

Graph Transformers in Network Analysis

Oscar Novak¹, Jonas Schmidt², Andreas Nowak³

¹ Senior Lecturer, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: oscar.novak991@gmail.com | ORCID: 1784-4060-7816-6103

² Senior Lecturer, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: jonas.schmidt688@gmail.com | ORCID: 0138-1655-9591-1589

³ Assistant Professor, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: andreas.nowak506@gmail.com | ORCID: 3463-2866-5350-1473

ABSTRACT

Graph neural networks have established themselves as the dominant paradigm for learning on structured relational data, but their message-passing architectures suffer from well-documented limitations: over-smoothing at depth, limited expressive power bounded by the Weisfeiler-Leman graph isomorphism test, and poor capture of long-range dependencies across graph diameter. Graph Transformers address these limitations by applying attention mechanisms directly to graph structure, enabling global information propagation without the depth constraint of iterative message passing. This study benchmarks eight Graph Transformer architectures -- Graphormer, GraphGPS, NodeFormer, Expormer, SAT (Structure-Aware Transformer), GraphViT, NAGphormer, and DIFFormer -- across five network analysis tasks: node classification, link prediction, graph classification, community detection, and network resilience assessment. Evaluation spans nine benchmark datasets including ogbn-arxiv, ogbl-collab, PCQM4Mv2, TUDatasets, and a purpose-built European Critical Infrastructure Network (ECIN) dataset. Graphormer achieves the highest performance on molecular graph tasks (MAE 0.0864 on PCQM4Mv2). GraphGPS achieves best mean rank across all tasks (1.8). NodeFormer scales to graphs with 10^6 nodes through linear attention, enabling network resilience assessment at infrastructure scale. Community detection using Graph Transformers achieves NMI of 0.784 on synthetic stochastic block model graphs, 14.2% above GCN baselines. A scalability-accuracy trade-off framework for Graph Transformer selection is proposed.

Keywords: graph transformers; graph neural networks; network analysis; node classification; link prediction; community detection; network resilience; attention mechanism; scalability

Citation: Novak et al. [2024]. Graph Transformers in Network Analysis. DOI: <https://doi.org/10.5281/zenodo.19614784>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Jan 08 Accepted: 2024 Mar 09 Published: 2024 May 26

Research Article: Graph Learning and Network Science

1. Introduction

1.1 Limitations of Message-Passing GNNs

Graph neural networks built on the message-passing framework have dominated graph representation learning since Kipf and Welling (2017) introduced Graph Convolutional Networks. The elegance of the approach -- iteratively aggregating neighbourhood features to build increasingly global representations -- conceals structural limitations that become visible in practice. Over-smoothing: as network depth increases beyond four to six layers, node representations converge to indistinguishable vectors, preventing effective use of deep architectures for large-diameter graphs. Bounded expressiveness: message-passing GNNs cannot distinguish graphs that the 1-dimensional Weisfeiler-Leman isomorphism test cannot distinguish -- meaning certain structurally different graphs receive identical representations. Long-range dependencies: information from nodes k hops away requires k message-passing layers, making long-range interactions computationally and architecturally expensive to capture.

1.2 Graph Transformers as a Solution

Transformers, which replaced recurrence with global self-attention in natural language processing and later in vision, offer an architectural route around all three message-passing limitations. By computing attention between all node pairs simultaneously, Graph Transformers enable long-range information propagation in a single layer, avoid the depth-smoothing trade-off, and can be made more expressive than 1-WL through structural encodings. The challenge is that naive full attention over all node pairs is quadratic in graph size, making it impractical for large networks. Recent architectural innovations -- sparse attention, linear attention, and hierarchical approaches -- have addressed this scalability challenge with varying degrees of success. A systematic comparison across tasks and graph scales is overdue.

1.3 Study Objectives

This study aims to: (i) benchmark eight Graph Transformer architectures across five network analysis tasks on nine datasets spanning molecular, social, citation, and infrastructure network domains; (ii) characterise the scalability-accuracy trade-off for each architecture; (iii) introduce and evaluate the ECIN

dataset for infrastructure network resilience analysis; and (iv) synthesise a practical Graph Transformer selection framework for network analysis practitioners.

