

Problem and Motivation

GNNs learn graph structure via message passing, where nodes iteratively aggregate neighbor information. On large KGs, this is inherently inefficient due to three compounding factors:

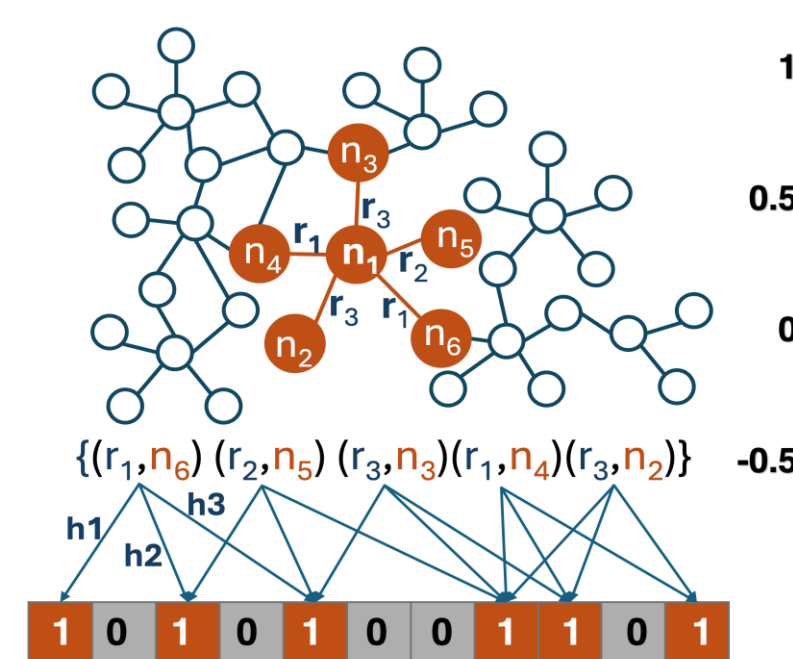
- mini-batch training requires neighbor sampling, discarding structural information.
- relation-aware models become memory-prohibitive as relation types grow.
- information from distant and infrequent relations degrades across aggregation rounds.

Key insight:

Shift complexity of capturing KG structure from message-passing to a **lightweight preprocessing step** that computes structural features encoding local and global graph structure.

