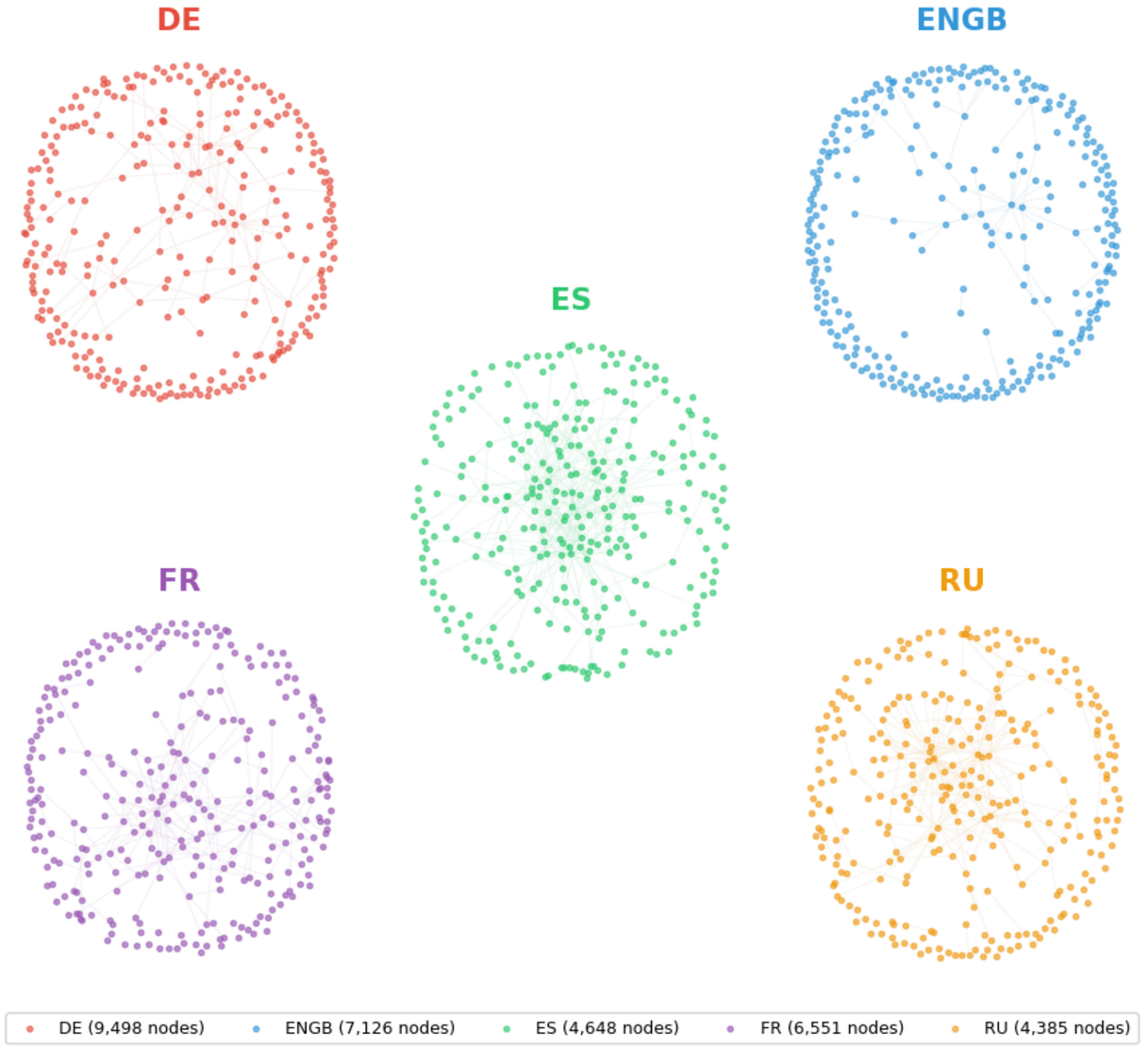
Shallow to Deep: Exploring the power of Graph Embeddings Across Two Domains

Hadiqa Alamdar Bukhari
Marwah Sulaiman

FEB 26

Dataset 1: Twitch Gamers

- 5 completely disconnected language graphs:
  - German - DE
  - English - ENGB
  - Spanish - ES
  - French - FR
  - Russian - RU
- # nodes = 32,208 nodes
- # edges = 397,814 edges
- No cross-language edges exist

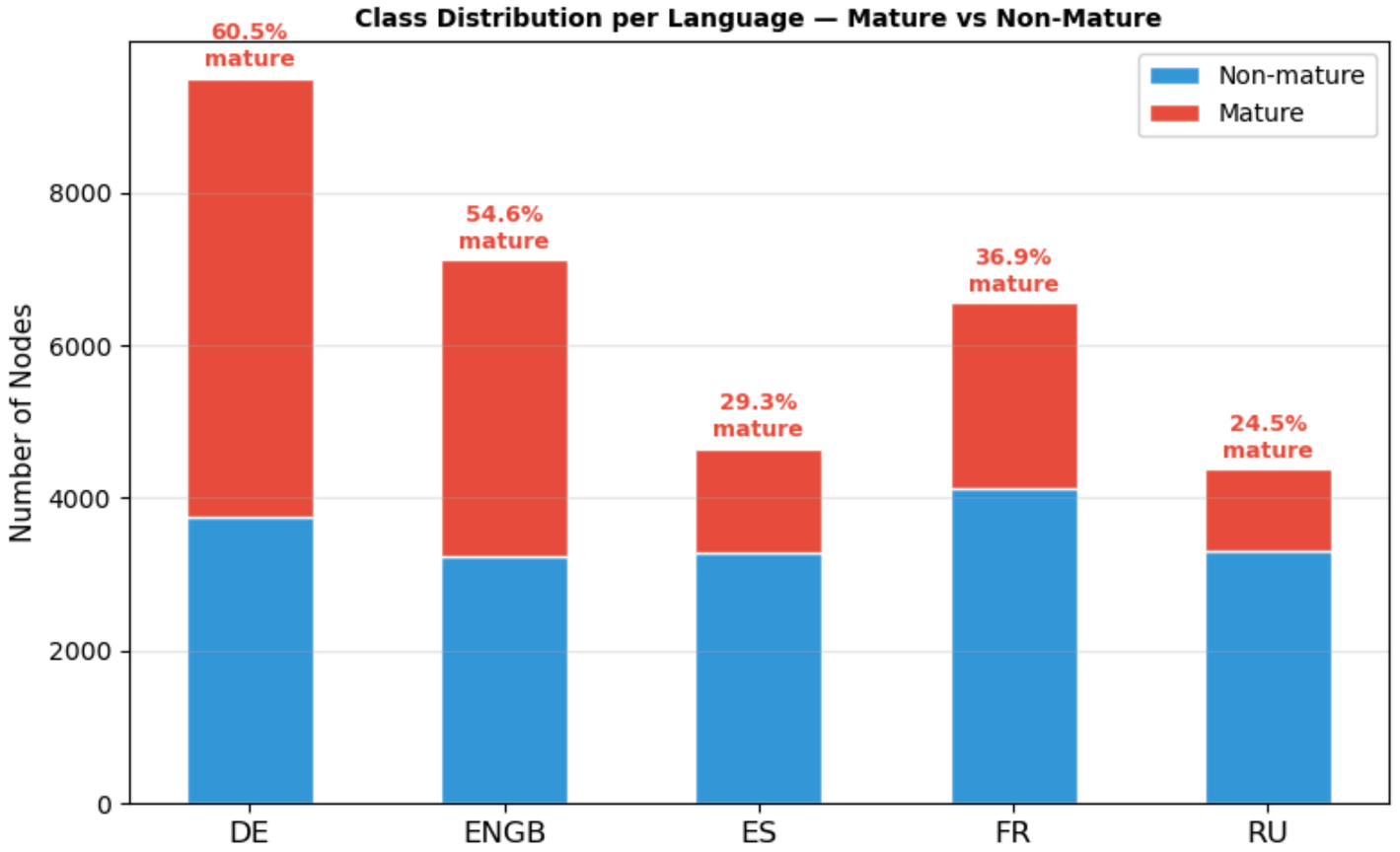


The Bridge from Part 1

1. Classical methods could only analyse each language in isolation
2. Every streamer is described by the same 3,170-dimensional game vocabulary.
3. Can a model trained on German streamers predict behaviour in Russian ones?

Transfer Learning

- Task: predict mature label — does this streamer use explicit language?
- Features: 3,170-dim multi-hot game vector (same vocabulary across all languages)
- Class imbalance: RU 24.5% mature, ES 29.3% — motivates AUC over accuracy
- Train on one source language
- Predict on all other languages



Overall Pipeline

INPUT

- 3,170-dim multi-hot game vector
- 1 = game played
- 0 = game not played

GCN NODE CLASSIFIER

- 2-layer message passing
- In domain AUC
- Transfer learning

T-SNE

- why does transfer work?

