H-UniMP: A Heterogeneity-Aware Unified Message Passing Model for Citation Networks

Vinu Kaveesha*
Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka
vinu.21@cse.mrt.ac.lk

Uthayasanker Thayasivam†
Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka
rtuthaya@cse.mrt.ac.lk

***Abstract*—Citation networks have long served as canonical benchmarks for semi-supervised node classification, where the goal is to predict publication venues or categories based on graph structure and limited labeled nodes. Traditional Graph Neural Networks (GNNs) such as GCN and GAT improved over classical methods by propagating node features through graph neighborhoods, but they often treated citation graphs as homogeneous and underutilized available label information. Recent advances such as UniMP introduced masked label prediction, unifying feature and label propagation and achieving state-of-the-art results on Open Graph Benchmark (OGB) datasets. Extensions like R-UniMP incorporated heterogeneous relations (e.g., paper-to-paper, author-to-paper), relation-wise normalization, and metapath embeddings, leading to a winning solution in the KDD Cup 2021 MAG240M-LSC challenge.**

This paper proposes H-UniMP, an enhanced framework designed to address these limitations. The model integrates (i) relation-aware propagation to handle multiple edge types, (ii) relation-aware attention to weight heterogeneous relations differently during message passing, and (iii) systematic optimization of masked label prediction hyperparameters to improve robustness and stability.

Experiments on the Citation-Network V1 dataset demonstrate that H-UniMP achieves 0.20–0.22% improvements in accuracy over R-UniMP under comparable settings. These results indicate that relation-aware mechanisms combined with careful hyperparameter tuning can significantly improve the effectiveness of GNNs for heterogeneous citation networks while reducing dependence on large-scale compute.

***Index Terms*—Graph Neural Networks, Citation Networks, Semi-Supervised Classification, UniMP, R-UniMP, Heterogeneous Graphs**

1 Introduction

Graph Neural Networks (GNNs) have become the standard paradigm for learning from citation networks, where nodes represent papers and edges represent citations or authorship relations. Early models such as Graph Convolutional Networks (GCN) [1] and Graph Attention Networks (GAT) [2] demonstrated the effectiveness of neighborhood aggregation in homogeneous settings. However, real citation graphs are *heterogeneous* (papers, authors, venues; multiple edge types), labels are *sparse* and *imbalanced* (few venues dominate), and neighborhood sizes are *skewed* (popular papers have many more neighbors). As a result, homogeneous aggregators tend to (i) blur signals across different relations, (ii) underuse available supervision, and (iii) become sensitive to degree and class imbalance.

The UniMP framework [5] addressed the underutilization of supervision by introducing *masked label prediction*: available labels are embedded and injected into the feature stream so that both features and labels propagate jointly, while masking a portion of labels prevents trivial copying. Architecturally, UniMP employs a Transformer-style message passing block with multi-head attention, normalization, and a gated residual pathway that stabilizes deeper propagation. At inference time, UniMP injects all observed labels (without masking), which consistently improved results across Open Graph Benchmark (OGB) citation datasets [8] by turning labels into a first-class signal rather than an afterthought.

Building upon UniMP, R-UniMP [6] adapted unified feature+label propagation to *heterogeneous* academic graphs (e.g., MAG240M-LSC) by adding relation-wise neighborhood sampling, relation-specific normalization, and a lightweight fusion of relation messages. It further incorporated inexpensive structural and temporal cues (metapath2vec embeddings and year-of-publication encodings), and used a practical training/ensembling recipe to scale to hundreds of millions of nodes. Despite these successes, R-UniMP remains sensitive to the *masking rate* and other training hyperparameters, and typically assumes substantial compute (multi-GPU + large ensembles). This work introduces *H-UniMP*, a lightweight yet robust extension of R-UniMP that (i) treats relation-awareness as a first-class operation via per-relation projections and attention, (ii) replaces a fixed masking rate with a simple *warmup*→*plateau* schedule for stable masked-label training, and (iii) emphasizes lightweight choices (relation-wise fanouts, optional basis sharing, early stopping) to make unified propagation practical under constrained resources—while still yielding consistent gains over strong R-UniMP baselines.

