GraphGPT: Generative Pre-trained Graph Eulerian Transformer

Qifang Zhao, Weidong Ren, Tianyu Li, Hong Liu, Xingsheng He, Xiaoxiao Xu @ Alibaba Inc.

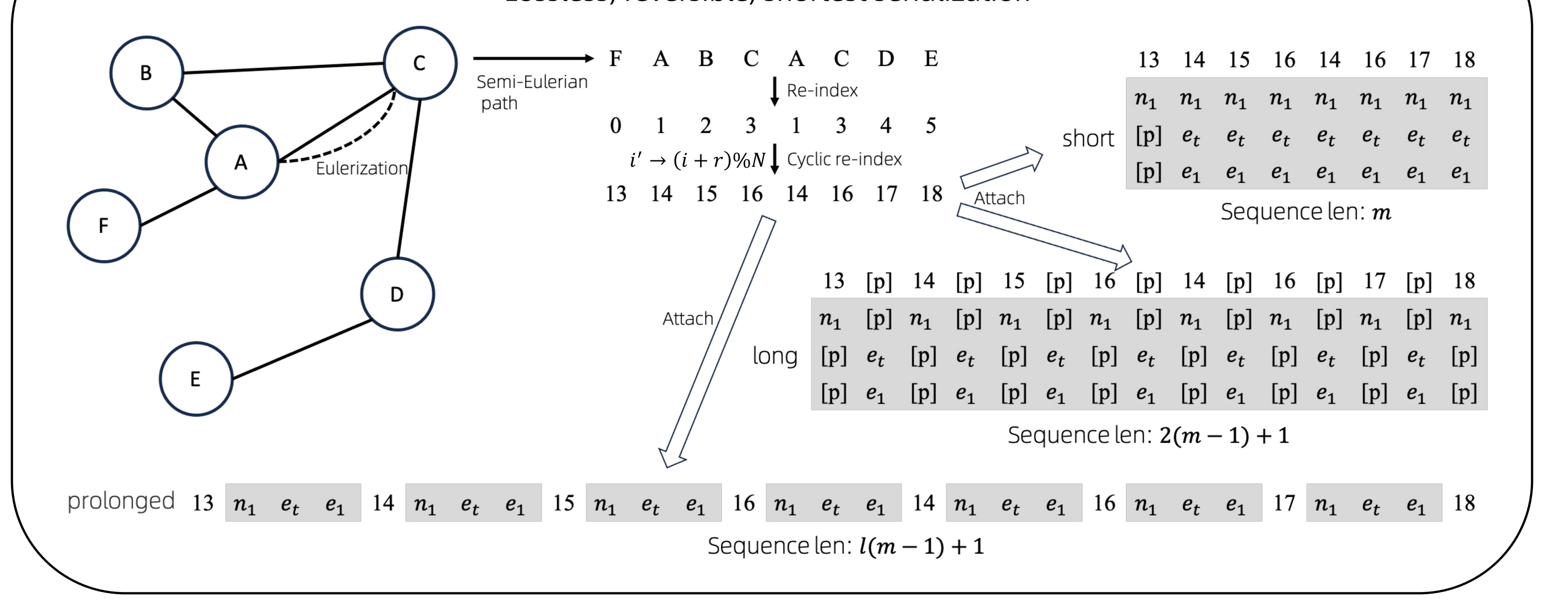
Motivation

- Graph has not benefited from the transformer architecture, like NLP/CV/Audio
- Unifying graph with other modalities is problematic due to inconsistent architecture
- Graph has not benefited from scaling up model sizes

