Heterophilic Methods on Multi-Label Graphs

How do methods designed for heterophilic graphs compare for multi-label node classification?

Cristian Turcan · Responsible Professor: Megha Khosla · Supervisor: Elena Congeduti · EEMCS, Delft University of Technology · CSE3000 Research Project

Background & question

- **Node classification:** predict each node's label(s) from its features and the surrounding graph.
- **GNNs** pass messages along edges, GCN smoothing is a *low-pass* filter - it helps when neighbours share labels (**homophily**, high edge ratio h).
- Under **heterophily** (low h) smoothing blends a node's signal with unrelated neighbours and *hurts*.

Risks the multi-label setting adds:

1. **Homophily is ill-defined** - Jaccard vs. per-label agreement disagree, and Jaccard overlap is intrinsically capped low.
2. **Correlations may absorb the signal** - co-occurring labels already encode structure that signed/high-pass filters target, shrinking the expected heterophily advantage.
3. **Labels are not mutually exclusive** - predicting one label does not rule out another, so multi-class intuitions break.

How do different methods designed for heterophilic graph datasets compare for multi-label node-classification datasets?

Definitions & metrics

Message passing (GCN-style aggregation):

$$h_i^{(l+1)} = \mathrm{UPD}\Big(h_i^{(l)},\ \mathrm{AGG}\big\{h_j^{(l)} : j \in \mathcal{N}(i)\big\}\Big)$$

Edge homophily (multi-class):

$$h = \frac{\big|\{(i,j) \in \mathcal{E} : y_i = y_j\}\big|}{|\mathcal{E}|}$$

Multi-label: replace equality with *Jaccard overlap* of label sets $\frac{|y_i \cap y_j|}{|y_i \cup y_j|}$ (intrinsically capped well below 1).

- **AP** - area under precision-recall curve, imbalance-aware. *Primary metric.*
- **F1-micro / F1-macro** - overall vs. per-class (rare-label) accuracy.
- **ROC-AUC** - ranking quality, averaged over labels.

Contribution & prior work

Our contribution is the *gap* between two recent works:

- **Zheng et al. (2024)** - survey organising the heterophily methods into three families *(used here)*, but evaluated almost only on **multi-class**.
- **Zhao et al. (2023)** - a **multi-label** benchmark, but testing only classical GNNs and a *single* heterophily model (H2GCN).

This work: 6 heterophily methods × multi-label - 5 real graphs, a multi-class control, and 2 controlled synthetic sweeps - read through one **feature** ↔ **structure** spectrum.

Heterophily methods: three architectural families (taxonomy of Zheng et al.)

Architecture refinement

change how messages are combined

Keep the neighbour set, change the filter:

- **FAGCN** - signed $\pm\alpha$ per edge: add similar, *subtract* dissimilar neighbours.
- **ACM-GCN** - low-pass, high-pass and identity channels, selected per node.
- **GPR-GNN** - learned signed per-hop weights (generalised PageRank).

Non-local extension

change who counts as a neighbour

Keep aggregation, change the neighbourhood (high-order mixing, potential-neighbour discovery):

- **Ordered GNN** - orders message passing by hop distance into separate, non-overlapping blocks of the hidden vector.

Hybrid

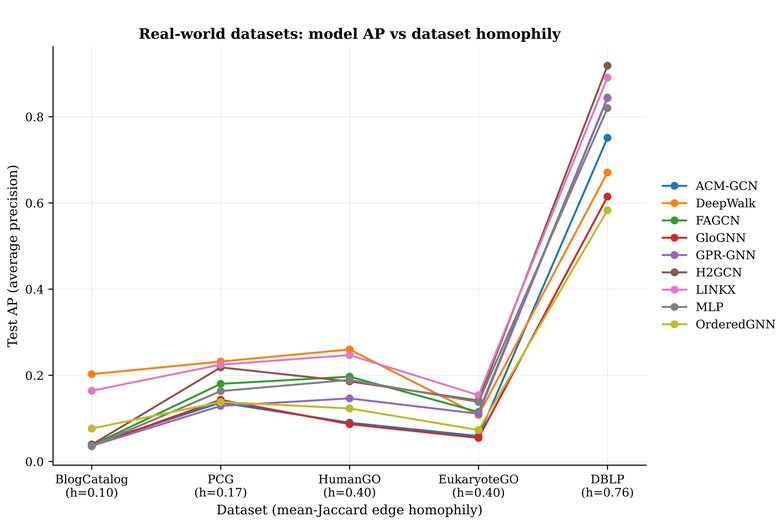
combine both

Mix neighbourhood and combination changes:

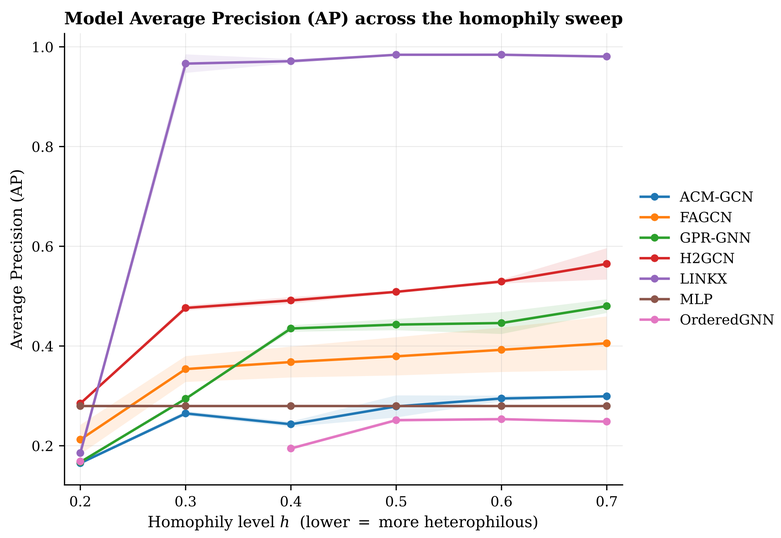
- **H2GCN** - ego/neighbour separation + higher-order hops + layer concatenation.
- **LINKX** - separate MLPs on node features and adjacency rows, combined only at the end.

Two baselines bracket the designs: **MLP** (features only) and **DeepWalk** (structure only). On the feature ↔ structure spectrum, the models that send features through *unguarded* neighbour aggregation (ACM-GCN, GPR-GNN) have the most to lose under low homophily.

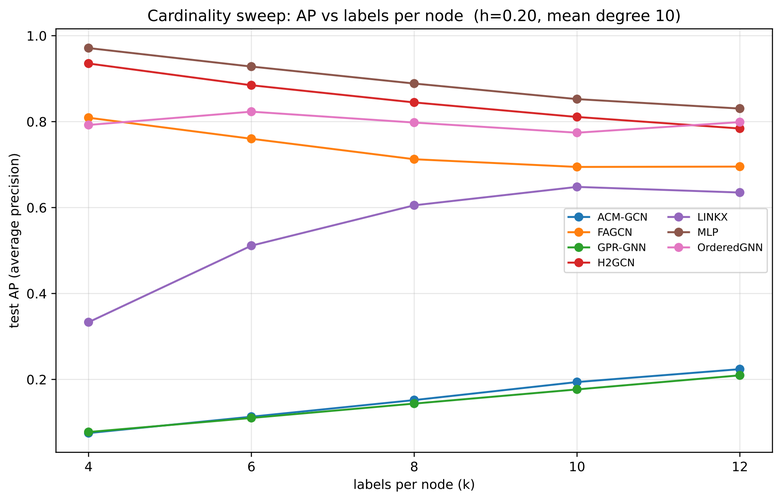
Key finding - performance tracks homophily, not model sophistication



Across real graphs. AP for the heterophily GNNs is low on heterophilous graphs and rises only on the most homophilous one (DBLP).



Controlled homophily sweep. Every GNN's AP rises with h, none gains an advantage as the graph grows more heterophilous.



Controlled cardinality sweep. The feature-only MLP stays on top across labels-per-node k, no heterophily GNN overtakes it.

Results: test average precision (AP)

Columns ordered left → right by increasing homophily h, **best per column** in bold.

Family	Model	Blog 0.10	PCG 0.17	Human 0.42	Euk 0.46	DBLP 0.76	Roman 0.05*
Baselines	MLP	0.036	0.163	0.189	0.138	0.820	0.545
	DeepWalk	**0.203**	**0.232**	**0.260**	0.108	0.671	0.058
Arch.	FAGCN	0.038	0.180	0.197	0.115	0.844	0.515
	ACM-GCN	0.040	0.137	0.090	0.059	0.751	0.474
	GPR-GNN	0.036	0.129	0.147	0.111	0.843	0.575
Non-local	Ordered GNN	0.077	0.138	0.123	0.073	0.583	0.125
Hybrid	H2GCN	0.039	0.218	0.186	0.142	**0.919**	**0.757**
	LINKX	0.164	0.225	0.247	**0.154**	0.891	0.455

Structure-only **DeepWalk** and late-fusion **LINKX** lead on the heterophilous graphs (left), message-passing **H2GCN** takes over once homophily is high (DBLP, and the Roman-Empire multi-class control*).