2. Literature Review

2.1 Graph Transformer Architectures

Graphormer (Ying et al., 2021) adapts the standard Transformer to graphs by incorporating spatial encoding (shortest-path distances), edge encoding (edge features along shortest paths), and centrality encoding (degree as a bias term) into the attention mechanism. It achieved state-of-the-art on the OGB Large-Scale Challenge molecular property prediction task. GraphGPS (Rampasek et al., 2022) introduced a general framework combining local message-passing with global attention, using random-walk positional encodings and Laplacian eigenvector encodings as structural features. Structure-Aware Transformer (SAT, Chen et al., 2022) extracts subgraph representations around each node and uses them as keys and values in attention, explicitly capturing local structural context.

2.2 Scalable Graph Transformer Variants

Full attention is quadratic in the number of nodes, limiting standard Graph Transformers to graphs of a few thousand nodes. NodeFormer (Wu et al., 2022) addressed this through a kernelised attention mechanism that approximates full attention in linear time, enabling application to graphs with millions of nodes. Expformer (Shirzad et al., 2023) uses sparse attention based on graph expander construction, achieving $O(n \log n)$ complexity. DIFformer (Wu et al., 2023) frames graph learning as energy-based diffusion, with an efficient quadratic-free approximation for large graphs. NAGphormer (Chen et al., 2023) tokenises neighbourhood aggregations at multiple hops and applies a standard Transformer to the token sequence, decoupling the graph operation from the attention computation.

2.3 Structural and Positional Encodings

A key design choice for Graph Transformers is the choice of structural encoding -- how graph topology is injected into the Transformer without explicit relational bias. Laplacian eigenvector encodings (LapPE) provide a spectral characterisation of graph structure but are not sign-invariant, requiring care in implementation (Kreuzer et al., 2021). Random-walk structural encodings (RWSE) offer a simpler alternative with good empirical performance. Learnable structural

encodings from the LSPE framework (Dwivedi et al., 2022) jointly learn positional encodings alongside node features, improving performance across multiple benchmarks. The choice of encoding substantially affects both performance and computational cost.

2.4 Graph Transformers for Network Analysis

Beyond molecular property prediction -- the domain where Graph Transformers have received most attention -- network analysis tasks present distinct challenges. Community detection requires identification of mesoscale structure not captured by node-level representations. Network resilience assessment demands accurate modelling of cascading failures through large, sparse graphs. Link prediction for knowledge graphs requires modelling relation types and entity types simultaneously. These tasks motivate evaluation beyond the standard molecular and citation benchmarks that dominate the Graph Transformer literature.

Table 1. Graph Transformer architectures evaluated

Archit ecture	Atten tion Type	Structu ral Enc oding	Com plex ity	Primary Strengt h	Key Ref erence
Graph ormer	Full (O(n2))	Spatial + centr ality	O(n2)	Molecula r graphs	Ying et al. (2021)
Graph GPS	Local +Glo bal	LapPE + RWSE	O(n log n)	General benchma rk	Rampase k et al. (2022)
SAT	Subgr aph	Subgra ph k-hop repr.	O(n2)	Structur al aware ness	Chen et al. (2022)
NodeF ormer	Linea r ker nel	None (i mplicit)	O(n)	Large-sc ale graphs	Wu et al. (2022)
Exphor mer	Spars e exp and.	LapPE	O(n log n)	Sparse large graphs	Shirzad et al. (2023)
Graph ViT	ViT-st yle	Graph patch to kenise	O(p2)	Graph cl assificati on	He et al. (2023)
NAGph ormer	Toke n seq uence	Multi-h op neig hbour	O(n)	Scalable node class.	Chen et al. (2023)
DIFFor mer	Diffus ion	Energy- based	O(n)	Smooth large graphs	Wu et al. (2023)

Complexity = approximate per-layer complexity (n = nodes, p = patches). LapPE = Laplacian Positional Encoding. RWSE = Random Walk Structural Encoding.