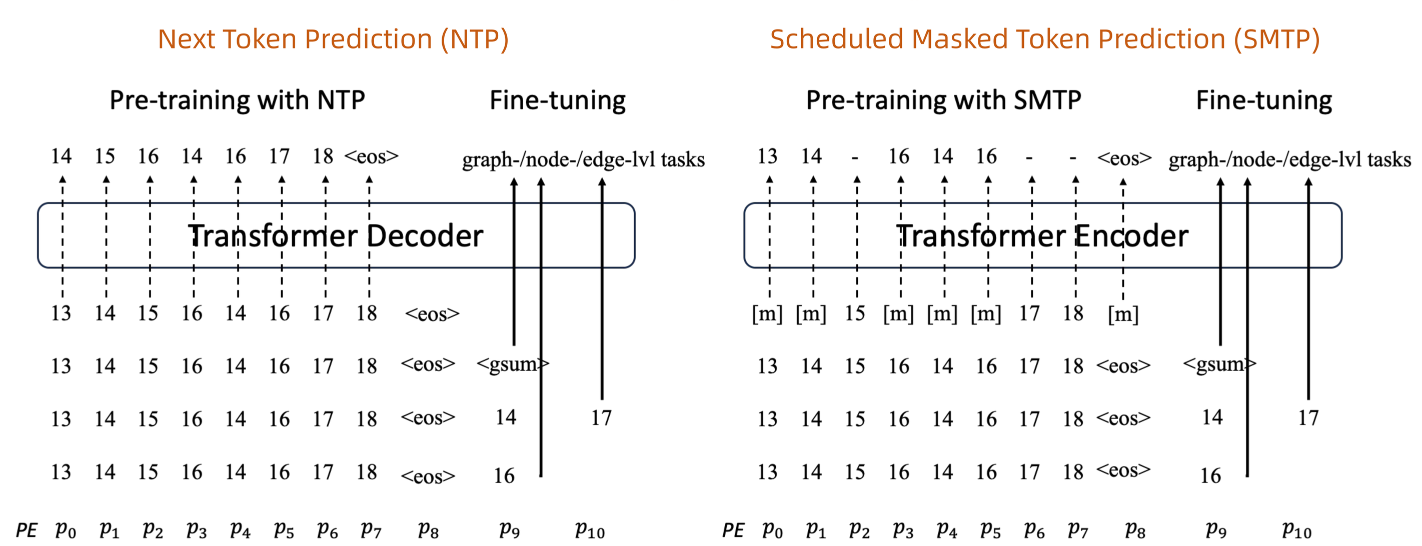
Method

Graph => Eulerian path => Sequence of tokens w/o attributes

Lossless, reversible, shortest serialization



Pre-training & Fine-tuning



Graph-level Task

- *PCQM4Mv2* contains > 3.7 million organic molecules from PubChemQC (Nakata & Shimazaki, 2017). Nodes represent atoms (9D attributes: atomic number, chirality, etc.), and edges denote chemical bonds (3D attributes: bond type, stereochemistry, conjugation).
- *ogbg-molpcba* is a smaller molecular dataset (Wu et al., 2017) with the same node/edge attributes.

PCQM4M-v2: 3.7M molecules

datasets	# of graphs	avg # of nodes	avg # of edges	task-type	metrics
PCQM4Mv2	3,746,619	14.14	14.56	regression	MAE
ogbg-molpcba	437,929	26.0	28.1	multi-label BC	AP

Table 1. Results of the graph regression task on the PCQM4Mv2 dataset. The metric is mean absolute error (MAE), the smaller the better. 86% of the valid dataset is added to training after hyper-parameters selection. The best results are in bold, and second-best are underlined.