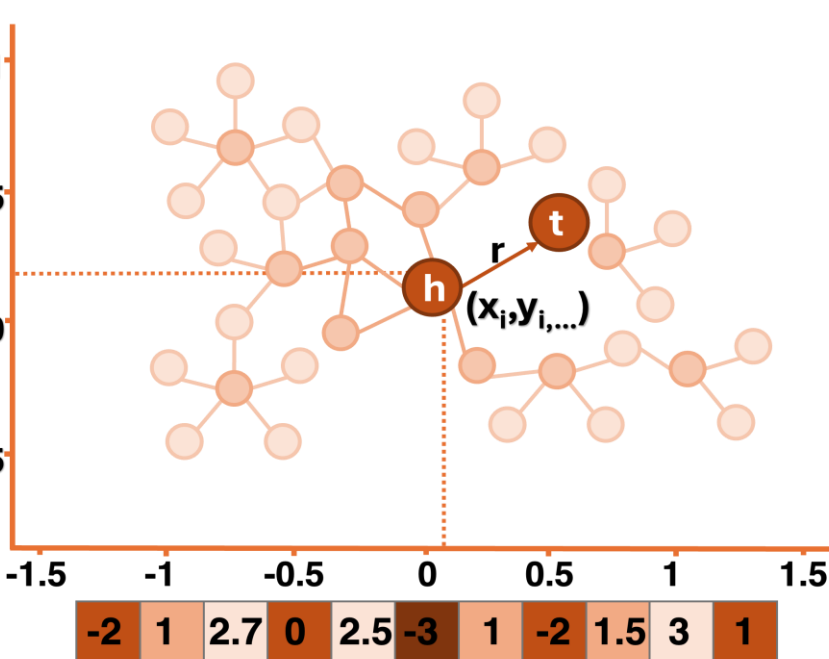
Structural Encoding

Bloom Filter – Local View



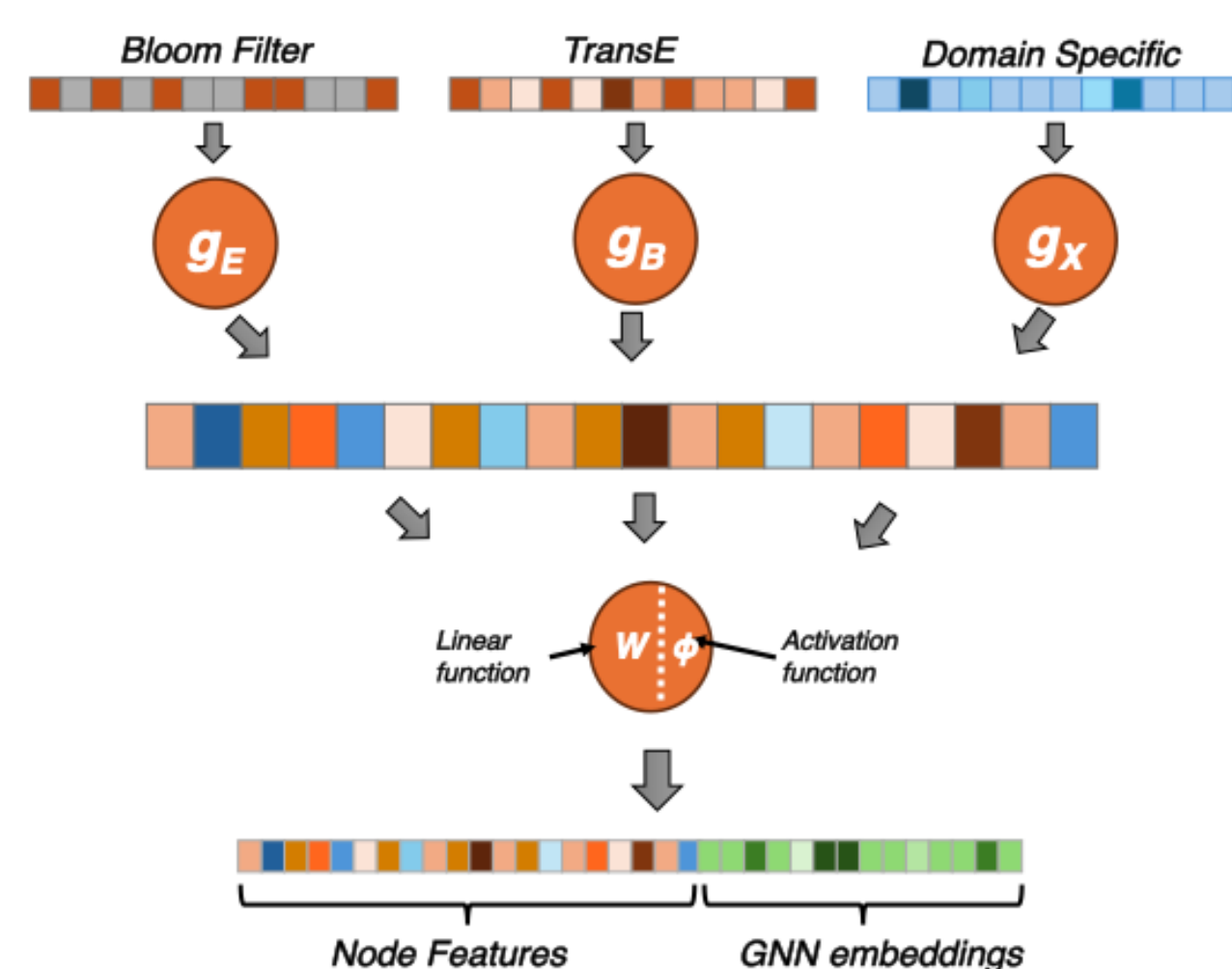
Encodes 1-hop neighbors + relation types into fixed-length bit vector. One pass over triples.