Datasets & pipeline

Dataset	Domain	Nodes	Lbl	h
BlogCatalog	Social	10 312	39	0.10
PCG	Protein	3 233	15	0.17
HumanGO	Gene Ont.	3 106	14	0.42
EukaryoteGO	Gene Ont.	7 766	22	0.46
DBLP	Co-author	28 000	4	0.76
Roman-Emp.	Word (MC)	22 662	18	0.05

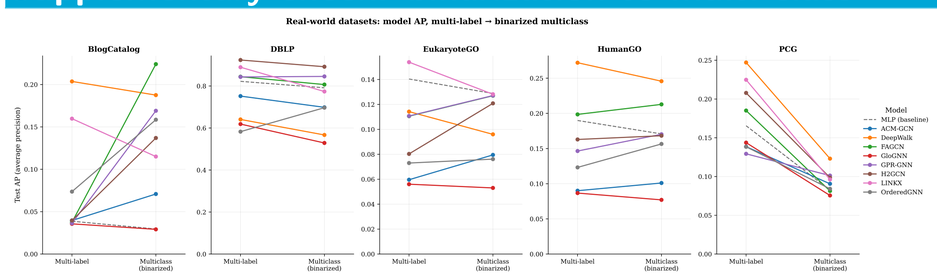
Five real multi-label graphs + Roman-Empire as a multi-class heterophily control.

- **One shared pipeline** per model (same load, preprocessing, splits, budget, metrics) - so differences are attributable to the model.

Synthetic sweeps - how the graphs are built. A multi-label generator (Zhao et al.) gives each node a fixed-size label set from L classes, then wires edges to hit a *target* Jaccard homophily, holding everything else fixed:

- **Homophily sweep:** vary $h = 0.2 \to 1.0$ ($N{=}3000$, $L{=}20$, $\approx$3.2 labels/node, 20-d features fixed). Targets h only, so mean degree co-varies ($\approx 1499 \to 32$) - a flagged confound.
- **Cardinality sweep:** fix $h = 0.20$ & mean degree 10, vary labels/node $k = 2\ldots12$ ($L{=}60$, 64-d features).
- Fresh 60/20/20 split, seeds $\{42, 43, 44\}$.

Supplementary check: binarization to multi-class



How we collapse. On the *same wiring* (A, X unchanged), relabel each node with its single *most-common* label - the label carried by the most nodes (ties → lowest index), unlabelled nodes are dropped. Only the targets y change. If multi-label were the only handicap every panel would slope *up* - it does not. The jump is concentrated on **BlogCatalog**, where the collapse incidentally *raises homophily*. Confounded by homophily and an easier target, so we treat it as a **diagnostic only** and keep it out of the main claims.

Findings & conclusion

- **Low homophily:** heterophily GNNs rarely beat MLP or DeepWalk, ACM-GCN the most frequent low point, LINKX often competitive.
- **High-h DBLP & the Roman-Empire control:** the message-passing advantage returns, with **H2GCN strongest**.
- **Both sweeps confirm it** under controlled conditions.

The advantages reported for heterophily methods on multi-class benchmarks may not transfer automatically to the low-homophily multi-label regime.

Future work - four threads

1. Deepen the homophily analysis. Compute per-class Cross-Class Neighbourhood Similarity (CCNS) under our pipeline and relate it to where each method helps or fails.

2. Strengthen the statistics. Extend multi-split runs to *all* datasets (EukaryoteGO and HumanGO are currently single-split) and report variance / confidence intervals.

3. Probe the co-occurrence hypothesis. Test directly whether label correlations explain the lost heterophily advantage, controlling for co-occurrence structure.

4. Broaden coverage. Complete the large Yelp benchmark (where several models, incl. H2GCN, currently fail) and add further heterophily models from the taxonomy.

Key references: Zheng et al., *GNNs for graphs with heterophily: a survey* (2024) · Zhao et al., *Multi-label node classification* (2023) · Platonov et al., *A critical look at heterophily benchmarks* (2023) · Zhu et al. H2GCN (2020) · Bo et al. FAGCN (2021) · Chien et al. GPR-GNN (2021) · Luan et al. ACM-GCN (2022) · Song et al. Ordered GNN (2023) · Lim et al. LINKX (2021) · Perozzi et al. DeepWalk (2014).