Models	MAE ↓ Valid	Test	Params
GCN[1]	0.1379	0.1398	2.0M
GIN[2]	0.1195	0.1218	3.8M
GCN[1]-VN[3]	0.1153	0.1152	4.9M
GIN[2]-VN[3]	0.1083	0.1084	6.7M
TokenGT[4]	0.0910	0.0919	48.5M
Graphformer[5]	0.0864	N/A	48.3M
GPS-Deep[6]	0.0852	0.0862	138.1M
GPS++ (no 3D)[7]	0.0818	N/A	40.0M
GPTrans-L[8]	0.0809	0.0821	86.0M
GraphGPT-B_{12}	0.0807	N/A	113.6M
GraphGPT-B_{24}	**0.0793**	N/A	227.3M
GraphGPT-B_{48}	**0.0792**	**0.0804**	453.4M

[1]Kipf & Welling (2017), [2]Xu et al. (2019), [3]Gilmer et al. (2017), [4]Kim et al. (2022), [5]Ying et al. (2021), [6]Rampásek et al. (2022), [7]Masters et al. (2023), [8]Chen et al. (2023b)

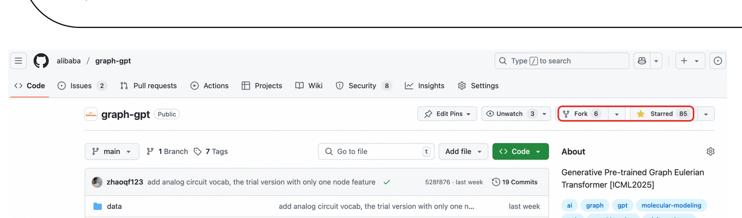
OGBG-MOLPCBA: 438K molecules

Table 2. Results of the graph classification task on the ogbg-molpcba dataset. All the baseline results are from the OGB leader-board or the corresponding papers. † indicates the model is pre-trained on PCQM4M-v2 dataset.

Models	Average Precision (%) ↑ Test	Valid	Params
GCN[1]	$20.20_{\pm0.24}$	$20.59_{\pm0.33}$	0.57M
GIN[2]	$22.66_{\pm0.28}$	$23.05_{\pm0.27}$	1.92M
GINE[3]-VN[4]	$29.17_{\pm0.15}$	$30.65_{\pm0.30}$	6.1M
NGIN[5]-VN[4]	$30.07_{\pm0.37}$	$30.59_{\pm0.56}$	44.19M
PDF[6]	$30.31_{\pm0.26}$	$31.15_{\pm0.20}$	3.84M
Graphormer-L[†7]	$31.40_{\pm0.32}$	$32.27_{\pm0.24}$	119.5M
EGT-Larger[†8]	$29.61_{\pm0.24}$	N/A	110.8M
GRPE-Large[†9]	$31.50_{\pm0.10}$	N/A	118.3M
GPTrans-L[†10]	$\mathbf{32.43}_{\pm0.22}$	N/A	86.0M
GraphGPT-M[†]	$30.13_{\pm0.25}$	$31.62_{\pm0.24}$	37.7M
GraphGPT-$B^{\dagger}_{12}$	$31.28_{\pm0.23}$	$32.27_{\pm0.15}$	113.6M
GraphGPT-$B^{\dagger}_{24}$	$31.81_{\pm0.1}$	$\mathbf{32.54}_{\pm0.2}$	227.3M

[1]Kipf & Welling (2017), [2]Xu et al. (2019), [3]Brossard et al. (2020), [4]Gilmer et al. (2017), [5]Zhang & Li (2021), [6]Yang et al. (2023), [7]Ying et al. (2021), [8]Hussain et al. (2022), [9]Park et al. (2022), [10]Chen et al. (2023b)

- *SOTA*: On PCQM4Mv2, GraphGPT achieves a test MAE of 0.0804, significantly outperforming the previous SOTA (*0.0821*, GPTrans)
- *vs GTs*: GTs like TokenGT, Graphformer, GPS, and GPTrans requires handcrafted features or intricate architectures to encode structural information, while GraphGPT attains superior performance without manual feature engineering.
- *vs GNNs*: GraphGPT surpasses GNNs by a substantial margin.
- *Parameter Efficiency*: GraphGPT's larger parameter count may reflect its capacity to implicitly learn features that other GTs encode manually. Generative pre-training also allocates model capacity to generation tasks, potentially limiting discriminative performance of models at smaller scales.



arXiv
https://arxiv.org › cs

Generative Pre-trained Graph Eulerian Transformer

by Q Zhao · 2023 · Cited by 16 — We introduceGraphGPT, a novel self-supervised generative pre-trained model for graph learning based on the Graph Eulerian Transformer (GET).