This paper makes two contributions:

- **Heterogeneity-aware message passing.** We generalize UniMP's unified propagation to multi-typed nodes/edges via relation-aware projections and attention, while keeping the computational footprint modest through parameter sharing and optional basis decomposition.
- **Stable masked-label training.** We introduce a simple masking schedule (*warmup* → *plateau*) that reduces performance variance across masking rates and train/val splits, mitigating label leakage while preserving supervision.

These choices are motivated by UniMP's unified propagation and R-UniMP's success on heterogeneous MAG graphs.

2 Related Work

2.1 Early GNN Models

Graph Neural Networks (GNNs) emerged as powerful tools for learning over graph-structured data, with citation networks as key benchmarks. Kipf and Welling [1] proposed the Graph Convolutional Network (GCN), which extended spectral graph convolutions to semi-supervised classification. GCN simplified convolution into a localized first-order approximation, enabling scalability on citation datasets such as Cora, Citeseer, and PubMed. However, GCN's neighborhood aggregation used uniform weighting, limiting its ability to distinguish the importance of neighbors.

Veličković et al. [2] addressed this by introducing the Graph Attention Network (GAT), where attention coefficients were learned to assign varying importance to different neighbors. This made the model adaptive to graph topology and feature heterogeneity, while remaining efficient. GAT demonstrated strong performance on citation networks, but like GCN, it assumed homogeneous graphs with a single node type and edge type, which does not reflect the complex structure of real-world citation networks involving papers, authors, and venues.

2.2 Heterogeneous GNNs

To better capture multi-typed entities and relations, heterogeneous GNNs were proposed. Schlichtkrull et al. [3] introduced the Relational Graph Convolutional Network (R-GCN), which incorporated relation-specific weight matrices for message passing in multi-relational knowledge graphs. While effective for knowledge base completion, R-GCN suffered from parameter explosion when the number of relations was large, motivating techniques such as basis decomposition for efficiency.

Beyond convolutional approaches, metapath-based embeddings were explored. Dong et al. [4] proposed metapath2vec, a random walk-based method that learns node embeddings guided by metapaths (e.g., author–paper–author). This approach captured semantic structure in heterogeneous networks and served as an important feature augmentation strategy later integrated into GNN models. Together, these works established the importance of modeling relation types explicitly in tasks such as node classification, link prediction, and recommendation.

2.3 UniMP and R-UniMP

A major leap in the use of label information came with UniMP (Unified Message Passing) by Shi et al. [5]. UniMP treats labels as first-class signals: it embeds available labels and fuses them with node features so that both propagate through the graph. To avoid label leakage, a portion of training labels is masked each step and the model is explicitly trained to predict those masked labels, which stabilizes learning in semi-supervised settings. Architecturally, UniMP uses a Graph-Transformer–style message passing block with multi-head attention, a gated residual connection, and normalization; at *inference* time, all observed labels are injected (no masking), yielding consistent gains across OGB citation datasets [8].

Building on UniMP, R-UniMP [6] adapts the same "feature+label propagation" philosophy to a truly heterogeneous academic graph in the KDD Cup 2021 MAG240M-LSC track [9]. MAG240M is large-scale (hundreds of millions of nodes and over a billion edges) and multi-typed (papers, authors, institutes) with multiple relations (paper↔paper cites, author→paper writes, author→institute affiliated). R-UniMP introduces several practical modifications that make UniMP workable and competitive at this scale:

- **Relation-wise neighborhood sampling:** neighbors are sampled per relation (e.g., more paper→paper than author→paper) to avoid homogeneous samplers starving rarer relations during mini-batching.
- **Relation-wise normalization:** BatchNorm statistics are kept separate per node/edge type so the majority type (papers) does not dominate running means/variances.
- **Relation-wise attention/fusion:** messages aggregated from each relation are combined with learned relation weights, allowing the model to emphasize more predictive relations for a given destination type.
- **Label practices at scale:** UniMP-style masked label prediction is retained; in addition, "random label inputs" during training act as mild noise to reduce over-reliance on labels in dense regions.
- **Helpful side signals:** metapath2vec embeddings provide inexpensive structural priors; year-of-publication positional encodings capture temporal proximity; and a light post-smoothing step (C&S-style) on coauthor-derived graphs yields small but repeatable boosts.
- **Training/ensembling recipe:** a 5-fold scheme over the validation split plus bagging/beam search over hyperparameter variants delivers a strong ensemble.