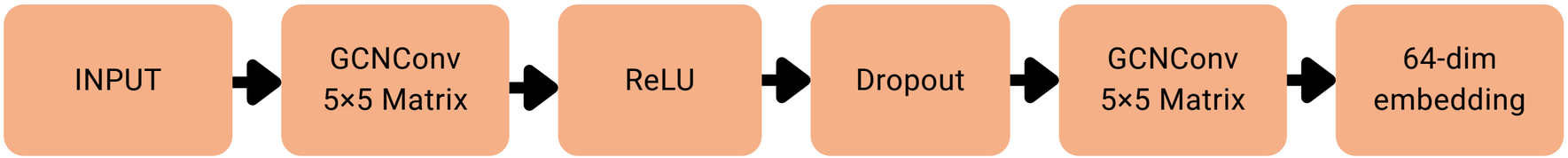
GAT NODE CLASSIFIER

- Learned attention weights over neighbours
- In domain AUC
- Transfer learning

LR NODE CLASSIFIER

- Used as a baseline
- Raw features only
- No graph data

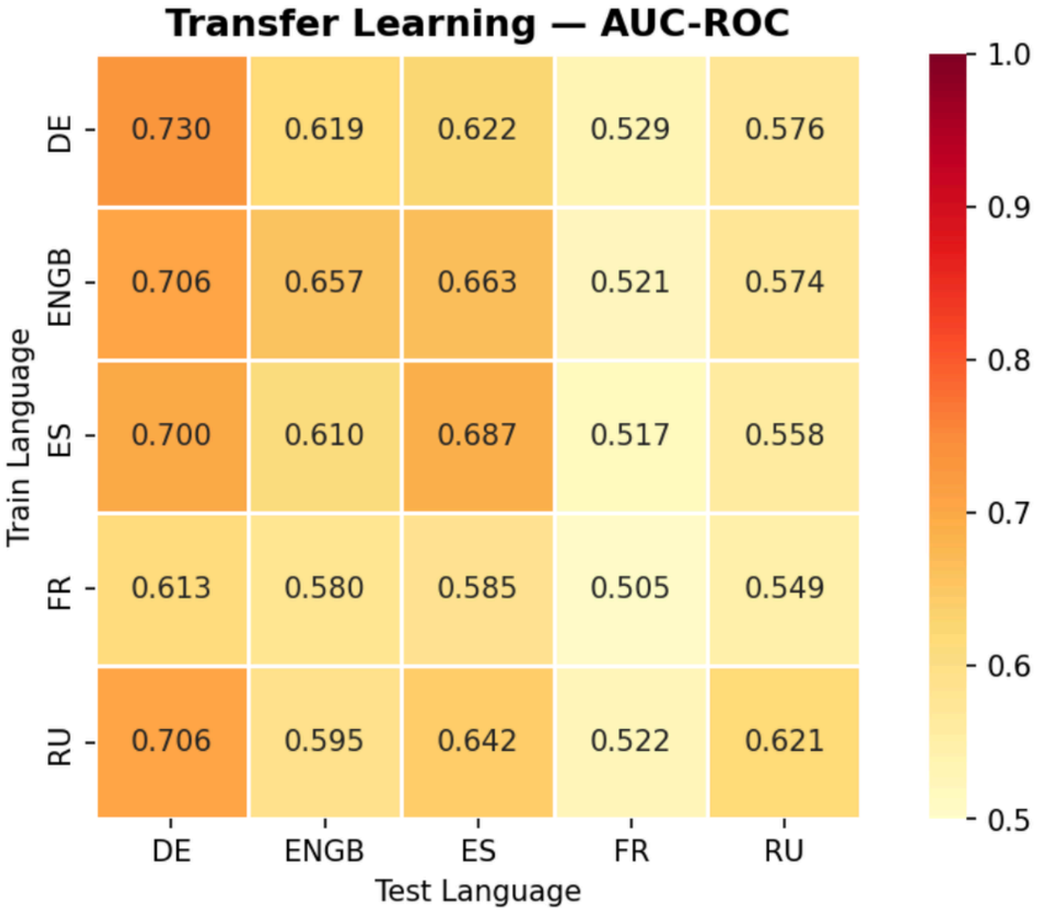
GCN Architecture and Method



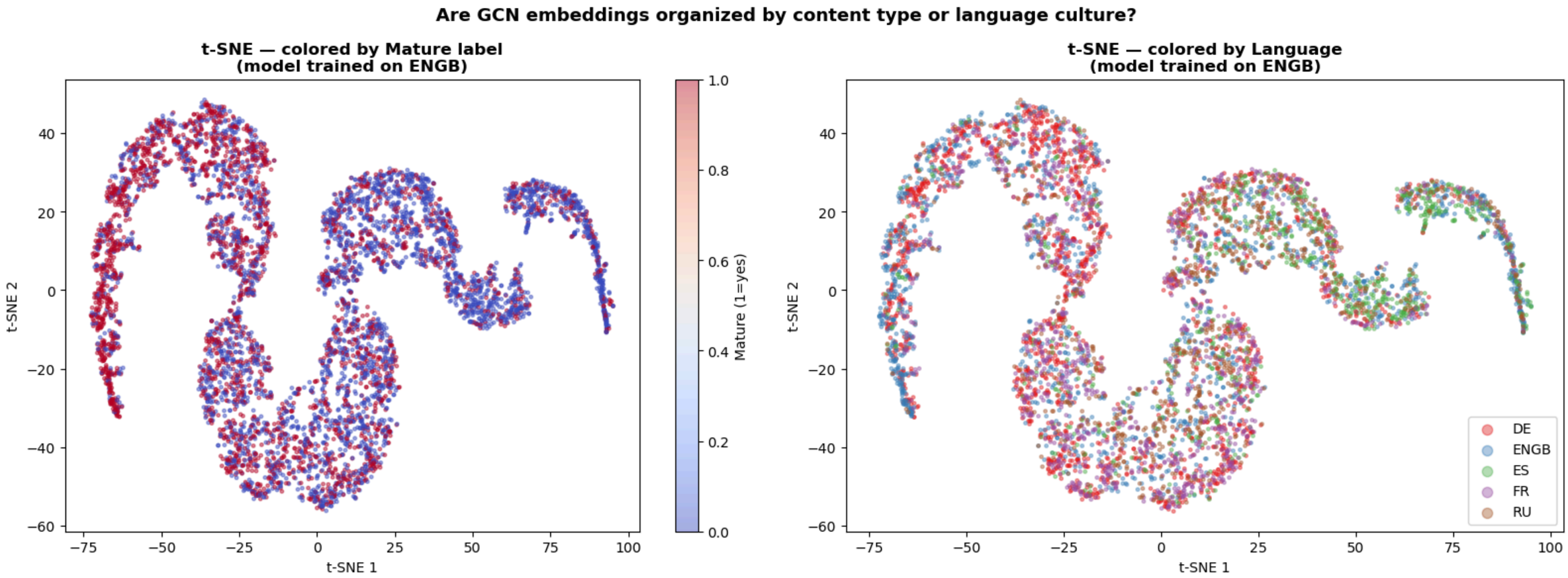
- For each source language:
  - Train a NodeClassifier on a 70/15/15 train/val/test split
  - Evaluate on its own test set (in-domain performance — matrix diagonal)
  - Apply the frozen, trained model directly to every other language (zero-shot transfer — off-diagonal)

GCN Results

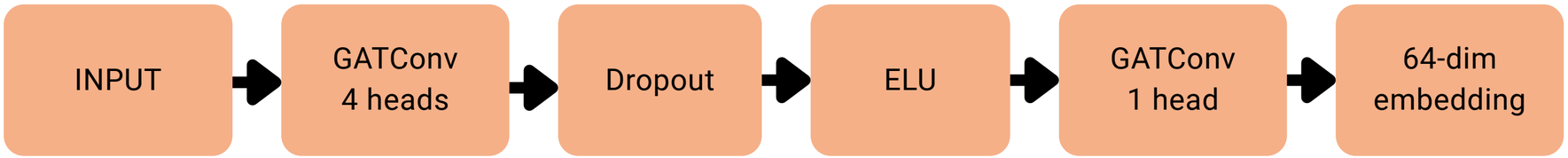
- **Best source**: ENGB and RU (avg transfer AUC 0.616 each) — not DE despite being the largest graph. Size alone does not determine transfer quality.
- **Easiest target**: DE (avg AUC 0.681 received from all sources) — near-balanced classes (60.5% mature) make it easy to predict regardless of source.
- **Hardest target**: FR (avg 0.515, barely above random).
- **Transfer beats in-domain**: ENGB→DE (0.706) exceeds ENGB in-domain (0.657). DE is such a rich, balanced target that even an imperfect model does well on it.