Scaling up model sizes

consistent improvement up to 200M parameters

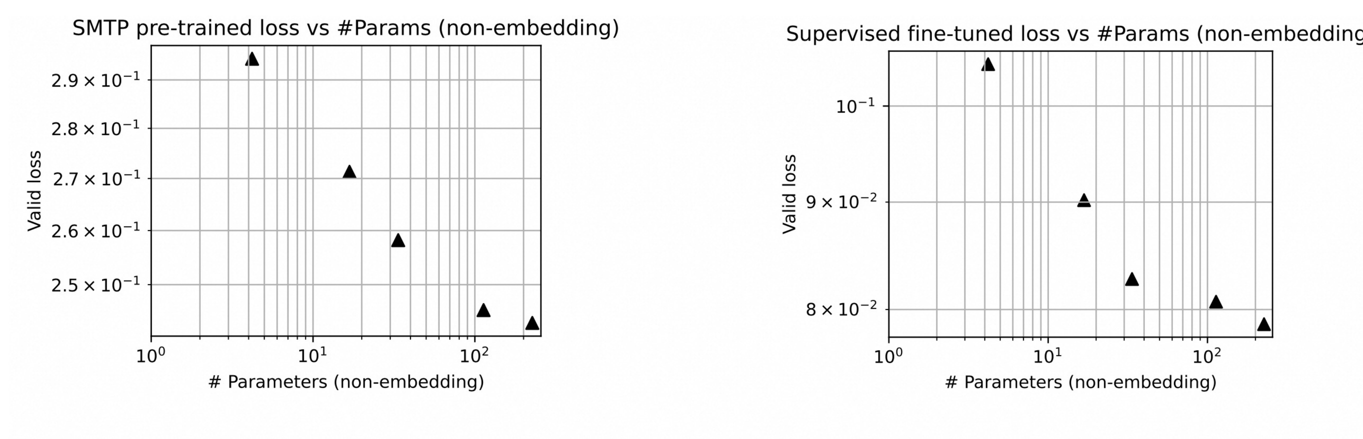


Figure 3. Log-log plot of pre-training loss and supervised fine-tuning loss versus the number of non-embedding parameters for the Mini/Small/Medium/Base/Base24 model configurations (see Table 11) on the PCQM4M-v2 dataset.

Graph Structure Understanding (GSU)

Dataset stats: totally 45000 graphs

	train	valid	test-small	test-large
indices	[0, 30000)	[30000, 35000)	[35000, 40000)	[40000, 45000)
connected graphs	63.13%	62.78%	62.76%	13.78%
avg # node	15.58	15.54%	15.61	63.08
min # node	4	4	4	26
max # node	25	25	25	100

- Pre-training (PT) is highly beneficial: 32% –> 92% ✦
- PT on other datasets also improves GSU on this dataset, sometimes even better: a vs $b/c/d$ ▷
  - This holds true even when PT includes node and edge attribute prediction: a vs $e/f/g$ ◀
- PT on real graphs outperforms random graphs: c vs d, $(a+c)$ vs $(a+b+d)$
- More data & diverse PT enhance generalization: a vs $(a+b)/(a+c)/(b+c)$ vs $(a+b+c)/(a+b+d)$ ◁

Graph triangle counting: 1~10 (10-class classification)