TransE Embeddings – Global View



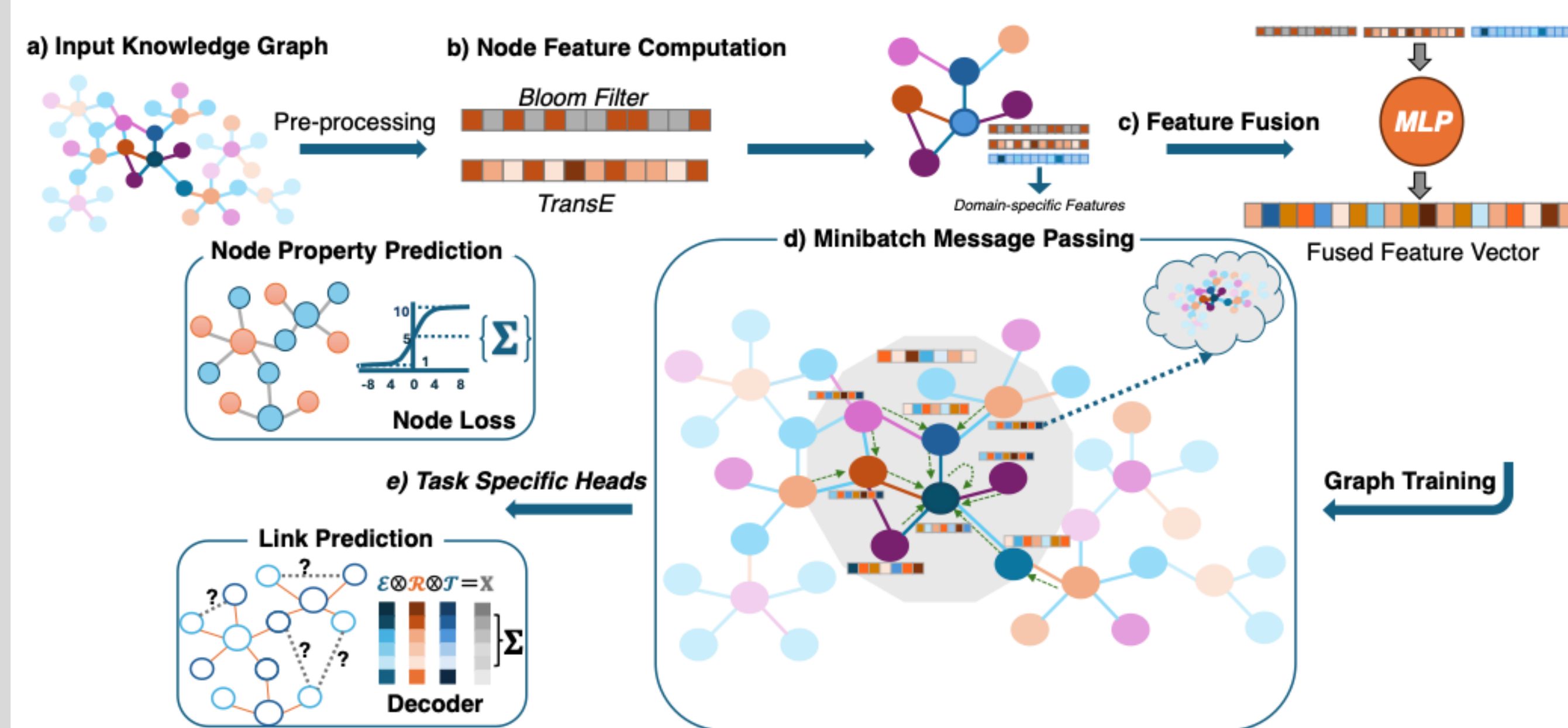
Models relations as translations $h+r=t$. Captures global position & semantic context