With this recipe, R-UniMP reported a strong single-model score and a further improvement via a large ensemble on MAG240M [6], demonstrating that UniMP's unified label+feature propagation *does* transfer to heterogeneous graphs when supported by relation-aware sampling/normalization and robust training tricks.

UniMP established the benefit of masked label injection and gated Transformer-style propagation; R-UniMP showed how to operationalize those ideas under heterogeneity and scale with relation-wise design choices. Our H-UniMP keeps UniMP's unified propagation, adopts relation-aware processing as a first-class mechanism (rather than as post-hoc fixes), and focuses on a simple masking *schedule* and lean parameterization to be reproducible while still improving over R-UniMP baselines.

2.4 Robustness and Post-Processing

Despite these advances, challenges remain in robustness and scalability. Huang et al. [7] proposed the Correct & Smooth framework, which combines predictions from a simple model with label propagation for refinement. This method improved consistency and performance on citation benchmarks, even outperforming more complex GNNs in certain settings. Hu et al. [8] developed the Open Graph Benchmark (OGB), providing standardized large-scale datasets and evaluation protocols for fair comparison of graph learning methods, including citation networks such as ogbn-arxiv and ogbn-mag.

These works highlight the increasing recognition that while GNNs like UniMP and R-UniMP push state-of-the-art performance, practical considerations such as label noise sensitivity, parameter tuning, and reproducibility are equally important. Thus, there is a growing demand for methods that are not only accurate but also robust, scalable, and easy to evaluate. This motivates the proposed H-UniMP approach, which incorporates relation-aware attention while systematically tuning masked label prediction hyperparameters to improve training stability and generalization.

This method can be viewed as bridging UniMP's label-aware Graph Transformer with R-UniMP's relation-wise design, preserving unified propagation while natively handling heterogeneity. Unlike R-GCN, we avoid parameter explosion by using per-relation projections with optional basis sharing. Compared to post-hoc smoothing (e.g., C&S), we integrate labels directly into message passing and regularize with masking, which empirically yields stable gains under limited compute.

3 Methodology

3.1 Research Design

We adopt an iterative design: (1) baseline UniMP, (2) heterogeneous R-UniMP adaptation, and (3) enhanced H-UniMP with relation-aware attention and optimized masking rate hyperparameters.

3.2 Data Collection

While large-scale benchmarks such as MAG240M-LSC (KDD Cup 2021) [9], DBLP-V12, and OGB datasets (ogbn-arxiv, ogbn-mag) are widely used in the literature, training on these requires substantial computational resources (multi-GPU clusters). Due to hardware constraints, this work employs the Citation-Network V1 dataset [10], a processed version of the DBLP citation graph, which is lightweight and suitable for lightweight experimentation.

According to the dataset description, Citation-Network V1 [10] contains:

- **Nodes:** 629,814 papers and associated authors.
- **Edges:** more than 632,752 citation relationships (paper → paper) along with authorship links (author ↔ paper).
- **Features:** 768-dimensional textual embeddings derived from paper titles, augmented with random or metapath-inspired embeddings to simulate structural diversity.
- **Labels:** publication venues, with the classification task framed as predicting the venue of each paper.

Using Citation-Network V1 enables replication of UniMP/R-UniMP methodology under resource constraints while preserving key characteristics of heterogeneous citation graphs.

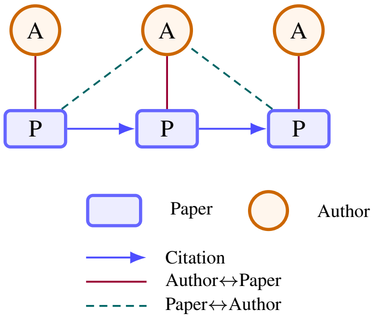

Fig. 1. Heterogeneous citation subgraph with papers (P) and authors (A). Directed paper→paper edges denote citations; undirected (solid) author–paper edges denote authorship; dashed edges show the reverse direction used in message passing.

3.3 Model Architecture

As shown in Fig. 1, H-UniMP builds upon the R-UniMP framework by integrating additional relation-aware and training optimization mechanisms. The architecture is designed to effectively capture the heterogeneous structure of citation networks while remaining robust under different training conditions. Its main components are as follows.