Models	Accuracy (%) ↑ T-small	T-large	Params
GIN[1]	$71.53_{\pm0.94}$	$33.54_{\pm0.30}$	0.15M
Transformer[2]	$12.08_{\pm0.31}$	$10.01_{\pm0.04}$	0.2M
Transformer-LapPE[3]	$78.29_{\pm0.25}$	$10.64_{\pm2.94}$	0.2M
Transformer-RWSE[3]	$\mathbf{99.40}_{\pm0.10}$	$54.76_{\pm7.24}$	0.2M
Graphormer[4]	$\mathbf{99.09}_{\pm0.31}$	$42.34_{\pm6.48}$	0.2M
✦ GET-B	$32.60_{\pm1.86}$	$13.99_{\pm1.78}$	113.5M
▷ ✦ GraphGPT-B^{a} ◀ ◁	$92.16_{\pm0.28}$	$26.51_{\pm1.01}$	113.5M
▷ GraphGPT-B^{b}	$81.38_{\pm0.27}$	$37.68_{\pm0.99}$	113.5M
▷ GraphGPT-B^{c}	$\mathbf{99.08}_{\pm0.14}$	$38.80_{\pm3.60}$	113.5M
▷ GraphGPT-B^{d}	$90.93_{\pm0.51}$	$40.79_{\pm1.40}$	113.5M
GraphGPT-B^{e} ◀	$64.28_{\pm0.33}$	$17.38_{\pm0.61}$	113.5M
GraphGPT-B^{f} ◀	$86.14_{\pm7.38}$	$26.94_{\pm4.80}$	113.5M
GraphGPT-B^{g} ◀	$86.57_{\pm2.74}$	$23.45_{\pm1.44}$	113.5M
GraphGPT-B^{a+b} ◁	$84.83_{\pm0.81}$	$39.62_{\pm1.84}$	113.5M
GraphGPT-B^{a+c} ◁	$98.68_{\pm0.18}$	$50.07_{\pm3.28}$	113.5M
GraphGPT-B^{b+c} ◁	$98.26_{\pm0.30}$	$52.33_{\pm2.61}$	113.5M
GraphGPT-B^{a+b+d} ◁	$89.98_{\pm0.54}$	$33.45_{\pm2.51}$	113.5M
GraphGPT-M^{a+b+c}	$95.07_{\pm0.67}$	$51.72_{\pm1.12}$	33.7M
GraphGPT-B^{a+b+c} ◁	$98.63_{\pm0.18}$	$\mathbf{58.96}_{\pm1.90}$	113.5M

[1]Xu et al. (2019), [2]Vaswani et al. (2017), [3]Rampásek et al. (2022), [4]Ying et al. (2021)

Pre-trained with: [a]Triangles (45K), [b]Reddit-threads (0.22M), [c]Internal dataset (3.1M), [d]Random graphs (3.1M), [e]PCQM4M-v2 (3.7M), [f]ogbl-ppa (1), [g]ogbn-proteins (1).

Edge-level Task

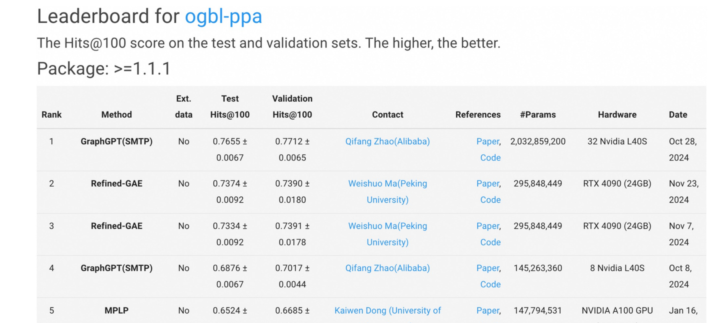
datasets	# of graphs	avg # of nodes	avg # of edges	task-type	metrics
PCQM4Mv2	3,746,619	14.14	14.56	regression	MAE
ogbg-molpcba	437,929	26.0	28.1	multi-label BC	AP
reddit-threads	203,088	23.9	24.9	BC	ROC-AUC
Triangles	45,000	20.9	32.7	multi-class classification	ACC
Internal dataset	3,100,000	24.8	54.7	N/A	N/A
Random Graph$_{p=0.05}$	3,100,000	67.0	124.7	N/A	N/A
Random Graph$_{p=0.03}$	3,100,000	67.1	74.8	N/A	N/A
Random Graph$_{p=0.01}$	3,100,000	67.1	25.0	N/A	N/A
ogbl-ppa	1	576,289	30,326,273	BC	HR@100
ogbl-citation2	1	2,927,963	30,561,187	BC	MRR
ogbn-proteins	1	132,534	39,561,252	multi-label BC	ROC-AUC
ogbn-arxiv	1	169,343	1,166,243	multi-class classification	ACC