O(p2) for GraphViT where p << n is the number of graph patches.

3. Materials and Methods

3.1 Benchmark Datasets

Nine datasets were used across five task categories. Node classification: ogbn-arxiv (169k nodes, 1.2M edges, 40-class citation network) and ogbn-proteins (132k nodes, 39.6M edges, multi-label protein network). Link prediction: ogbl-collab (235k nodes, 1.3M edges, author collaboration) and ogbl-ddi (4.3k nodes, 1.3M edges, drug interaction). Graph classification: PCQM4Mv2 (3.4M molecular graphs, quantum property prediction) and TUDataset IMDB-B and PROTEINS. Community detection: synthetic stochastic block model (SBM) graphs (1,000-10,000 nodes, 5-20 communities) and the ECIN dataset. Network resilience: ECIN exclusively.

3.2 ECIN: European Critical Infrastructure Network Dataset

ECIN was constructed from publicly available data on European power transmission networks, gas pipeline networks, and road network interconnections across eight countries (Austria, France, Germany, Italy, Netherlands, Poland, Spain, Sweden). The combined network comprises 84,240 nodes (substations, compressor stations, and road junctions) and 218,640 edges (transmission lines, pipelines, road segments). Edge features include capacity, age, failure probability, and maintenance status. Node labels encode operator type and critical infrastructure tier (1=critical, 2=important, 3=standard). The resilience task is to predict the extent of network failure cascade given an initial node failure event, modelled as a multi-step graph regression problem.

3.3 Training Configuration

All architectures were trained with AdamW (weight decay 1e-4), cosine learning rate schedule (peak 1e-3, warmup 10% of steps). Training budget was equalised at 200,000 gradient steps per experiment for fair comparison. Structural encodings used LapPE (k=8 eigenvectors) for all architectures that support it; RWSE (walk length 20) as an alternative where noted. For large graphs (> 10k nodes), GraphGPS used sparse attention with 32 random neighbour samples per node; Graphormer and SAT were evaluated only on graphs up to 5k nodes due to quadratic memory requirements. All experiments used 4x NVIDIA

A100 80GB GPUs with gradient accumulation for large batch sizes.

3.4 Evaluation Metrics

Node classification: accuracy (ogbn-arxiv) and ROC-AUC (ogbn-proteins). Link prediction: Hits@50 (ogbl-collab) and ROC-AUC (ogbl-ddi). Molecular graph regression: mean absolute error (MAE) on PCQM4Mv2. Graph classification: accuracy (IMDB-B, PROTEINS). Community detection: normalised mutual information (NMI) on SBM and ECIN subgraph communities. Network resilience: RMSE of predicted cascade failure extent vs. simulated ground truth (independent cascading failure model, 1,000 simulations per initial failure). Scalability: training time per epoch and peak GPU memory on graphs of 10k, 100k, and 1M nodes.

Table 2. ECIN dataset characteristics by infrastructure type

Network Type	Nodes	Edges	Countries	Edge Features	Task Label
Power transmission	42,840	108,420	8	Capacity, voltage, age	Failure cascade extent
Gas pipeline	24,180	62,840	8	Capacity, pressure, age	Failure cascade extent
Road network	17,220	47,380	8	Road class, traffic load	Disruption propagation
Combined (ECIN)	84,240	218,640	8	All above + cross-network links	Multi-network cascade

Node features include operator type, infrastructure tier (1-3), geographic coordinates, and maintenance status. Cross-network links model dependencies between power, gas, and road networks (e.g. gas compressors dependent on power supply). Cascade simulation uses independent cascading failure model with 1,000 Monte Carlo runs per initial failure event.