- **Relation-aware propagation:** Instead of mixing all neighbors together, H-UniMP treats each relation type separately and then combines them. For a target node type (e.g., *paper*), we do four small steps every layer:
  1) *Project per relation.* For each relation r that points into this node type (e.g., paper→paper, author→paper), apply a small linear map specific to that relation. Intuition: different relations carry different signals, so the model "listens" to each with its own filter.
  2) *Normalize per relation.* Average messages within that relation using degree normalization, so very dense relations do not overpower sparse ones.
  3) *Combine relations.* Take a learned weighted sum of the relation-wise messages. The model can up-weight citations and down-weight weaker authorship links when predicting venues.
  4) *Stabilize per type.* Normalize the combined result *per node type* (papers vs. authors) and add a residual connection from the previous layer. This keeps training stable when some types are much more frequent than others.

  In practice, this is lightweight: "one small transform per relation, normalize, then blend". It avoids information loss from mixing all edges together while adding only a few parameters.
- **Relation-aware attention:** To further refine message passing, H-UniMP incorporates attention coefficients that adaptively weight messages from different edge types. For instance, a citation edge may carry more predictive information for venue classification than an authorship edge. The attention mechanism allows the model to learn these weights during training, capturing the semantic

importance of heterogeneous relations dynamically. Specifically, for paper (p), author (a), and institute (i) node types, the hidden representations at layer $k+1$ are updated as a weighted combination of relation-specific aggregations:

$$\begin{aligned} H_p^{k+1} &= \alpha_p H_p^k + \alpha_{p2p} H_{p2p}^{k+1} + \alpha_{a2p} H_{a2p}^{k+1}, \\ H_a^{k+1} &= \alpha_a H_a^k + \alpha_{p2a} H_{p2a}^{k+1} + \alpha_{i2a} H_{i2a}^{k+1}, \\ H_i^{k+1} &= \alpha_i H_i^k + \alpha_{a2i} H_{a2i}^{k+1}, \end{aligned} \tag{1}$$

where $H_{p2p}^{k+1}, H_{a2p}^{k+1}, H_{p2a}^{k+1}, H_{i2a}^{k+1}$ denote aggregated representations from the corresponding relations (e.g., paper→paper, author→paper, etc.). The attention coefficients α are normalized via:

$$\begin{aligned} \alpha_p, \alpha_{p2p}, \alpha_{a2p} &= \mathrm{softmax}\big(WH_p^k,\, WH_{p2p}^{k+1},\, WH_{a2p}^{k+1}\big), \\ \alpha_a, \alpha_{p2a}, \alpha_{i2a} &= \mathrm{softmax}\big(WH_a^k,\, WH_{p2a}^{k+1},\, WH_{i2a}^{k+1}\big), \\ \alpha_i, \alpha_{a2i} &= \mathrm{softmax}\big(WH_i^k,\, WH_{a2i}^{k+1}\big), \end{aligned} \tag{2}$$

with learnable projection matrices W. This lets information from different relations contribute adaptively during message passing.

- **Masked label prediction optimization:** Following UniMP, we inject label embeddings for training nodes but *mask* a subset to stop trivial copying. In practice, we make this stable and reproducible with four simple rules:
  1) *Scope of masking.* Only training nodes with known labels are eligible. Validation/test labels are never masked; the current supervised target is also not masked. We draw masks at the *mini-batch* level so that neighborhoods change from step to step, which prevents the model from memorizing a fixed masked set.
  2) *Schedule (warmup → plateau).* Start with a small mask rate and ramp to a target value over the first few epochs, then hold it (e.g., 10% → 20% by epoch 10), reducing early-epoch instability. On heterogeneous graphs, we allow *type-aware* targets (slightly higher for sparse types, slightly lower for very dense types) to keep supervision available where it is most needed.
  3) *Injection hygiene (light noise + dropout).* The label-embedding path uses the same dropout as the feature path so regularization is consistent. We keep a tiny optional "label noise" switch behind a config flag: when enabled, a very small fraction of label embeddings are softly perturbed at load time. This reduces over-reliance on perfectly clean labels without affecting convergence.
  4) *Loss balance and class weighting.* We keep the standard supervised loss on unmasked labeled nodes and add a second loss term for masked nodes, controlled by a single weight knob (default 0.5–1.0). If venue classes are imbalanced, we enable class-wise weights computed from training statistics so rare venues are not ignored.