Table 4. Results of the link prediction task on the ogbl-ppa and ogbl-citation2 datasets.

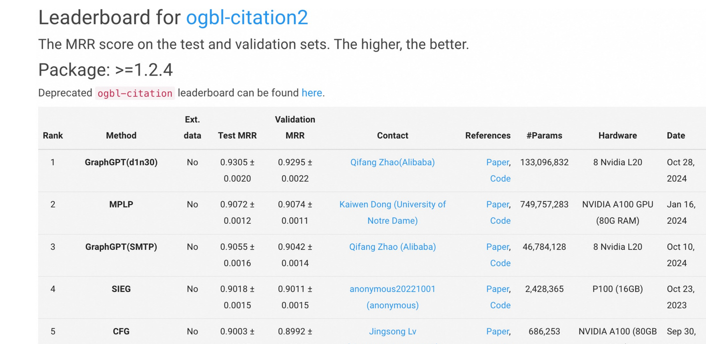
Models	ogbl-ppa HR@100 (%) ↑	ogbl-citation2 MRR (%) ↑
Common Neighbor	$27.65_{\pm0.00}$	$51.47_{\pm0.00}$
Adamic Adar	$32.45_{\pm0.00}$	$51.89_{\pm0.00}$
Resource Allocation[1]	$49.33_{\pm0.00}$	$51.98_{\pm0.00}$
Node2Vec[2]	$22.26_{\pm0.83}$	$61.41_{\pm0.11}$
Matrix Factorization[3]	$32.29_{\pm0.94}$	$51.86_{\pm4.43}$
GCN[4]	$18.67_{\pm1.32}$	$84.74_{\pm0.21}$
GraphSAGE[5]	$16.55_{\pm2.40}$	$82.60_{\pm0.36}$
SEAL[6]	$48.80_{\pm3.16}$	$87.67_{\pm0.32}$
AGDN[7]	$41.23_{\pm1.59}$	$85.49_{\pm0.29}$
SIEG[8]	$63.22_{\pm1.74}$	$90.18_{\pm0.15}$
MPLP[9]	$65.24_{\pm1.50}$	$90.72_{\pm0.12}$
RefinedGAE[10]	$73.74_{\pm0.92}$	$84.55_{\pm0.15}$
GraphGPT-M	$65.44_{\pm0.43}$	$92.82_{\pm0.27}$
GraphGPT-B	$68.76_{\pm0.67}$	$\mathbf{93.05}_{\pm0.20}$
GraphGPT-XXL	$\mathbf{76.55}_{\pm0.67}$	N/A

[1]Zhou et al. (2009), [2]Grover & Leskovec (2016), [3]Mnih & Salakhutdinov (2008), [4]Kipf & Welling (2017), [5]Hamilton et al. (2017), [6]Zhang et al. (2021), [7]Sun et al. (2020), [8]Shi et al. (2024), [9]Dong et al. (2023), [10]Ma et al. (2024)

Leaderboard for ogbl-ppa



Leaderboard for ogbl-citation2



- *Performance Superiority*: GraphGPT significantly outperforms all baseline methods, including GNNs, heuristic methods, and latent-factor approaches, across both datasets.
- *Scalability*: GraphGPT scales seamlessly to 2 billion parameters, achieving sustained performance gains with increasing model size.
- *Transformer Efficacy*: To our knowledge, GraphGPT is the first transformer-based model to achieve SOTA results on ogbl-ppa and ogbl-citation2, demonstrating the viability of sequence-driven architectures for large-scale edge-level tasks.