Feature Fusion



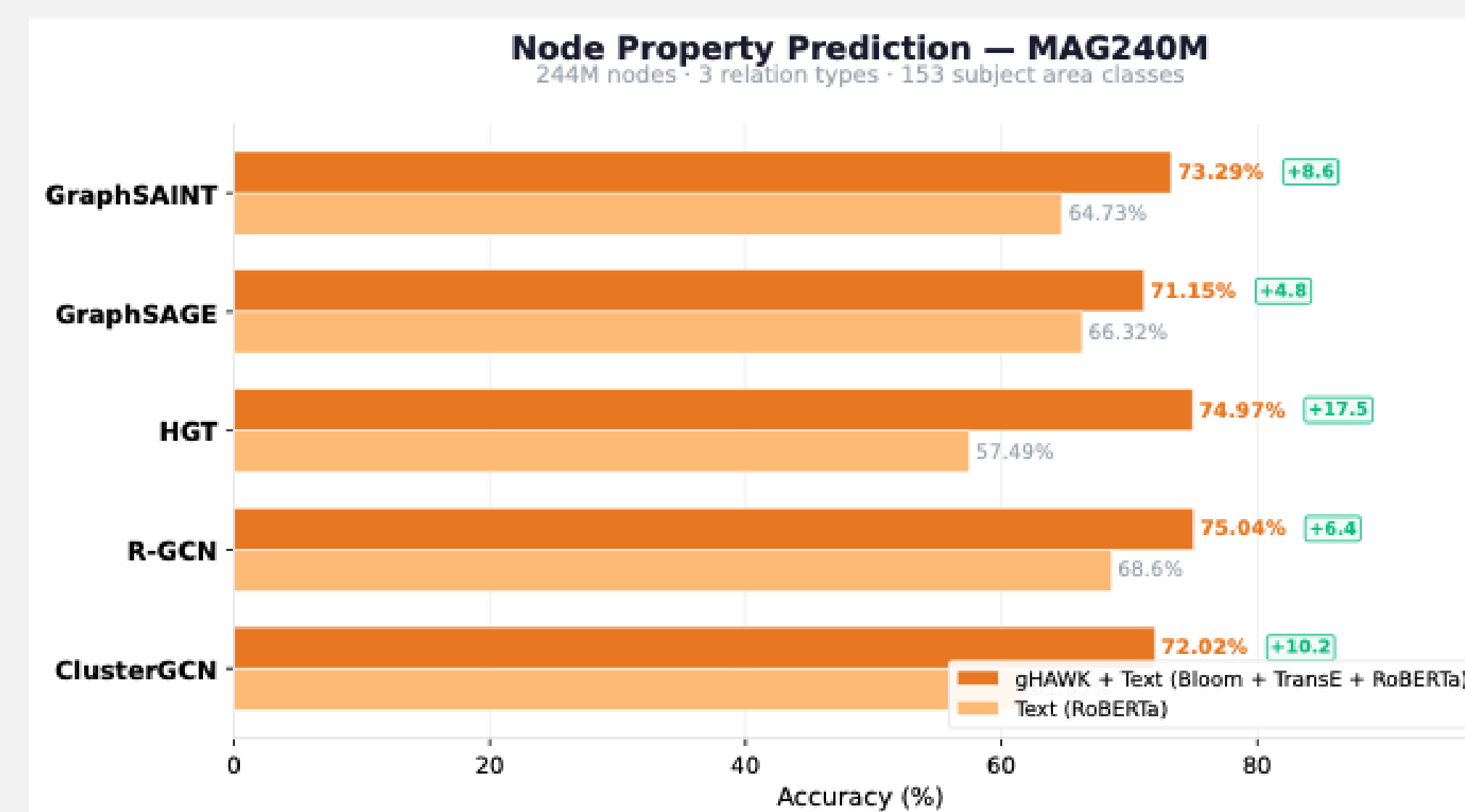
Learned projections harmonize heterogeneous feature spaces. Trained end-to-end with GNN