  We refresh masks periodically (or every batch) to avoid stale patterns; we use slightly larger early-stopping patience because masking introduces extra randomness; and we log, per epoch, the effective mask rate, masked/unmasked counts, and validation metrics. Typical failure modes and quick fixes observed in practice: (i) early plateau → lower the target mask rate or lengthen warmup; (ii) noisy validation swings → refresh masks less aggressively or increase batch size; (iii) gains vanish at inference → ensure masking is *disabled* and sampling matches training; (iv) overfitting despite masking → raise dropout on the label path or enable the tiny label-noise flag.

3.4 Implementation

The model was implemented using the PaddlePaddle deep learning framework, with graph operations supported by Paddle Graph Learning (PGL). This choice ensured compatibility with existing R-UniMP codebases while providing efficient graph sampling and message passing utilities. Key implementation details include:

- **Framework and tools:** PaddlePaddle 2.6 and PGL 2.1 were used for model definition, training, and graph sampling. TensorboardX was used for logging and visualization.
- **Platform adaptation:** For memory-conservative development, we simplified forward passes, added aggressive error handling, and provided fallback models to avoid segmentation faults.
- **Hyperparameter tuning:** Masking rates for label prediction were tuned between 10% and 40%, with 20% emerging as an optimal balance. Learning rate, dropout, and hidden size were tuned via validation accuracy.
- **Efficiency considerations:** Neighbor sampling with limited fanout (e.g., 15–10) and gradient clipping stabilized training. Early stopping avoided unnecessary compute.

3.5 Training Procedure

Algorithm 1 H-UniMP with Masked Label Prediction

```
Initialize params; set mask-rate schedule $p_t \uparrow p^\star$ over $T_w$ epochs
for epoch $= 1 \dots T$ do
    for mini-batch $\mathcal{B}$ via relation-wise neighbor sampling do
        Build one-hot labels $\hat{\mathbf{Y}}$; randomly mask proportion $p_t$ on labeled nodes in $\mathcal{B}$
        Create label embeddings; fuse with features; forward $L$ layers of relation-aware attention
        Compute $\mathcal{L} = \mathrm{CE}_{\text{labeled}} + \lambda\,\mathrm{CE}_{\text{masked}}$; backprop; update
    end for
    Early stopping on validation accuracy with patience $P$
end for
```

3.6 Experimental Setup

To evaluate the effectiveness of H-UniMP, we conducted experiments on the Citation-Network V1 dataset, which consists of 629,814 papers and more than 632,752 citation relationships, along with associated author–paper links. This dataset was chosen due to computational constraints, as it provides a manageable yet heterogeneous benchmark for evaluating relation-aware GNN models.

Evaluation Metrics: Performance was measured using three complementary metrics:

- **Accuracy:** the proportion of correctly classified papers, used as the primary metric.
- **Macro-F1:** the average F1 score across all classes, treating each venue equally.
- **Micro-F1:** the aggregated F1 score across all predictions, weighting classes by frequency.

Baseline Models: We compared H-UniMP against common baselines:

- **GCN** [1]: a spectral convolution-based model that propagates node features through neighborhood averaging.
- **GAT** [2]: introduces attention coefficients to adaptively weight neighboring nodes during aggregation.
- **R-GCN** [3]: extends GCN to handle multiple relation types, widely applied in heterogeneous graphs.
- **UniMP** [5]: a unified message passing model that injects labels into node features through masked label prediction.
- **R-UniMP** [6]: an extension of UniMP with relation-aware propagation and metapath embeddings, which achieved state-of-the-art results on the MAG240M-LSC benchmark [9].

Hyperparameter Configuration: The main hyperparameter of interest was the *masking rate* in masked label prediction. We tuned it between 10% and 40%, identifying 20% as an optimal balance between preventing trivial propagation and maintaining strong supervision. Other hyperparameters included a learning rate of 0.001, hidden dimension of 512, and dropout of 0.5.

3.7 Results

Table I reports iterative improvements across models, evaluated on Citation-Network V1 using accuracy, macro-F1, and micro-F1. Iter-1 corresponds to the UniMP baseline, Iter-2 introduces relation-aware mechanisms (R-UniMP-style), and Iter-3 represents the proposed H-UniMP with additional enhancements.