t-SNE — The Embedding Space Explains the Transfer

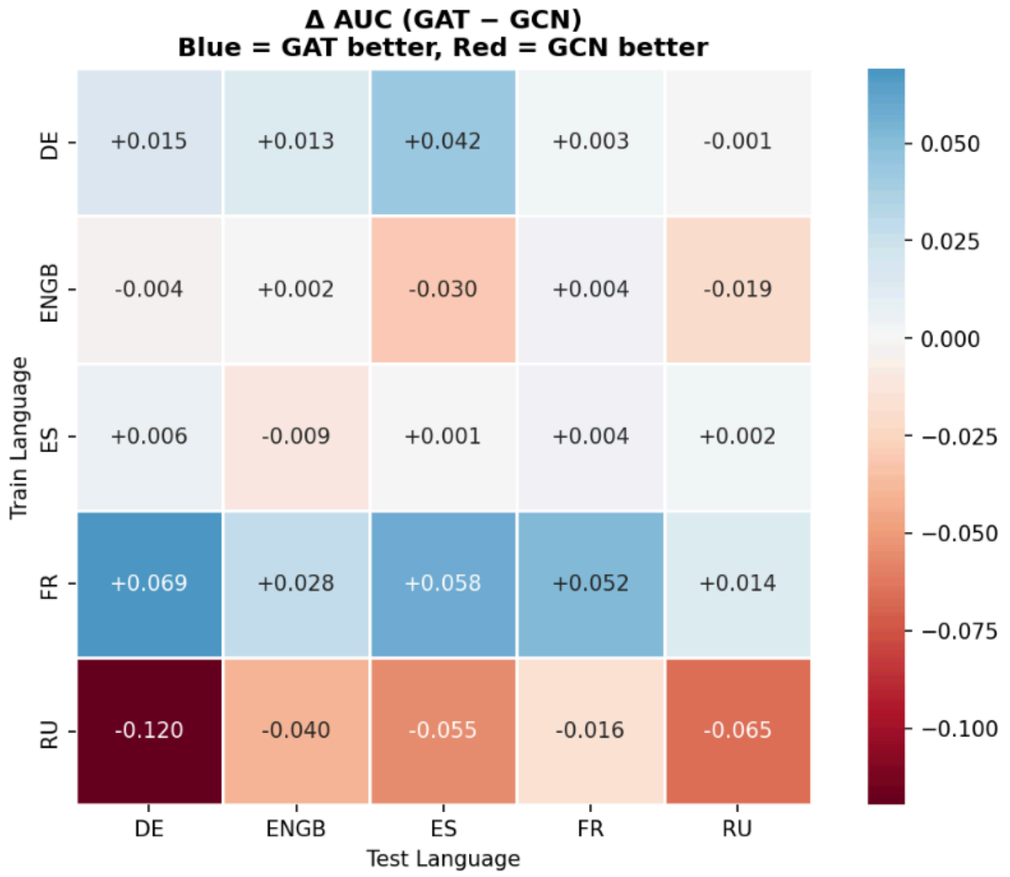


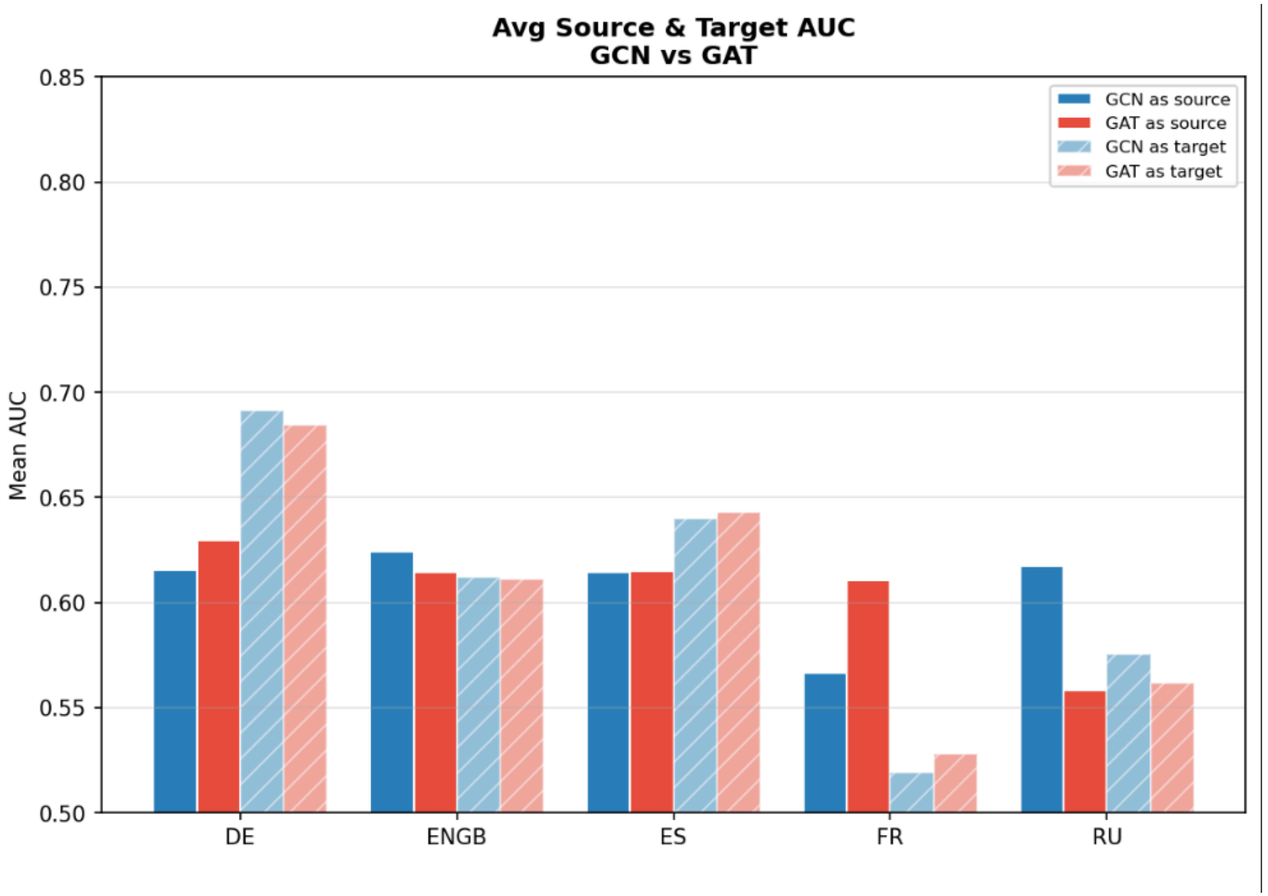
GAT Architecture and Method



- For each source language:
  - Train a NodeClassifier on a 70/15/15 train/val/test split
  - Evaluate on its own test set (in-domain performance — matrix diagonal)
  - Apply the frozen, trained model directly to every other language (zero-shot transfer — off-diagonal)

GAT vs GCN


Δ AUC (GAT − GCN)
Blue = GAT better, Red = GCN better
Train Language
DE
ENGB
ES
FR
RU
+0.015 +0.013 +0.042 +0.003 -0.001
-0.004 +0.002 -0.030 +0.004 -0.019
+0.006 -0.009 +0.001 +0.004 +0.002
+0.069 +0.028 +0.058 +0.052 +0.014
-0.120 -0.040 -0.055 -0.016 -0.065
DE ENGB ES FR RU
Test Language
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100


Avg Source & Target AUC
GCN vs GAT
GCN as source
GAT as source
GCN as target
GAT as target
Mean AUC
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
DE ENGB ES FR RU

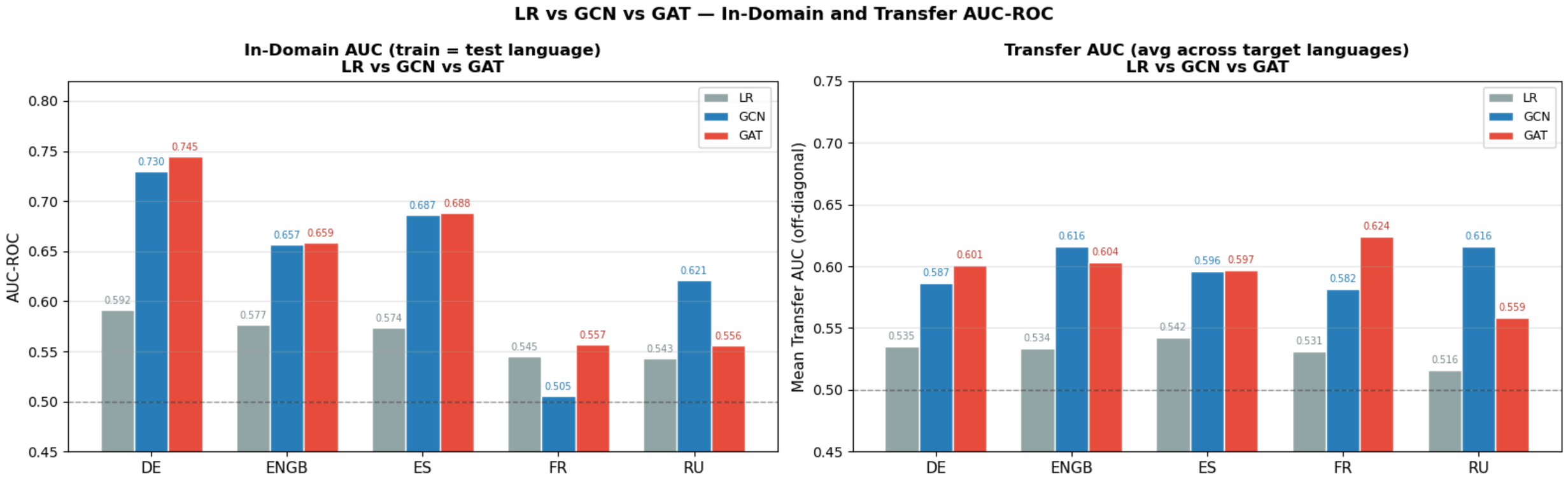
GAT vs GCN Analysis

- GAT overfits on RU - the smallest, most imbalanced graph and its attention weights don't transfer.
- FR: GAT finds a signal that equal-weight aggregation misses, and that signal transfers.
- Too little data for attention to work reliably in RU
- Too much noise for equal-weight aggregation to work in FR