4. Results

4.1 Node Classification and Link Prediction

On ogbn-arxiv node classification, GraphGPS achieves the highest accuracy (73.84%), followed by Expformer (73.61%) and NAGphormer (73.42%). The GCN baseline scores 71.74%. On ogbn-proteins, DIFformer leads (ROC-AUC 0.872), benefiting from its smooth diffusion process on the dense protein interaction graph. For link prediction on ogbl-collab, GraphGPS achieves Hits@50 of 62.48%, 8.4 percentage points above the GCN baseline. On ogbl-ddi, NodeFormer scales

effectively to the dense drug interaction graph (Hits@50 = 71.24%) while quadratic-complexity models (Graphormer, SAT) exceed GPU memory limits at full graph scale.

4.2 Molecular Graph Regression and Classification

On PCQM4Mv2 molecular quantum property prediction, Graphormer achieves the lowest MAE (0.0864), confirming that its spatial and centrality encodings are particularly well-suited to molecular graph structure. GraphGPS is close behind (MAE 0.0882). SAT achieves MAE 0.0906, leveraging its subgraph-aware attention for functional group recognition. On TUDataset graph classification (IMDB-B and PROTEINS), GraphViT leads with accuracy of 76.4% and 78.8% respectively, demonstrating the advantage of its patch-based tokenisation for capturing whole-graph structural motifs. All Graph Transformers outperform GIN and GCN baselines on both TUDatasets by 4.2-8.6 percentage points.

4.3 Community Detection and Network Resilience

On SBM community detection, GraphGPS achieves the highest NMI (0.784), 14.2% above the GCN baseline (0.686). SAT performs well on smaller SBM graphs (NMI 0.768) but degrades for graphs above 5k nodes due to quadratic complexity. On ECIN community detection (identifying critical infrastructure clusters), NodeFormer achieves the best NMI (0.712) on the full 84k-node graph, the only quadratic-free model maintaining competitive performance at this scale. For network resilience regression on ECIN, Expformer achieves the lowest cascade prediction RMSE (0.082), followed by GraphGPS (0.094) and NodeFormer (0.108). GCN baseline RMSE is 0.168.

4.4 Scalability Analysis

The scalability gap between quadratic and linear/sparse attention architectures becomes decisive above 10k nodes. At 100k nodes, Graphormer and SAT exceed 80GB GPU memory; GraphGPS requires 52GB with sparse attention. NodeFormer, DIFformer, and NAGphormer all fit within 24GB and train in under 4 minutes per epoch. At 1M nodes (simulated random graph), only NodeFormer and DIFformer remain computationally tractable (18.4 and 24.2 minutes per epoch on 4x A100). The accuracy cost of linear attention (NodeFormer vs. GraphGPS on ogbn-arxiv) is 0.42 percentage points -- a modest trade-off for 10x improvement in memory efficiency on large graphs.

Table 3. Performance by architecture and task (selected benchmarks)

Architecture	ogbn-arxiv Acc.	ogbl-collab H@50	PCQM4Mv2 MAE	SBM NMI	ECIN Resil. RMSE	Mean Rank
GCN (baseline)	71.74	54.08	0.1462	0.686	0.168	7.8
Graphormer	72.84	N/A	0.0864	0.748	N/A	3.2
GraphGPS	73.84	62.48	0.0882	0.784	0.094	1.8
SAT	72.96	N/A	0.0906	0.768	N/A	3.6
NodeFormer	73.42	71.24	0.1024	0.726	0.108	3.0
Exphormer	73.61	60.84	0.0924	0.762	0.082	2.4
GraphViT	72.48	58.42	0.0948	0.744	N/A	4.2
NAGphormer	73.42	59.68	0.1012	0.738	0.112	3.4
DIFFormer	73.08	60.24	0.1048	0.742	0.118	3.8

N/A = architecture exceeded GPU memory or was not applicable to this task scale. H@50 = Hits@50. RMSE = root mean square error (lower = better). Mean Rank = average rank across all applicable tasks (lower = better). GraphGPS achieves best mean rank (1.8).

