Per-Node *k*-Hop Label Homophily Predicts GNN Accuracy in Multi-Label Node Classification

The most predictive label-similarity scale is **two hops**, set by the **graph** and not the model. GCN and GAT, depth 1 to 4, observational plus causal.

Veaceslav Guzun
CSE3000 Research Project · TU Delft, EEMCS
Supervisors: Megha Khosla, Elena Congeduti

1 The setting and question

Many real graphs are **multi-label**: a protein joins several biological processes, a user carries several interest tags, a paper spans several topics. The standard predictors of these label sets are GNNs such as **GCN** and **GAT**, which classify each node by aggregating its neighbours.

Zhao et al. (TMLR 2023) built the MLGNC benchmark and proposed a clean **one-hop** label-homophily metric that tracks GNN performance. Yet a two-layer GNN already reads **two hops** wide, and deeper models reach further. The metric and the receptive field are mismatched.

> **Research question.** How does label similarity between a node and its neighbourhoods at increasing distance (one hop versus two, three, and beyond) affect GNN accuracy on multi-label node classification, and at which scale is that similarity most predictive of **per-node** accuracy?

- **SQ1** Does similarity beyond one hop predict per-node accuracy, and where is it strongest?
- **SQ2** Exact-distance shell or cumulative neighbourhood?
- **SQ3** Is the predictive scale set by the graph or by the model's depth?
- **SQ4** Does higher-order homophily cause the accuracy change, or merely co-vary?

2 Metric: per-node *k* -hop homophily

For each node v take its **exact-k-hop shell** $N_k(v)$, the nodes whose *shortest* path to v has length exactly k. The per-node score averages label-set similarity over that shell:

$$s_k(v) = (1 / |N_k(v)|) \sum_{u \in N_k(v)} \mathrm{Jaccard}(\ell(v), \ell(u))$$

mean label-set similarity at exactly k hops

Averaging over nodes, weighted by shell size, gives the dataset-level number, exactly the mean Jaccard over all distance-k pairs:

$$H_k = (1 / |P_k|) \sum_{(v,u) \in P_k} \mathrm{Jaccard}(\ell(v), \ell(u))$$

$P_k = \{ (v, u) : d(v, u) = k \}$

Two generalisations of Zhao et al. at once: per node, so $s_k(v)$ pairs with one node's own accuracy; and per distance, so it reads any scale. At $k = 1$ it collapses exactly to their Definition 1 (verified: $H_1 = 0.755$ on DBLP versus their 0.76).

3 Method

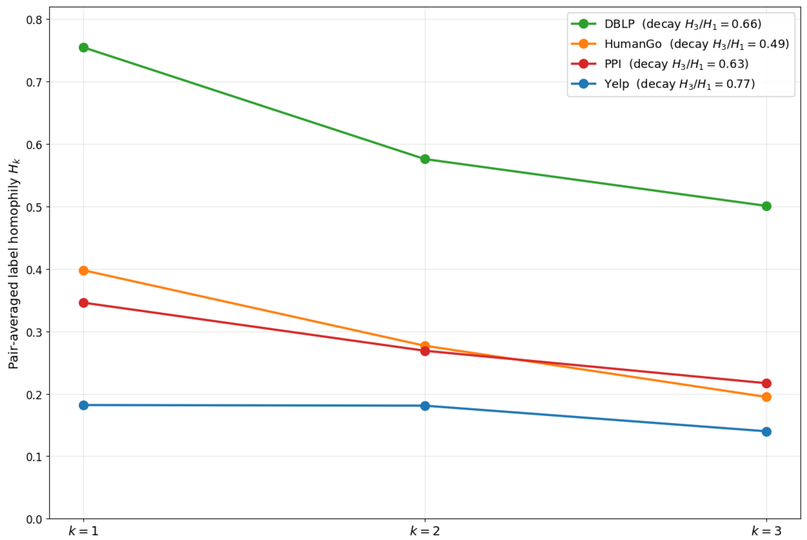
Datasets. Three MLGNC graphs, analysed per node: **HumanGo** (1029 test nodes), **DBLP** (5411), **PPI** (5298, pooled). Plus two synthetic generators for causal evidence (§7).

Models. GCN and GAT at $L \in \{1, 2, 3, 4\}$, Zhao et al.'s exact hyperparameters, sigmoid output, BCE loss, Adam, early stopping. Every (dataset, model, depth) trained under **3 seeds** (42, 123, 456).

Per-node analysis. For each test node pair its shell homophily $s_k(v)$ with its **per-node Average Precision** (threshold-free, the primary outcome), then take the within-dataset **Spearman r**.

Per node, not per dataset: only three datasets overlap the published scores, so a per-dataset correlation is structurally blind. Letting every test node be a point turns 3 points into thousands and lets the correlation vary with k. **Spearman** because the claim is monotonic, AP saturates near 1, and $s_k(v)$ is right-skewed.

