Evaluating Graph Neural Additive Networks for Multi-Label Node Classification

Arsenie Vlas | Supervisors: Elena Congeduti, Megha Khosla | Responsible Professor: Megha Khosla
EEMCS, Delft University of Technology | CSE3000 Research Project | a.vlas@student.tudelft.nl

1. Motivation

Multi-label node classification: each node carries several labels simultaneously (e.g. researcher spanning multiple fields; protein with multiple functions). Underexplored vs. single-label setting.

Standard GNNs are **black boxes**: they cannot explain *why* a node gets a label — a critical limitation in biology and medicine.

GNAN [2] is *interpretable by design*: explanations are a direct read-off of model parameters, not a post-hoc approximation. Yet GNAN has never been evaluated in the multi-label setting.

→ *This work fills that gap.*

2. What is GNAN?

Node i's k-th representation entry:

$$[h_i]_k = \sum_j \frac{1}{\#\mathrm{dist}(j,i)} \cdot \rho\left(\frac{1}{1+d(j,i)}\right) \cdot f_k([x_j]_k)$$

- $\rho(\cdot)$ — **distance function**: weight per hop; reveals local vs. global reliance
- $f_k(\cdot)$ — **shape functions**: one network per feature; readable as a curve

Multi-label adaptation: replace softmax with sigmoid + BCEWithLogitsLoss. Each label gets its own shape and distance behaviour. All other components unchanged.

3. Datasets

Two datasets from the MLGNC benchmark [1]:

Dataset	Nodes	Labels	Homophily
DBLP	28 702	4	0.76 (high)
PCG	3 233	15	0.17 (low)

DBLP: co-authorship network; TF-IDF word features. Primary for explanation analysis.
PCG: protein–cancer gene network. Low homophily challenges GNAN's additive structure.

4. Experimental Setup

Baselines [1]: MLP (no graph), GCN [3], GraphSAGE [4].
Metric: macro-AP — threshold-free, robust under label sparsity. AUROC excluded: unreliable when labels are sparse [1].
Protocol: 60/20/20 splits, 3 seeds, no hyperparameter tuning; early stopping on validation loss.
Hypothesis: GNAN competitive on DBLP (clean neighbourhood signals); underperforms on PCG (complex feature interactions needed).

5. Performance Results

Method	DBLP (AP)	PCG (AP)
MLP	0.350	0.148
GraphSAGE	0.868	0.185
GCN	0.893	0.210
GNAN (ours)	$\mathbf{0.850 \pm 0.002}$	$\mathbf{0.160 \pm 0.009}$

Table 1. Macro-AP, mean ± std over 3 splits. Baselines from [1].

PCG F1 ≈ 0: calibration artifact — sparse labels keep sigmoid outputs below 0.5. AP tells the real story: 10 of 15 PCG labels are learnable. Per-label AP correlates with prevalence (*Spearman* +0.98) and homophily (+0.94).

Main Finding

GNAN is **competitive with GCN** on high-homophily DBLP (macro-AP 0.850, within ∼4 pts of GCN's 0.893), but **degrades to the feature-only baseline** on low-homophily PCG (AP 0.160 ≈ MLP 0.148). Its built-in explanations show exactly why.

6. Shape Functions (DBLP)

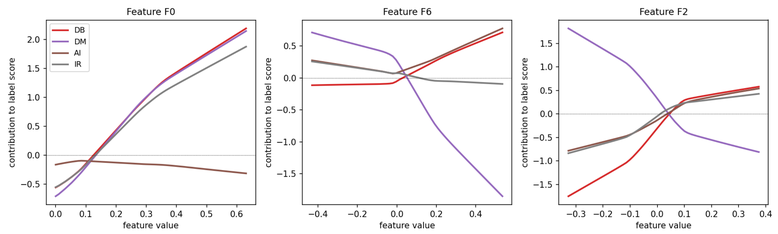

Figure 1. Three DBLP shape functions. x-axis = feature value; y-axis = additive contribution to label score. F0 raises DB/DM/IR but not AI; F6 separates DM; F2 sets DM apart. Influence in ∼26 of 300 features; top-10 identical across 3 seeds (Jaccard = 1.00). **DBLP**: stable; 90% of pairs in one direction. **PCG**: all 32 features, unstable (Jaccard = 0.37).