Ghawk Pipeline

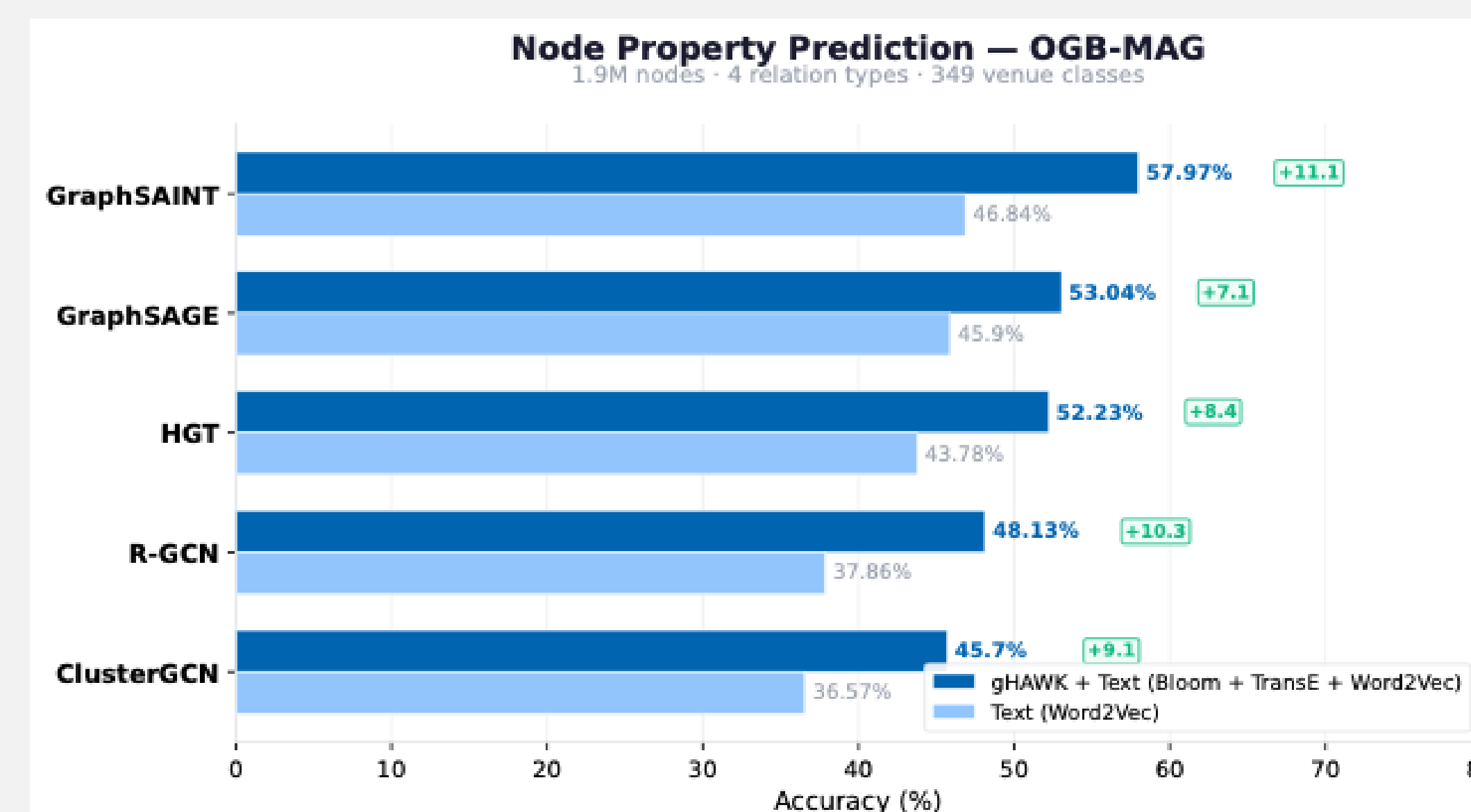


Node Classification Results

gHAWK improves accuracy by 7–11 percentage points across all GNN backbones. Structural features (Bloom + TransE) are more informative than text features alone.