4 Dataset profiles: H_k decays with k

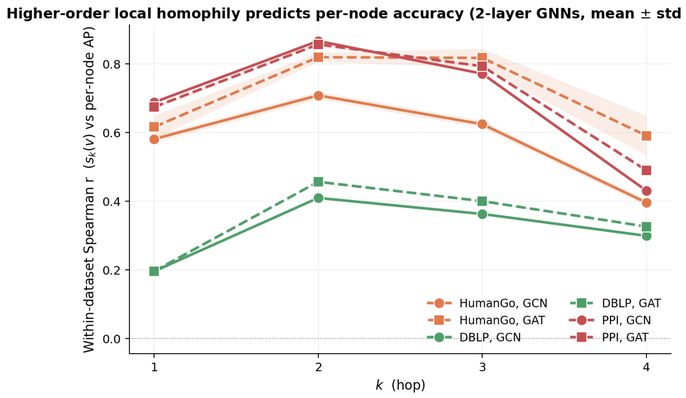


All four datasets show monotone decay of H_k. **Level and shape are independent features:** DBLP starts highest yet decays at a middling rate ($H_3/H_1 = 0.66$), HumanGo decays fastest (0.49), Yelp slowest (0.77).

Dataset	H_1	H_2	H_3	Ref. H_1
Yelp	0.182	0.181	0.140	0.22
PPI	0.346	0.269	0.217	n/a
DBLP	0.755	0.576	0.501	0.76
HumanGo	0.398	0.277	0.195	n/a

Our H_1 reproduces Zhao et al.'s reference to two decimals (Yelp differs only by an explicit self-loop policy), anchoring the pipeline.

5 SQ1 The predictive scale is two hops



Within-dataset Spearman r between per-node $s_k(v)$ and per-node AP, two-layer GNNs, mean over 3 seeds (bands one std, mostly thinner than the line). Every curve is strongest at $k = 2$ or 3, **never** at $k = 1$.

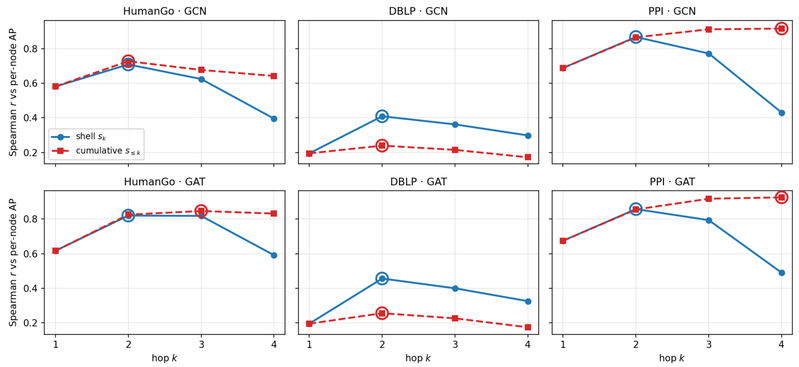
Data	Model	k=1	k=2	k=3	k=4
HumanGo	GCN	+0.58	**+0.71**	+0.62	+0.40
HumanGo	GAT	+0.62	**+0.82**	**+0.82**	+0.59
DBLP	GCN	+0.20	**+0.41**	+0.36	+0.30
DBLP	GAT	+0.20	**+0.46**	+0.40	+0.33
PPI	GCN	+0.69	**+0.87**	+0.77	+0.43
PPI	GAT	+0.67	**+0.86**	+0.79	+0.49

Highlighted cell is the strongest scale. HumanGo GAT ties at $k = 2$ and 3 (within the seed spread). The gain over $k = 1$ is +0.13 to +0.26; the spread there is at most ±0.02.

Finding 1 **Higher-order wins.** In all 6 (dataset, model) cells, some $k > 1$ beats $k = 1$, and $k = 1$ is never the best. Five of six are strongest at $k = 2$, the two-layer receptive field; the sixth is tied between $k = 2$ and 3. Two hops balances **coverage** (the one-hop shell is tiny and noisy) against **relevance decay** (far shells regress to the graph mean).

6 SQ2 Read it on the shell, not the ball

Does the unit of measurement matter? We recompute every correlation on the **cumulative ball** $s_{\le k}(v)$ (all neighbours within k hops) and compare against the **exact shell** $s_k(v)$, reusing the same trained models.

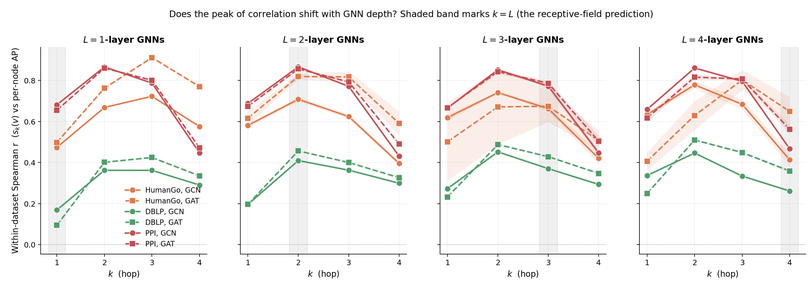


Shell s_k (solid) versus cumulative $s_{\le k}$ (dashed) as a predictor of per-node AP; circles mark where each curve is strongest. On **PPI** the cumulative keeps climbing to a misleading high at $k = 4$ (+0.91) while the shell drops back (+0.43); on **DBLP** the cumulative sits well below the shell at every k; on **HumanGo** the two track closely.