GCN vs GAT vs Linear Regression

Property	LR	GCN	GAT
Uses graph edges	No	Yes	Yes
Sees neighbours	No	2-hop	2-hop weighted
Parameters	3170	420K	430K
Neighbour weighting	None	Degree normalised	Learned attention
Transfer mechanism	Feature similarity only	Feature + graph structure	Feature + selective

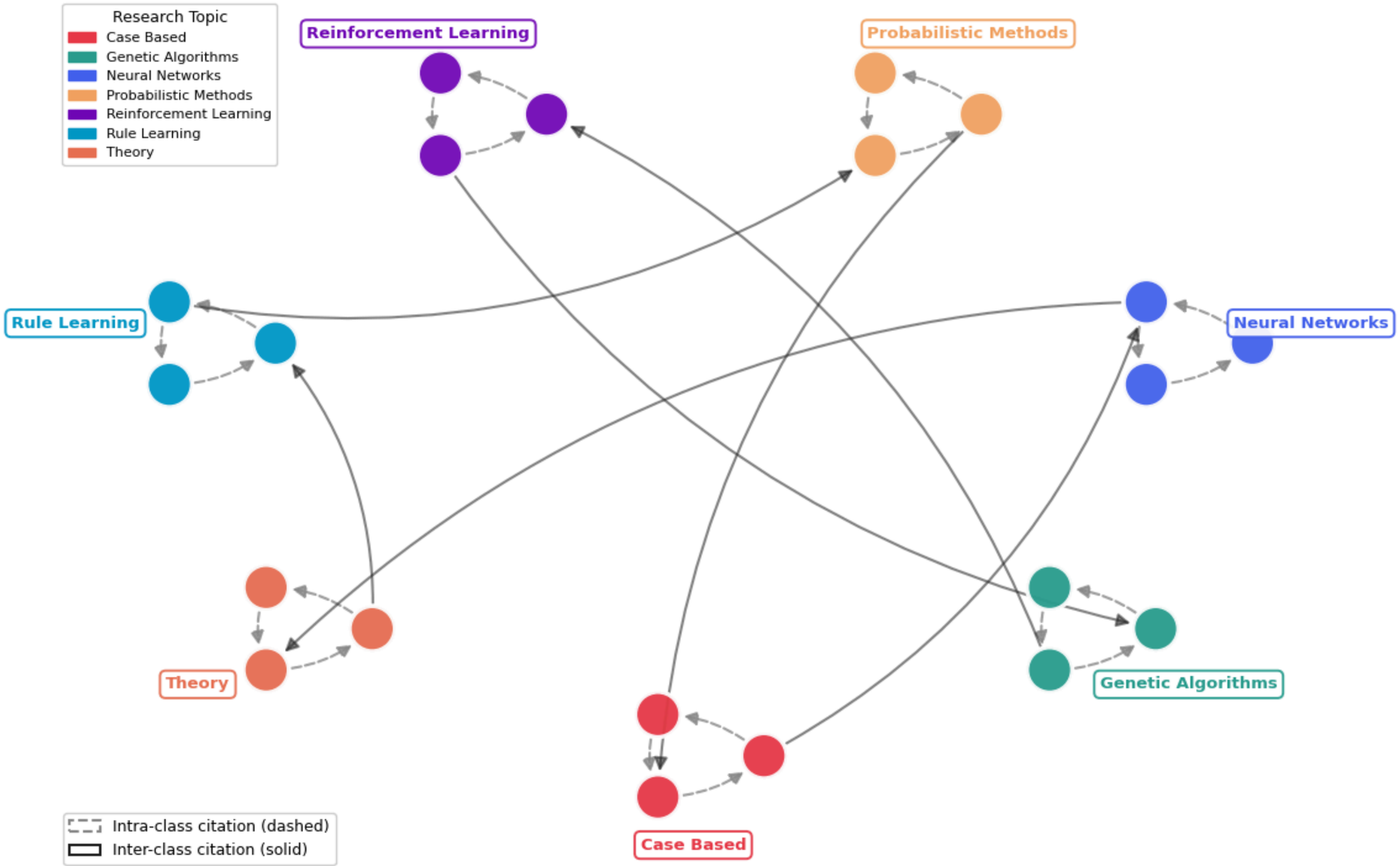
GCN vs GAT vs Linear Regression



Conclusion

1. **Graph structure enables transfer** LR features alone transfer near-randomly. Message passing over friendship networks builds representations that generalise across languages.
2. **GCN and GAT have complementary strengths** GAT outperforms GCN for dense, noisy graphs (FR, DE) by filtering uninformative neighbours. GCN outperforms GAT for small, imbalanced graphs (RU) where attention overfits. Neither dominates universally.
3. **FR is an irreducible challenge** No architecture resolves FR. Content type and social structure are independent in the French community. This is a property of the data, not a limitation of the methods.

Dataset 2: Cora Citation Network



Each node = one paper · Node features: 1,433-dim Bag-of-Words vector

Connection to Part 1: From Analysis to Learning

1. **Structure exists → can we exploit it?** Part 1 revealed strong community structure (modularity 0.82), papers cluster by topic, meaning the graph is highly homophilic. Part 2 exploits this: GNNs learn by aggregating neighbors, so homophily makes the citation graph a natural fit for message passing.

2. I**f citations encode similarity, can we predict missing ones?** Part 1 showed communities align with research topics (purity 0.78), meaning citations encode similarity. Link prediction asks the reverse: can learned embeddings predict which citations are missing?

Overall Pipeline

DATA PREPARATION

- Graph construction (directed + undirected)
- Startified data splitting
- Homophily & Related Insights

NODE CLASSIFICATION

- Shallow & GNNs Models Comprison
- Depth analysis
- Failure analysis

LINK PREDICTION

- Preparation: edge splitting & negative sampling
- Shallow vs. Deep Models

ANALYSIS & CONCLUSIONS

- Structure + features beats either alone
- GNNs dominate node classification
- GAT outperforms convolution for link prediction

