph Neural Networks: Unexpected Behaviors and Partial Remedies

Anonymous Submission
ICBINB Workshop @ ICLR 2025

**Abstract**

Despite recent progress in deploying graph neural networks (GNNs) to real-world tasks, we have encountered several pitfalls and inconclusive results. We highlight unexpected confusions in predictions, sensitivity to learning rate choices, and partial improvements that fail to generalize. These observations underscore the need for a more cautious perspective on GNN success stories in practical contexts.

## 1 Introduction

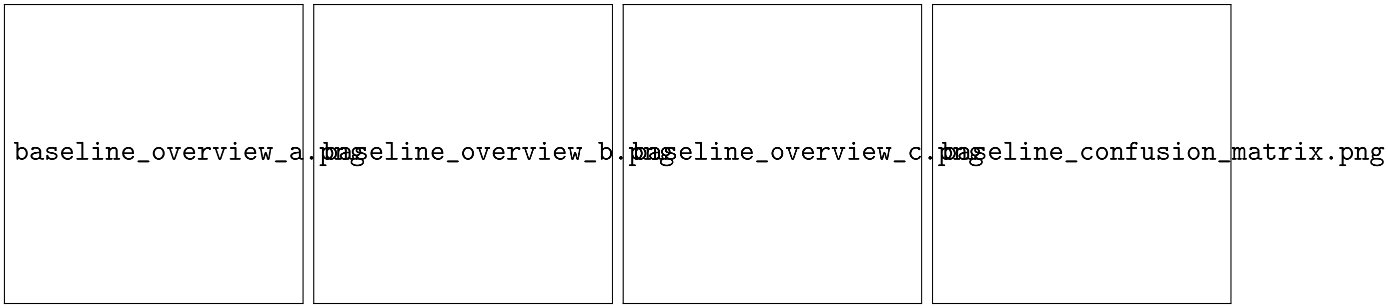
Graph neural networks (GNNs) have gained significant traction for diverse tasks such as node classification, link prediction, and molecular property identification [**??**]. However, their real-world deployment can be fraught with challenges, including brittle hyperparameter tuning and unexpected misclassifications. We investigate these issues through extensive experiments on a standard node classification benchmark, focusing on both Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs). Our contributions include a thorough analysis of pitfalls and partial remedies, aiming to inform practitioners about hidden caveats.

## 2 Related Work

Numerous studies have explored the stability and limitations of GNNs [**??**]. Some research highlights the vulnerability of GNNs to minor perturbations [**?**], while others focus on suboptimal generalization when domain shifts occur. Our examination extends this literature by emphasizing negative or inconclusive results, thereby providing valuable lessons regarding architecture choices and tuning practices.

## 3 Method and Pitfalls Observed

We trained a baseline GCN on a representative citation network under different learning rates. We tracked both training and validation performance, noting the frequent occurrence of overfitting or near-random performance when hyperparameters were not carefully tuned. We then introduced a GAT-based approach to examine whether richer attention mechanisms alleviated such pitfalls. Our findings exhibit partial improvements but also reveal persistent confusion among certain node classes.



(a) Train/Val Loss (b) Harmonic Mean (c) Validation HM (d) Confusion Matrix

Figure 1: Baseline GCN Results with varying hyperparameters.

# 4 Experiments

## 4.1 Baseline GCN Performance

Figure 1 presents an overview of our GCN results. Subfigures (a)–(c) show the training vs. validation loss, the harmonic mean of precision and recall, and how final validation scores vary across learning rates. In subfigure (d), we observe misclassifications concentrated on certain classes in the confusion matrix, illustrating how changes in hyperparameters fail to mitigate bias.

## 4.2 Advanced GAT Approach

We turned to GAT for potentially more refined representations. Figure 2 highlights similar pitfalls: while the learning curves improved marginally, the confusion matrix still revealed misclassifications, suggesting consistent struggles for certain node types.

# 5 Conclusion

We presented empirical evidence of unexpected behaviors and mixed results in GNN training. Our negative and inconclusive findings underscore how small hyperparameter shifts can trigger drastic fluctuations. Future work might investigate adaptive techniques or domain-specific calibration to mitigate these pitfalls and advance robust GNN deployment in real-world scenarios.

# A Additional Ablation Studies

Further ablation results, including learning rate sweeps and confusion matrices for each variant, are provided in Figures A1–A3 in the supplementary material. They confirm that baseline biases repeatedly emerge unless hyperparameters are carefully selected, often resulting in inconsistent performance gains.

# References

research_overview.png

Figure 2: GAT results exhibiting partial performance gains but persistent confusion among specific classes.