Finding 2 **The shell localises the scale; the ball does not.** A practitioner reading the cumulative curve would place PPI's scale at four hops and DBLP's near one, whereas the per-shell view places both at **two**. Widening the ball re-mixes a size-weighted blend rather than adding a fresh measurement, so where it looks strongest follows each graph's shell-size distribution, not its signal.

7 SQ3 The best scale does not track depth

Is two hops the predictive scale because the graphs make it so, or only because we use two-layer models? If the latter, that scale should move outward as we add layers. We sweep $L \in \{1, 2, 3, 4\}$, extending homophily to $k = 4$: **24** (dataset, model, depth) cells.

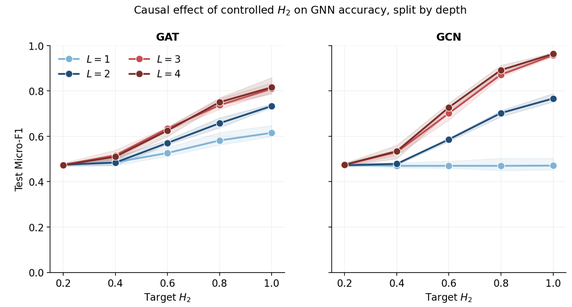


Spearman r versus k, one panel per depth L. The grey band marks $k = L$, where the receptive-field story predicts the strongest scale. Curves are strongest at **$k = 2$ in every panel** instead. Depth only sharpens or flattens the sides, and destabilises deep GAT on the small HumanGo graph (oversmoothing).

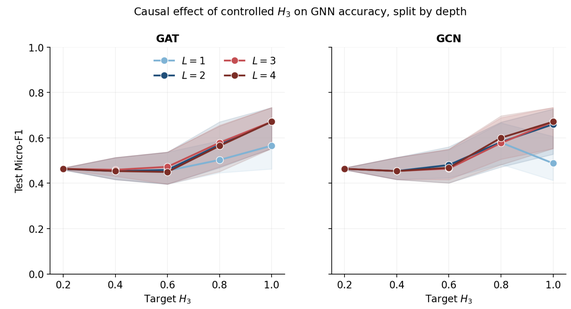
Finding 3 **Graph-driven, not architecture-driven.** Not one of the 24 cells is strongest at $k = L$ when $L \in \{1, 4\}$; the $k = 4$ column is never the highest, and is the lowest in every PPI cell. The two-hop scale is a property of the **data**. Report H_k as a **profile over distance**, not a single edge number.

8 SQ4 Causal evidence under intervention

Correlation cannot rule out a feature confound. Two generators plant label structure at **exactly one** target hop while holding the rest at baseline; features are **Gaussian noise**, so the only signal is topology. Five H_k levels × 3 seeds × 4 depths each.



Hub-spoke, controlling H_2. Raising H_2 from 0.2 to 1.0 lifts Micro-F1 from 0.47 (chance) to 0.96 for GCN at $L \ge 3$. The **L=1** GCN stays **exactly flat**: one hop cannot reach a two-hop signal.



Hexagon-cycle, controlling H_3. A smaller but real effect (≈ +0.20 Micro-F1). Depth separation is softer: a 6-cycle is already weakly position-identifying, a weaker isolation than hub-spoke.

Finding 4 **Higher-order homophily causes accuracy.** The response is graded across five controlled levels (dose-response), fixed in direction by construction, and gated by depth as the mechanism predicts. The one-layer GAT rises rather than staying flat, a shared-hub **shortcut** that attention exploits, not two-hop reach.

9 Conclusions

- **Two hops is the predictive scale** SQ1 , beating one hop in all 6 cells across two architectures and thousands of nodes.
- **Read on the exact shell** SQ2 ; the cumulative ball mislocates the scale by an amount the graph sets.
- **The scale is graph-driven** SQ3 ; it survives depth 1 to 4 rather than tracking the receptive field.
- **The relationship is causal** SQ4 ; controlled sweeps recover the gain under random features.
- **Next:** a shell-weighted GNN that targets the most predictive shell, turning a diagnostic into a design knob.

Contact & reproducibility

- **Veaceslav Guzun** · CSE3000 Research Project, June 2026
- TU Delft, EEMCS · Supervisors: Megha Khosla, Elena Congeduti
- Email: veaceslavguzun@tudelft.nl

A four-script pipeline reproduces every result from fixed seeds; the shell and cumulative readings differ by one `--cumulative` flag and share the same trained models. Code, splits, and per-node CSVs released.