7. Distance Function

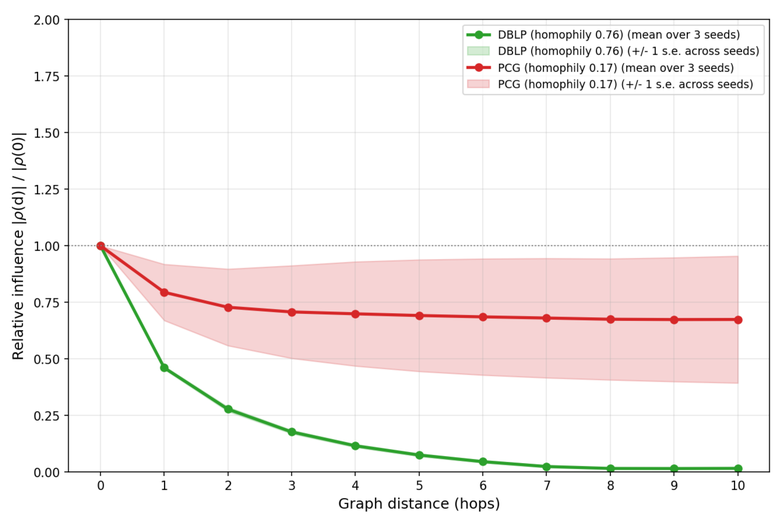

Figure 2. $|\rho(d)|/|\rho(0)|$ vs. hop distance. DBLP (green): steep decay, 78% within 2 hops, tight band (Pearson = 1.00). PCG (red): flat, wide band (Pearson = −0.26) — no consistent signal.

High homophily ⇒ steep, reproducible local decay. **Low homophily** ⇒ flat, unstable profile. → *The distance function adapts to data homophily.*

8. Feature × Distance Heatmap (DBLP)

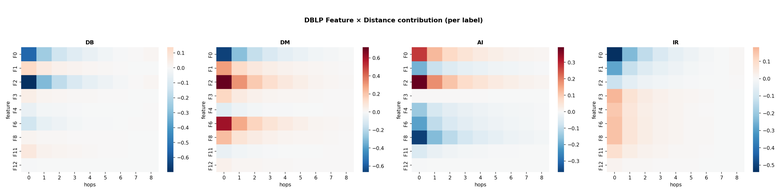

Figure 3. DBLP feature × distance contributions per label. Mass concentrated at low hops — signal comes from a few features in the immediate neighbourhood.

9. Local Node Importance (DBLP)

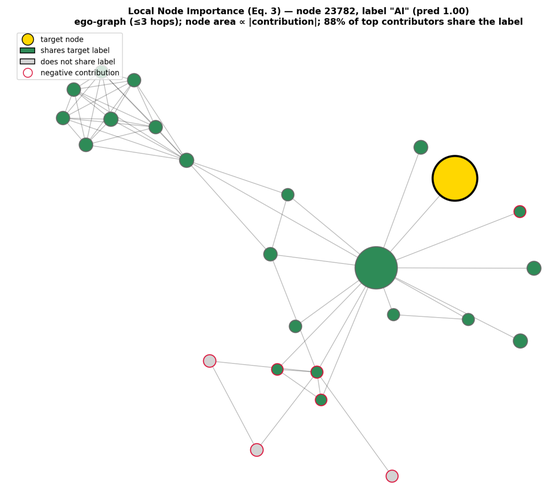

Figure 4. Ego-graph for label "AI". Node area ∝ contribution; 88% of top contributors share the label — confirms local neighbourhood reliance at the individual prediction level.

Node importance is derived directly from the model via Eq. (3) in [2] — no separate explanation method needed.

10. Conclusions

1. **Adaptation works.** Sigmoid output yields stable, non-trivial predictions across all seeds on both datasets.
2. **Performance follows homophily.** GNAN within ∼4 AP pts of GCN on DBLP; degrades to MLP level on PCG.
3. **Explanations reveal why.** Distance function concentrates locally on DBLP (reproducible) and flattens on PCG (unstable). Shape functions confirm: stable key features on DBLP; spread, noisy importance on PCG.

→ *Additive structure is an asset under high homophily and a liability under low — interpretability by design makes that trade-off transparent.*

11. Limitations & Future Work

Limitations:

- No TF-IDF vocabulary with DBLP; reading is structural not semantic.
- Two datasets; synthetic sweeps are future work.
- Fixed 0.5 threshold understates sparse-label performance.

Future work:

- Synthetic homophily & feature-quality sweeps (MLGNC generator).
- Compare explanations against GNNExplainer [5] / LIME.
- Per-label threshold calibration.

References

[1] T. Zhao et al. Multi-label node classification. *TMLR*, 2023. [2] M. Bechler-Speicher et al. The intelligible and effective GNAN. *NeurIPS*, 2024. [3] T. N. Kipf & M. Welling. GCNs. *ICLR*, 2017. [4] W. L. Hamilton et al. Inductive representation learning. *NeurIPS*, 2017. [5] Z. Ying et al. GNNExplainer. *NeurIPS*, 2019. [6] T. J. Hastie & R. J. Tibshirani. *Generalized Additive Models*. 1990. [7] R. Agarwal et al. Neural additive models. *NeurIPS*, 2021.