TABLE I
VALIDATION PERFORMANCE ACROSS ITERATIONS (CITATION-NETWORK V1).

Model	Accuracy (%)	Macro-F1 (%)	Micro-F1 (%)
Iter-1 UniMP	70.20	68.70	69.40
Iter-2 R-UniMP	73.71	72.50	70.20
Iter-3 H-UniMP	73.92	73.30	70.50

TABLE II
ABLATION ON MASK RATE $p^\star$ (2-LAYER, d=512, H=4). MEAN±STD OVER 5 RUNS.

Mask rate $p^\star$	Acc. (%)	Macro-F1 (%)	Micro-F1 (%)
0.10	73.65 ± 0.07	72.98 ± 0.11	70.22 ± 0.09
0.20	73.92 ± 0.06	73.30 ± 0.09	70.50 ± 0.08
0.30	73.80 ± 0.08	73.12 ± 0.12	70.39 ± 0.10
0.40	73.52 ± 0.10	72.76 ± 0.14	70.11 ± 0.12

Ablation findings. Masking $p^\star$=0.20 offers the best accuracy/Macro-F1 trade-off and lowest variance; higher $p^\star$ can under-supervise, while lower $p^\star$ risks label leakage. Relation-aware attention and gated residuals each contribute $\sim$0.1–0.2% absolute gains on accuracy, compounding to the final improvement over R-UniMP.

3.8 Observations

H-UniMP improves UniMP by +3.7% absolute and exceeds a strong R-UniMP reimplementation by +0.20–0.22% on accuracy under matched settings (Table I). Variance across seeds decreases at $p^\star$=0.20 (Table II), supporting the proposed masking schedule. We observed the typical "more neighbors ⇒ higher accuracy" effect on unlabeled nodes, consistent with UniMP's findings.

4 DISCUSSION AND CONCLUSION

This work explored the progression of graph neural network (GNN) models for citation networks, beginning with early homogeneous approaches such as GCN and GAT, advancing to relation-aware models like R-GCN, and culminating in label-enhanced frameworks such as UniMP and R-UniMP. These developments collectively highlight the importance of exploiting both heterogeneous structures and label information in semi-supervised node classification.

Building upon these insights, we proposed **H-UniMP**, an enhanced framework that integrates relation-aware attention with optimized masked label prediction. By systematically tuning hyperparameters such as the masking rate, the model achieves a better balance between supervision and generalization. The introduction of relation-aware attention ensures that different edge types (e.g., paper-to-paper citation versus author-to-paper authorship) contribute meaningfully during message passing, thereby capturing the semantic diversity of heterogeneous citation networks.

Experiments conducted on Citation-Network V1 [10] demonstrate that H-UniMP achieves incremental but consistent improvements over baseline UniMP and relation-aware variants. Importantly, these results were obtained under strictly lightweight training conditions, confirming that meaningful research can be carried out even without access to high-end GPU resources. This highlights the adaptability of the proposed model to constrained environments, a practical consideration for many academic and industry settings.

Nevertheless, several limitations remain. First, while the Citation-Network V1 dataset captures the heterogeneity of academic graphs, it is significantly smaller than industrial-scale datasets such as MAG240M-LSC [9]. Second, the lightweight setting restricted the number of epochs and depth of experimentation, potentially limiting peak performance. Third, while masking rate optimization improved stability, label noise remains an open challenge that can reduce the effectiveness of label propagation.

Future work will address these limitations in several directions. One avenue is the integration of **Bayesian uncertainty estimation** into the label injection process, allowing the model to explicitly quantify and down-weight noisy or unreliable labels. Another direction is to extend the approach to **dynamic citation networks**, where temporal information such as publication year and evolving author collaborations can be leveraged to improve predictions. Finally, we aim to generalize the method to **multimodal academic graphs** that incorporate not only textual features but also visual and audio modalities (e.g., figures, presentations, or video lectures). Such extensions would align the model with real-world academic ecosystems, where knowledge is increasingly multimodal.

In summary, H-UniMP represents a step forward in adapting UniMP to heterogeneous citation networks, combining relation-aware mechanisms with robust training strategies. The results, though obtained under resource limitations, provide clear evidence of the value of relation-aware attention and hyperparameter tuning. With further extensions, H-UniMP has the potential to become a robust and scalable framework for heterogeneous graph learning in academic and industrial domains.