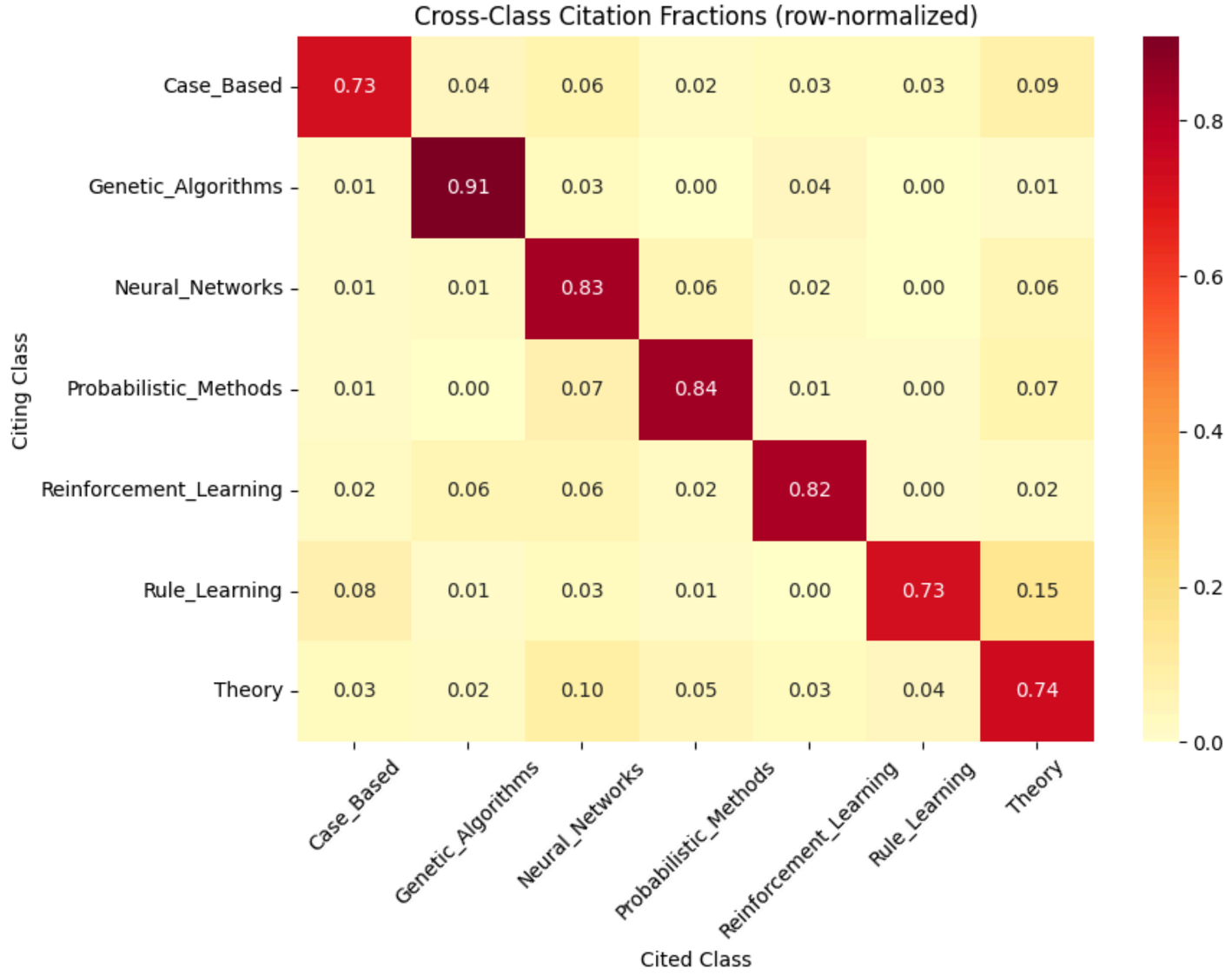
Graph Structure

Nodes: 2,708
Edges (Directed): 5,429
Classes: 7
Avg Degree: 4

Homophily: fraction of edges where both nodes share the same class label.

0.81

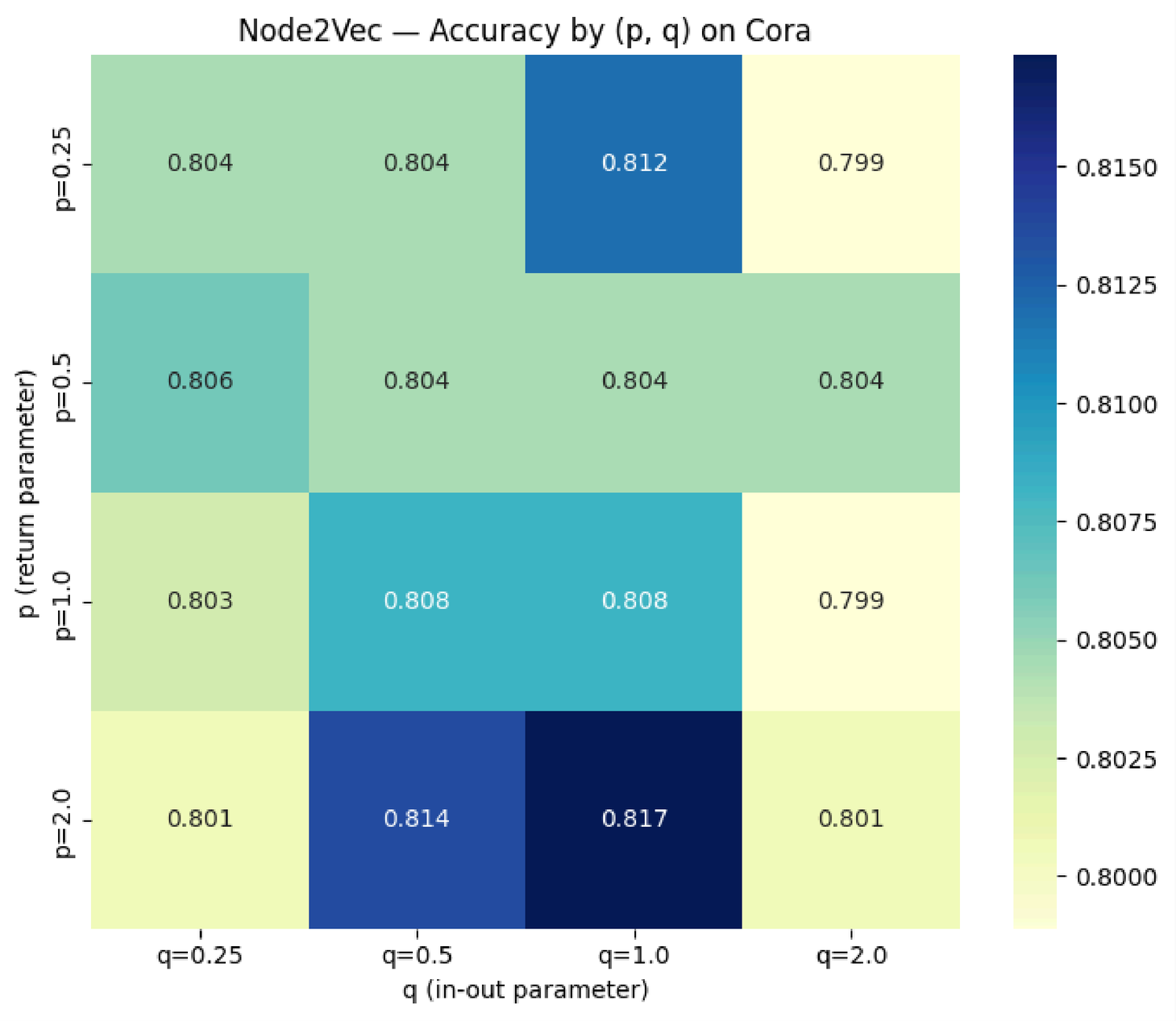
High homophily → GNNs should strongly benefit from message passing



Node Classification

With millions of papers published yearly, manual topic labeling doesn't scale. Automatically tagging papers by research topic enables indexing, search, and recommendation at scale.

Model	What it uses	Why / Motivation
BoW + Logistic Regression	Features only (no graph)	Baseline — how much do paper keywords alone predict topic?
Laplacian Eigenmaps + LR	Structure only (no features)	Spectral baseline — can graph topology alone recover topics? Connects to Part 1 community detection
Node2Vec + LR	Structure only (random walks)	Shallow embedding baseline — do random-walk neighborhoods capture class membership?
GCN	Structure + features	First deep model — does combining both improve over either alone? Standard message-passing benchmark
GraphSAGE	Structure + features	GCN variant with **inductive** neighbour sampling — generalises to unseen nodes, handles large graphs
GAT	Structure + features	Not all citations carry equal weight. Can the use of **attention** improve performance?
R-GCN	Directed structure + features	Treats cites and cited_by as separate relation types — does citation **direction** carry extra signal?