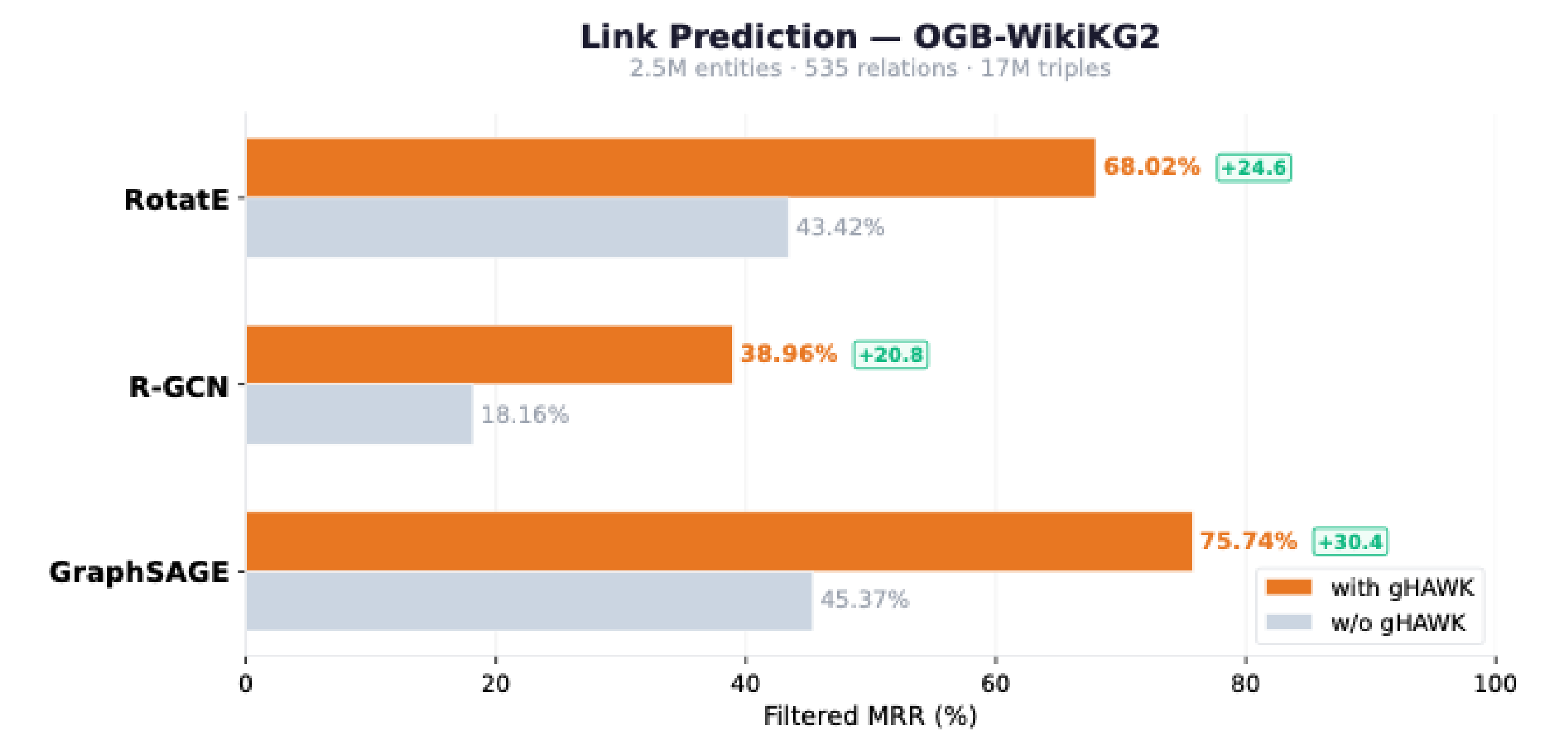


On the web-scale MAG240M (244M nodes), gHAWK improves accuracy by up to 17.5 percentage points. R-GCN with gHAWK achieves 75.04% — **ranking #1 on the OGB leaderboard.**



Link Prediction Results

gHAWK improves MRR by up to 30 points on OGB-WikiKG2, demonstrating that its structural features effectively capture relational semantics and provide every backbone a strong structural prior before training begins.



gHAWK ranks #1 on OGB leaderboards for OGB-WikiKG2 without external data or model ensembles.

Key Contributions

- Introduce a preprocessing framework that encodes local and global KG structure as node features, compatible with any GNN architecture.
- Compact Bloom filters represent each node's full 1-hop multi-relational neighborhood in a fixed-size bit vector.
- TransE embeddings capture global graph position, semantic context, and signals from infrequent relations.
- A learned fusion module combines structural and domain-specific features, trained end-to-end with the GNN.
- Ranks #1 on three OGB leaderboards without external data or ensembles.

Conclusion

- Preprocessing replaces expensive iterative structure learning with lightweight Bloom filters and TransE embeddings.
- Enables shallow (1–2 layer) GNNs to match or surpass deep, resource-heavy models.
- Architecture-agnostic — improves every GNN backbone tested.
- Scales to web-scale KGs with hundreds of millions of nodes.

References

- Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. NeurIPS 33 (2020), 22118–22133.
- Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In The Semantic Web: Int. Conf. ESWC. 593–607.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-Relational Data. NeurIPS 26 (2013).
- Will Hamilton, Zitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. NeurIPS 30 (2017).