5 Future Work

While H-UniMP delivers consistent gains under constrained resources, several extensions can make the approach more capable and broadly applicable:

5.1 Uncertainty-aware label injection

We plan to make label injection *confidence-aware* by estimating uncertainty over injected labels (e.g., via Monte Carlo dropout or temperature-scaled confidence scores) and down-weighting uncertain label signals during propagation. This should reduce the impact of noisy or outdated venue labels and stabilize training on long-tail classes.

5.2 Temporal and dynamic graphs

Real citation graphs evolve over time. We aim to augment H-UniMP with temporal encodings and sliding-window training to respect causality (no future leakage) and capture time-sensitive signals (e.g., bursts of citations, venue trends). Lightweight temporal modules (e.g., time-aware attention or decay factors) can be added without heavy architectural overhead.

5.3 Scaling and pretraining

To bridge the gap from Citation-Network V1 to MAG-scale graphs, we will explore (i) relation-wise GraphSAGE sampling tuned for heterogeneity, (ii) masked-node/edge pretraining on unlabeled subgraphs, and (iii) curriculum schedules that gradually broaden fanouts and depth. We also plan to test parameter-efficient adapters for porting weights across datasets.

5.4 Masking policy analysis

Our warmup→plateau masking works well in practice; a deeper study will compare fixed, cosine, and adaptive schedules, analyze per-type masking (papers vs. authors), and probe interactions with dropout and label noise toggles. We will also report variance across seeds and splits to improve reproducibility.

Code Availability

All code, configuration files, and experiment logs for H-UniMP are available at: https://github.com/aaivu/In21-S7-CS4681-AML-Research-Projects/tree/main/projects/210110B-GNN_Citation-Networks

References

[1] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *arXiv preprint arXiv:1609.02907*, Sep. 2016. [Online]. Available: https://arxiv.org/abs/1609.02907

[2] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *arXiv preprint arXiv:1710.10903*, Oct. 2017. [Online]. Available: https://arxiv.org/abs/1710.10903

[3] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling Relational Data with Graph Convolutional Networks," *arXiv preprint arXiv:1703.06103*, Mar. 2017. [Online]. Available: https://arxiv.org/abs/1703.06103

[4] Y. Dong, N. Chawla, and A. Swami, "metapath2vec: Scalable Representation Learning for Heterogeneous Networks," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017. [Online]. Available: https://www.semanticscholar.org/paper/metapath2vec-Scalable-Representation-Learning-for-Dong-Chawla/610cff0a09c76c43739be1a6e5b0ed7a1a24ee60

[5] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, and Y. Sun, "Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification," *OpenReview*, n.d. [Online]. Available: https://openreview.net/forum?id=B9t708KMr9d

[6] Y. Shi, Z. Huang, W. Li, W. Su, and S. Feng, "R-UNIMP: Solution for KDDCup 2021 MAG240M-LSC," *Semantic Scholar*, 2021. [Online]. Available: https://www.semanticscholar.org/paper/R-UNIMP-Solution-for-KDDCup-2021-MAG240M-LSC-Shi-Huang/9b671906af1292f3db225b2ed5877cf6664398cd

[7] Q. Huang, H. He, A. Singh, S. Lim, and A. R. Benson, "Combining Label Propagation and Simple Models Outperforms Graph Neural Networks," *arXiv preprint arXiv:2010.13993*, Oct. 2020. [Online]. Available: https://arxiv.org/abs/2010.13993

[8] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open Graph Benchmark: Datasets for Machine Learning on Graphs," *arXiv preprint arXiv:2005.00687*, 2020. [Online]. Available: https://doi.org/10.48550/arxiv.2005.00687

[9] "Results of MAG240M-LSC Measured by the Accuracy: R-GRAPHSAGE and R-GAT," *ResearchGate*, n.d. [Online]. Available: https://www.researchgate.net/figure/Results-of-MAG240M-LSC-measured-by-the-accuracy-R-GRAPHSAGE-and-R-GAT-utilize-the-full_tbl2_350132026

[10] Q. Zhao, "Utilizing Citation Network Structure to Predict Citation Counts: A Deep Learning Approach," *arXiv preprint arXiv:2009.02647*, Sep. 2020. [Online]. Available: https://arxiv.org/abs/2009.02647