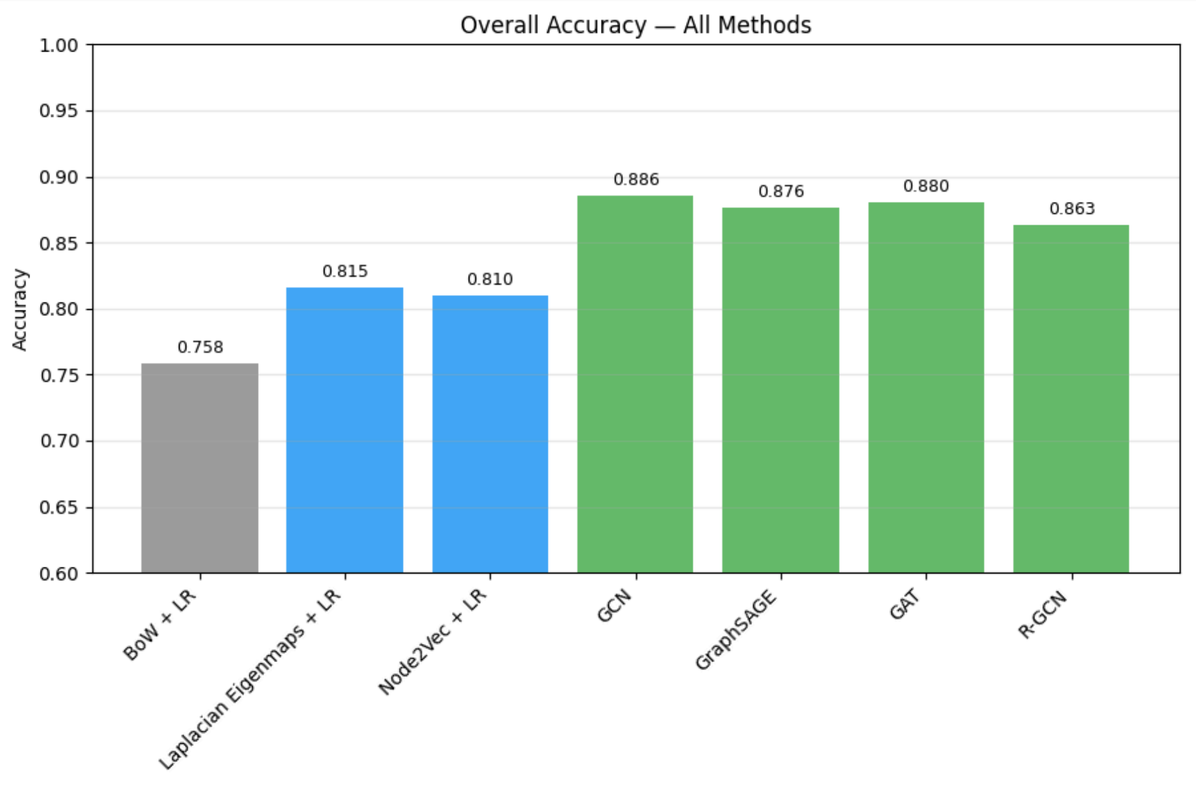
Node2Vec p & q study

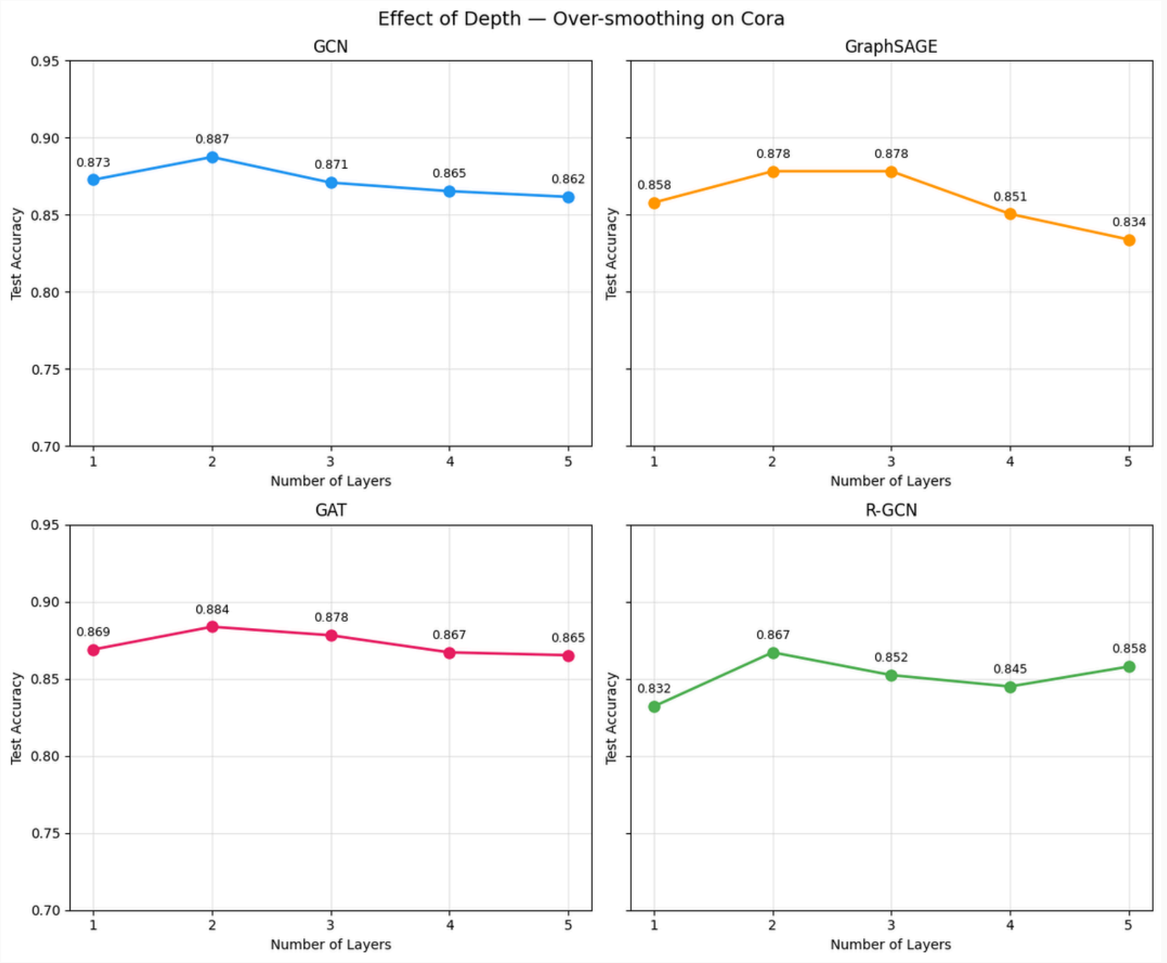
4×4 grid over p, q ∈ {0.25, 0.5, 1.0, 2.0}

p & q trade local vs. global walk exploration. Given Cora's high homophily, BFS-biased walks (low p, high q) should yield more class-coherent embeddings by staying within same-topic citation neighbourhoods.

Result: p & q matter less than expected: Grid search range: 0.799–0.817 accuracy — Cora doesn't strongly reward BFS vs. DFS walks, likely because its homophily (≈0.81) is high enough that most walk strategies stay within class clusters.

Node Classification - Results





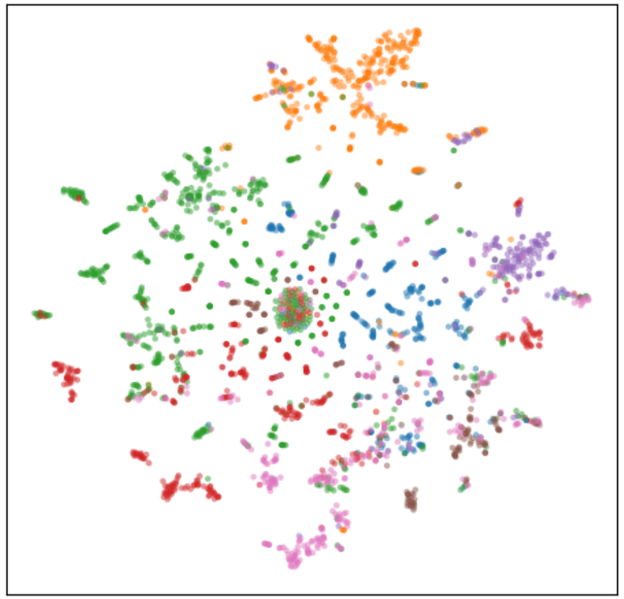
Structure + Features > Structure only > Features only

- GraphSAGE — Near-identical accuracy to GCN, but inductive (better)
- GAT — 3–5× longer to train, near-identical accuracy (88.0%)
- R-GCN — Direction adds nothing: citing and cited papers are already in the same domain. More parameters, same accuracy, complexity without gain.

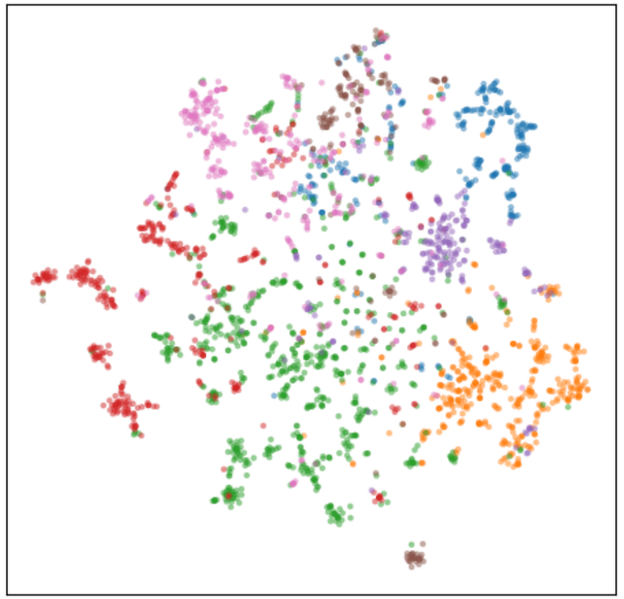
All GNNs peak at depth 2.

Deeper layers cause over-smoothing.