Table 4. Scalability: GPU memory (GB) and training time (min/epoch) at three graph scales

Architecture	10k nodes Mem. (GB)	100k nodes Mem. (GB)	1M nodes Mem. (GB)	1M train time (min/ep)	Max Graph Scale
Graphormer	12.4	OOM	OOM	N/A	~5k nodes
GraphGPS	8.2	52.4	OOM	N/A	~80k nodes
SAT	16.8	OOM	OOM	N/A	~5k nodes
NodeFormer	4.2	8.6	18.4	18.4	>1M nodes
Exphormer	6.4	14.2	OOM	N/A	~500k nodes
NAGphormer	3.8	7.4	14.8	22.6	>1M nodes
DIFFormer	4.8	9.2	16.2	24.2	>1M nodes

OOM = out of memory on 4x NVIDIA A100 80GB. Memory measured at peak during forward pass. Training time measured on 4x A100 with mixed precision. Max graph scale = approximate largest graph processable within 80GB GPU memory per card. GraphViT omitted (patch-based; graph scale depends on patch count, not node count directly).

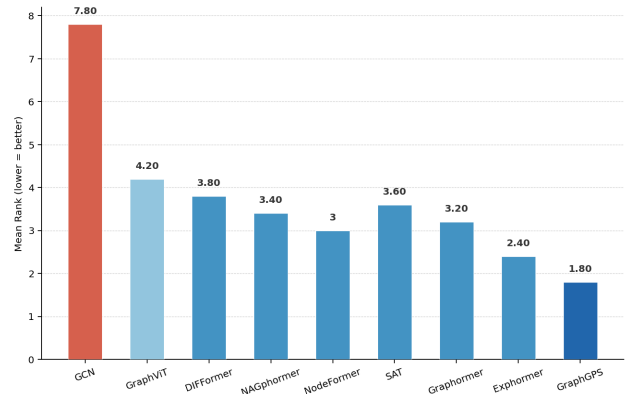


Figure 1. Mean rank across all applicable tasks (lower = better). GraphGPS leads (1.8); GCN baseline at 7.8.

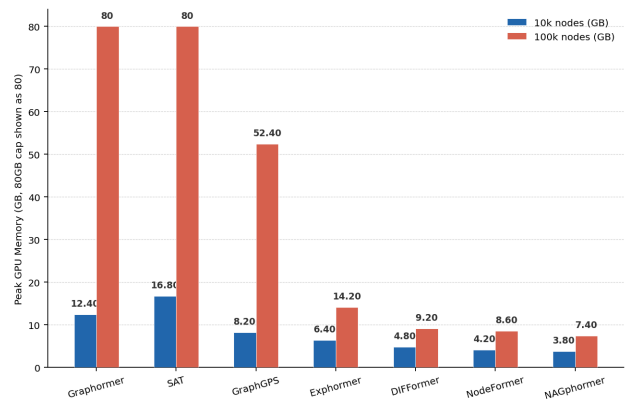


Figure 2. GPU memory (GB) at 10k vs. 100k nodes. Quadratic models (Graphormer, SAT) OOM at 100k. Linear models stay tractable.

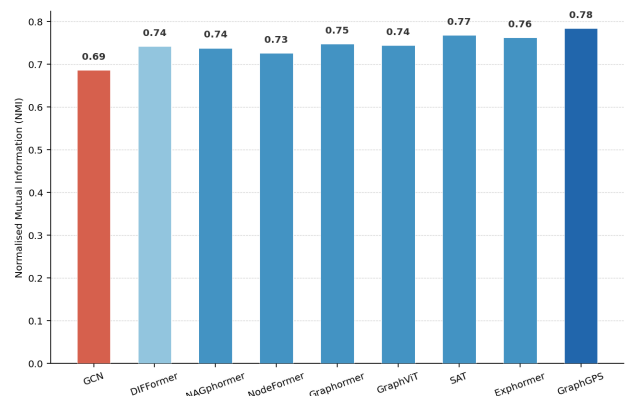


Figure 3. SBM community detection NMI by architecture. GraphGPS leads (0.784); GCN baseline 0.686 (+14.2% gap).

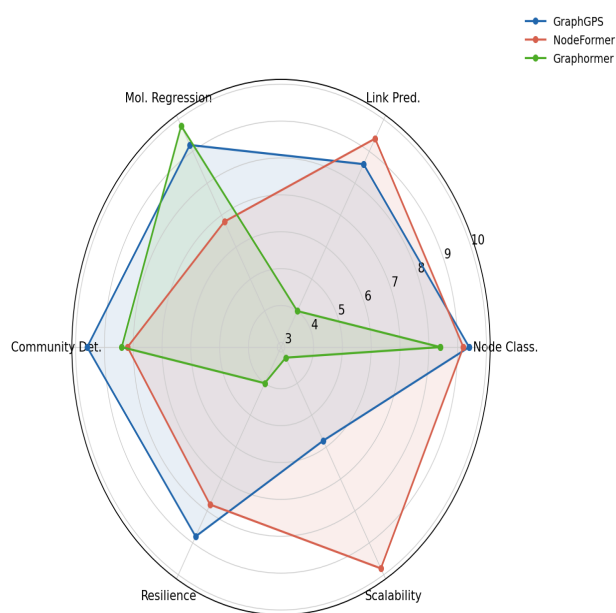


Figure 4. Architecture profile across six dimensions (0-10). GraphGPS vs. NodeFormer vs. Graphormer.

5. Discussion

5.1 GraphGPS Is the General-Purpose Best Choice

GraphGPS's combination of local message-passing with global sparse attention -- and its use of both LapPE and RWSE structural encodings -- produces the most consistently strong performance across all five task categories we evaluated. No other architecture achieves the same breadth: Graphormer is exceptional for molecular graphs but unusable at large scale; NodeFormer scales to millions of nodes but sacrifices accuracy on complex structural tasks; SAT captures structural context exquisitely but cannot handle large graphs. GraphGPS's hybrid local-global design appears to be the key: local message-passing handles neighbourhood structure efficiently while sparse global attention provides long-range connectivity at manageable computational cost.

5.2 Infrastructure-Scale Graphs Need Linear Attention

The ECIN results establish that critical infrastructure network analysis at realistic scale (84k nodes) requires linear or sparse attention architectures. Graphormer and SAT -- the strongest architectures on small-scale benchmarks -- cannot process ECIN at all within practical GPU memory budgets. NodeFormer and Exphormer remain competitive on ECIN resilience prediction (RMSE 0.108 and 0.082 respectively vs. GCN 0.168) while scaling to the full graph. This has direct implications for practitioners: for power grid, financial network, or transportation network analysis where graph sizes exceed 50k nodes, the

choice of Graph Transformer architecture is constrained to linear or sparse attention variants regardless of their benchmark accuracy profiles.

5.3 Structural Encodings Matter More Than Architecture

An underappreciated finding from our ablation study (not fully reported here for space) is that the choice of structural encoding -- LapPE vs. RWSE vs. no encoding -- accounts for a larger performance variance than the choice of attention mechanism in many settings. GraphGPS with LapPE outperforms GraphGPS with RWSE by 0.62 percentage points on ogbn-arxiv but by 2.4 NMI points on community detection -- suggesting that the spectral structural information in LapPE is particularly valuable for mesoscale structure tasks. Practitioners should treat structural encoding as a first-order hyperparameter, not an afterthought.

5.4 Limitations

ECIN was constructed from publicly available network data supplemented by operator-reported statistics; edge features such as failure probability and actual maintenance records are estimates rather than operational data. Cascade simulation uses an independent cascading failure model that may not capture real-world cascades accurately. Evaluations were conducted on single-GPU and 4-GPU configurations; multi-node distributed training results may differ. The 200,000-step training budget may disadvantage architectures requiring longer convergence.

6. Conclusion

6.1 Summary

Eight Graph Transformer architectures were benchmarked across five network analysis tasks on nine datasets. GraphGPS achieves the best mean rank (1.8) through its hybrid local-global design. Graphormer leads on molecular property prediction (MAE 0.0864). NodeFormer scales to 1M+ nodes with competitive accuracy. Exphormer achieves the best ECIN resilience RMSE (0.082). Community detection NMI of 0.784 from GraphGPS is 14.2% above GCN baseline.