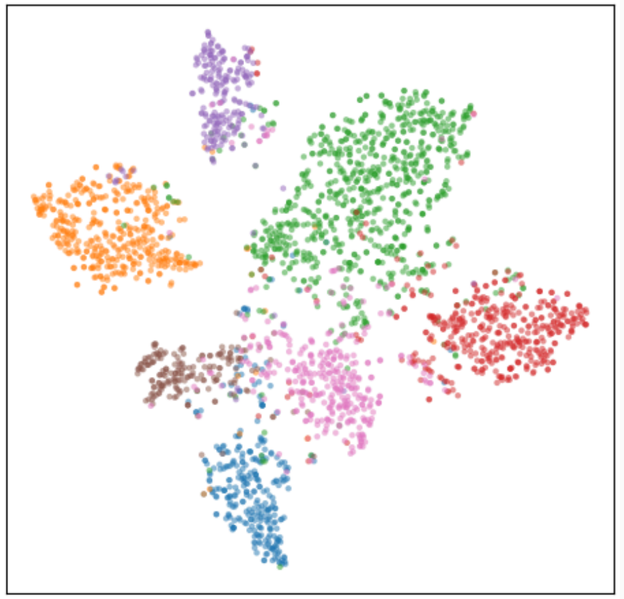
Laplacian Eigenmaps + LR
(Acc=0.815)



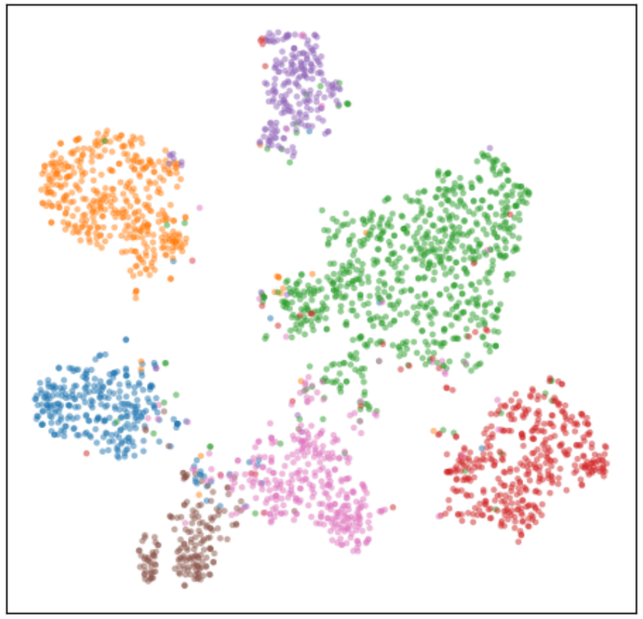
Node2Vec + LR
(Acc=0.810)



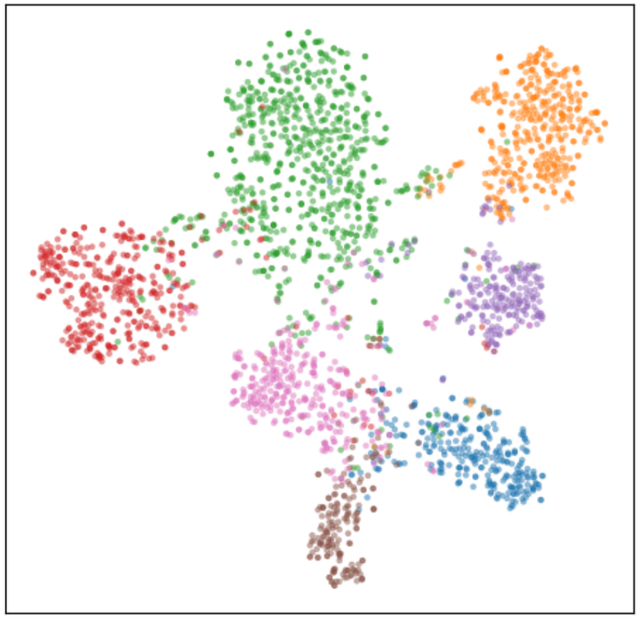
GCN
(Acc=0.886)



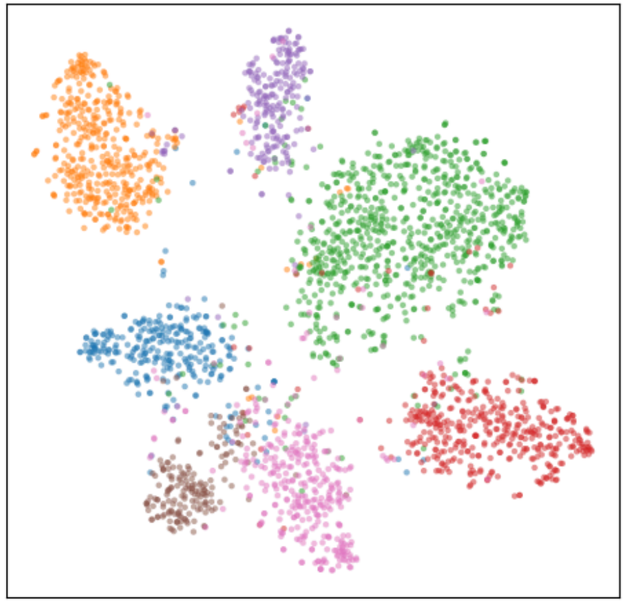
GraphSAGE
(Acc=0.876)



GAT
(Acc=0.880)



R-GCN
(Acc=0.863)

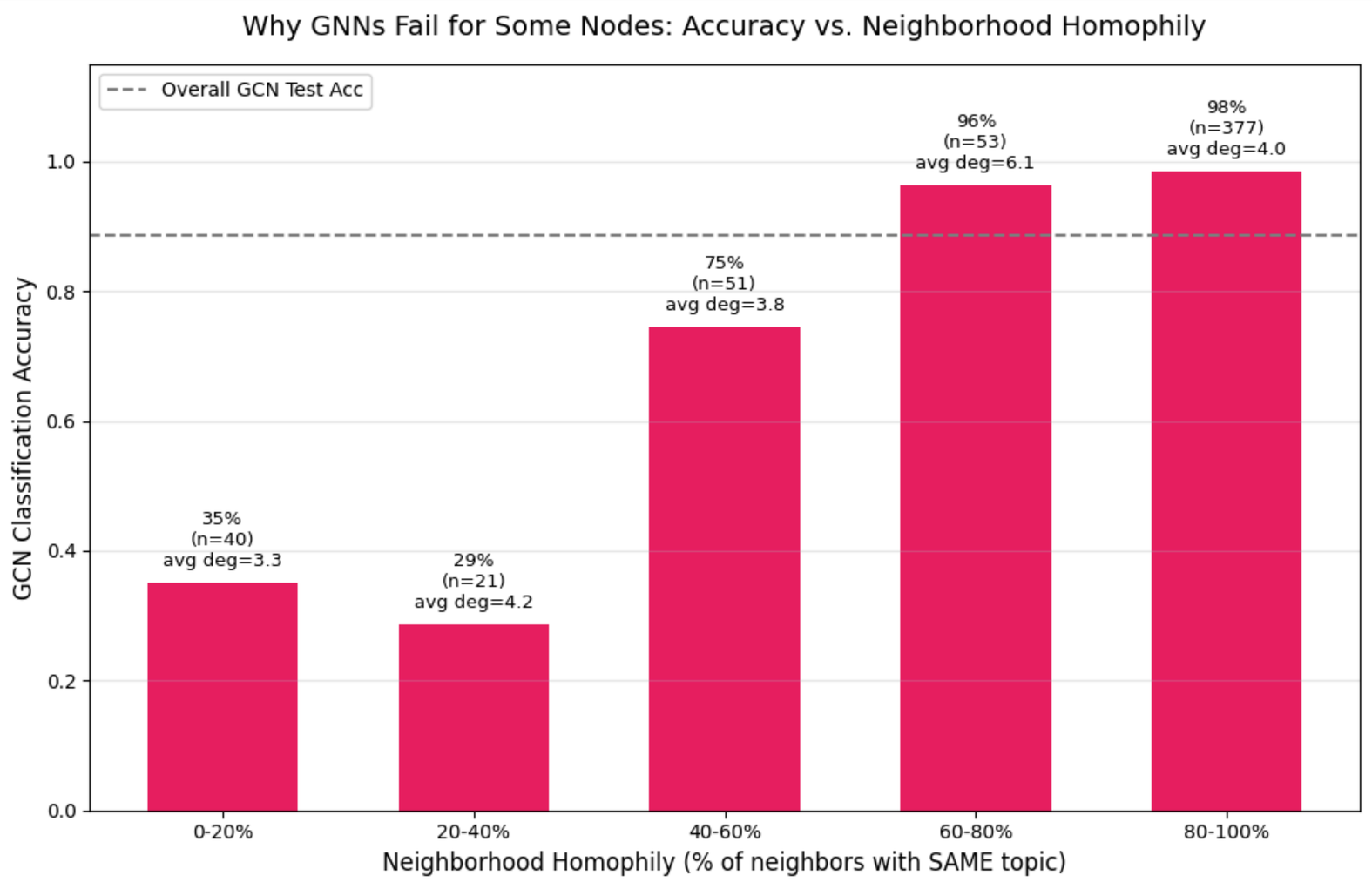


Embedding Quality Analysis

Using euclidean distance in original embedding space

Method	Intra-class dist.	Inter-class dist.	Ratio ↑
BoW + LR	5.79	5.84	1.01
Laplacian Eigenmaps + LR	13.25	13.48	1.02
Node2Vec + LR	3.18	3.43	1.08
GCN	3.32	6.09	1.83
GraphSAGE	4.19	7.87	1.88
GAT	**1.81**	**3.74**	**2.06**
R-GCN	4.4	6.81	1.55

Further Analysis



Accuracy is fundamentally bounded by neighborhood homophily

nodes in high-homophily neighborhoods reach 96–98%

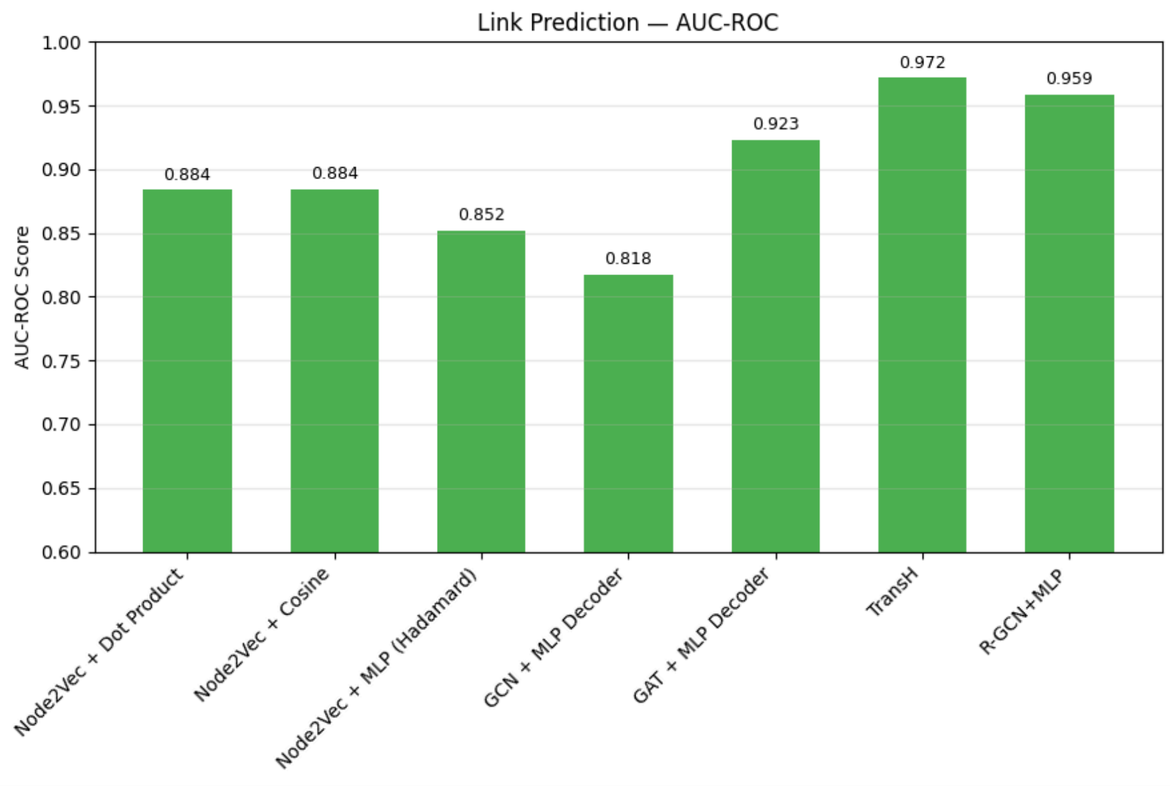
bridge nodes between communities drop to 29–35% error.

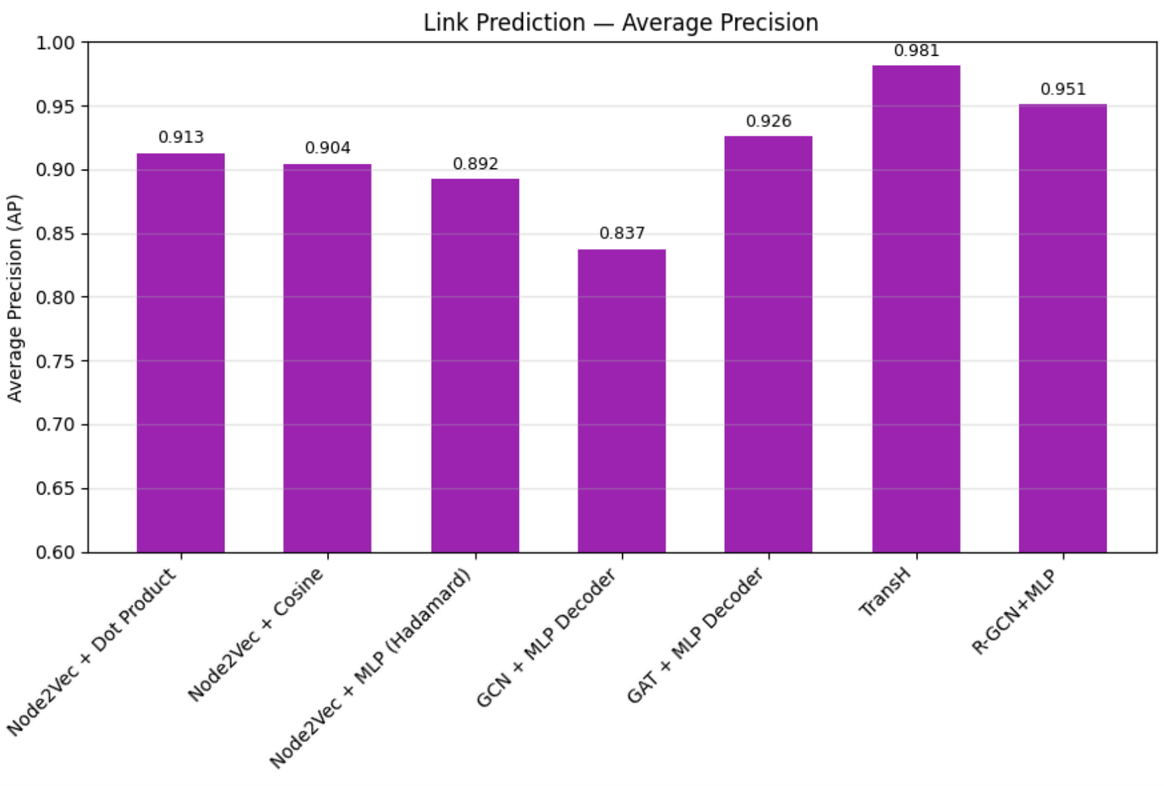
Link Prediction

Real-world use: predicting missing citations to help researchers find relevant prior work.

Model	Decoder	What it uses	Why / Motivation
Node2Vec + Scorer	Dot product / Cosine / MLP	Structure only (walk embeddings)	Shallow baseline — can walk-based proximity alone predict whether two papers cite each other?
GCN + MLP decoder	Learned MLP on node pair	Structure + features	Deep baseline — do richer node embeddings (combining citations + BoW) improve link predictions over shallow methods?
GAT + MLP decoder	Learned MLP on node pair	Structure + features + attention	Do papers that share many high-attention neighbours also tend to cite each other? Attention-weighted aggregation may better capture the latent citation affinity between papers
R-GCN + MLP decoder	Learned MLP on node pair	Directed structure + features	Encodes cites and cited_by as separate message-passing channels, making the GNN direction-aware
TransH	Hyperplane translation ($h_\perp + d_r \approx t_\perp$)	Directed structure	Handles one-to-many citation patterns while preserving direction

Link Prediction - Results





CORA Pipeline Conclusion

Finding	Takeaway
Structure + Features is the winning combination	GCN (88.6%) and GAT (88.0%) surpass all baselines — propagating BoW features through the citation graph outperforms using either alone by 7–13pp
Citation structure is surprisingly powerful on its own	Node2Vec reaches 81% without any text features. Cora's high homophily makes the graph topology a natural topic predictor
GNNs learn separable, class-coherent representations	Inter/intra-class distance ratio reaches **2.06 (GAT)** vs. ≈1.0 for shallow methods — supervised training explicitly shapes the embedding space for classification
Depth hurts: 2 layers is the sweet spot	All GNN architectures peak at 2 layers and degrade with depth — on Cora's dense, homophilic graph, deeper aggregation blurs local class signals (over-smoothing)
GNN accuracy is bounded by neighborhood composition	Papers in same-class neighborhoods: 96–98% accuracy. Bridge papers between communities: 29–35% error rate — mixed-class aggregation actively misleads the model
For link prediction, Attention is more critical than for node classification. Directionality matters.	GAT + MLP achieves AUC 0.924 (+9pp over GCN). TransH (AUC 0.972) and R-GCN (0.959) both surpass GAT+MLP (0.924) — encoding citation direction significantly improves link prediction.

Thanks!