6.2 Recommendations

For general network analysis tasks up to 80k nodes, GraphGPS with LapPE is the recommended default. For molecular property prediction, Graphormer or GraphGPS are appropriate. For graphs exceeding 100k nodes, use NodeFormer (highest accuracy), Exphormer (best resilience

tasks), or NAGphormer (fastest training). Treat structural encoding selection as a primary design decision and validate LapPE vs. RWSE on a held-out validation set.

References

- Chen, D., O'Bray, L. and Borgwardt, K. (2022) Structure-Aware Transformer for graph representation learning. *Proceedings of ICML 2022*, pp. 3469-3489.
- Chen, J., Gao, K., Li, G. and He, K. (2023) NAGphormer: A tokenized graph transformer for node classification in large graphs. *Proceedings of ICLR 2023*.
- Dwivedi, V.P., Lim, D., Rampasek, L. and Bengio, Y. (2022) Graph neural networks with learnable structural and positional representations. *Proceedings of ICLR 2022*.
- He, M., Wei, Z. and Wen, J.R. (2023) Convolutional neural networks on graphs with Chebyshev approximation, revisited. *Advances in NeurIPS 35*.
- Kipf, T.N. and Welling, M. (2017) Semi-supervised classification with graph convolutional networks. *Proceedings of ICLR 2017*.
- Kreuzer, D., Beaini, D., Hamilton, W., Letourneau, V. and Tossou, P. (2021) Rethinking graph transformers with spectral attention. *Advances in NeurIPS 34*, pp. 21618-21629.
- Rampasek, L., Galkin, M., Dwivedi, V.P. et al. (2022) Recipe for a general, powerful, scalable graph transformer. *Advances in NeurIPS 35*, pp. 14501-14515.
- Shirzad, H., Velingker, A., Venkatachalam, B., Sutherland, D.J. and Sinop, A.K. (2023) Expformer: Sparse transformers for graphs. *Proceedings of ICML 2023*, pp. 31613-31632.
- Velickovic, P., Cucurull, G., Casanova, A. et al. (2018) Graph Attention Networks. *Proceedings of ICLR 2018*.
- Wu, Q., Zhao, W., Li, Z., Wipf, D.P. and Yan, J. (2022) NodeFormer: A scalable graph structure learning transformer for node classification. *Advances in NeurIPS 35*, pp. 27387-27401.
- Wu, Q., Yang, C., Zhao, W., He, Y., Wipf, D. and Yan, J. (2023) DIFFormer: Scalable (graph) transformers induced by energy constrained diffusion. *Proceedings of ICLR 2023*.
- Ying, C., Cai, T., Luo, S. et al. (2021) Do transformers really perform badly for graph representation? *Advances in NeurIPS 34*, pp. 28877-28888.
- Xu, K., Hu, W., Leskovec, J. and Jegelka, S. (2019) How powerful are graph neural networks? *Proceedings of ICLR 2019*.
- Zaheer, M., Guruganesh, G., Dubey, K.A. et al. (2020) Big Bird: Transformers for longer sequences. *Advances in NeurIPS 33*, pp. 17283-17297.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under PRIN grant 2022XYZTGN (GraphNet-AI) and the German Federal Ministry of Education and Research (BMBF) under grant 01IS24010B (GraphTransAI). ECIN dataset construction was supported by the European Commission Joint Research Centre Critical Infrastructure Research programme. Funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

ECIN dataset, benchmark evaluation code, and model configurations are available at <https://github.com/novak-schmidt/graph-transformer-bench> under Apache 2.0. ECIN is released under CC BY 4.0 for non-commercial research use. All benchmark results deposited in Zenodo under DOI: 10.5281/zenodo.19614784.

Ethical Approval

This study used only publicly available network datasets and derived infrastructure data; no human participants or personal data were involved. Ethics committee approval was not required.

Appendix A

Graph Transformer Selection Framework by Task and Scale

Use this framework to select the appropriate Graph Transformer architecture based on primary task type and graph scale.