Node-level Task

Table 5. Results of the node classification task on the ogbn-proteins and ogbn-arxiv datasets.

Models	ogbn-proteins ROC-AUC (%) ↑	ogbn-arxiv Accuracy (%) ↑
GCN[1,2]	$77.29_{\pm0.46}$	$73.53_{\pm0.12}$
GraphSAGE[1,3]	$82.21_{\pm0.32}$	$73.00_{\pm0.28}$
GAT[1,4]	$85.01_{\pm0.46}$	$73.30_{\pm0.18}$
DRGAT[5]	N/A	$74.16_{\pm0.07}$
AGDN[6]	$88.65_{\pm0.13}$	$73.41_{\pm0.25}$
DeeperGCN[7]	$85.80_{\pm0.17}$	$71.92_{\pm0.16}$
GraphGPS[1,8]	$77.15_{\pm0.64}$	$71.23_{\pm0.59}$
NAGphormer[1,9]	$72.17_{\pm0.45}$	$70.88_{\pm0.24}$
Exphormer[1,10]	$77.62_{\pm0.33}$	$72.32_{\pm0.36}$
GOAT[1,11]	$79.31_{\pm0.42}$	$72.76_{\pm0.29}$
NodeFormer[1,12]	$77.86_{\pm0.84}$	$67.78_{\pm0.28}$
SGFormer[1,13]	$79.92_{\pm0.48}$	$72.76_{\pm0.33}$
Polynormer[1,14]	$79.53_{\pm0.67}$	$73.40_{\pm0.22}$
GraphGPT-S	$83.4_{\pm0.00}$	$71.2_{\pm0.00}$
GraphGPT-M	$84.3_{\pm0.00}$	$71.8_{\pm0.00}$
GraphGPT-B	$85.5_{\pm0.00}$	$72.2_{\pm0.00}$

[1]Luo et al. (2024), [2]Kipf & Welling (2017), [3]Hamilton et al. (2017), [4]Vaswani et al. (2017), [5]Zhang et al. (2023), [6]Sun et al. (2020), [7]Li et al. (2020), [8]Rampásek et al. (2022), [9]Chen et al. (2023a), [10]Shirzad et al. (2023), [11]Kong et al. (2023), [12]Wu et al. (2022), [13]Wu et al. (2024), [14]Deng et al. (2024)

- Ogbn-proteins: Undirected, weighted graph of 132,534 proteins (nodes) with 8D edge attributes encoding association strengths.
- Ogbn-arxiv: Citation network of 169,343 papers; tasks involve predicting 40 subject categories.

Github, paper, ckps

- GraphGPT outperforms/matches classic GNNs.
  - But still lags behind some customized GNN variants.
- It significantly improves or equals GTs.

Various Model Sizes

Model-size	Hidden-size	# of layers	# of heads	Params (excluding embed)
Mini	256	4	4	4.2M
S (Small)	512	4	8	16.8M
M (Medium)	512	8	8	33.6M
B / B_{12} (Base)	768	12	12	113.2M
B_{24} (Base24)	768	24	12	226.5M
B_{48} (Base48)	768	48	12	453.0M
L (Large)	1024	24	16	402.7M
XXL (XXLarge)	1600	48	25	2.0B

- Statistics of GraphGPT models of different sizes. The GraphGPT-Base is of the same scale as Bert-Base (Devlin et al., 2019).

Ablation

Pre-training

Table 6. Ablation study of pre-training on the datasets of various types of tasks. ∗ means both molpcba and PCQM4Mv2 datasets are used for SMTP pre-training, and † indicates that the model is further trained using PCQM4M-v2's regression task. For the PCQM4Mv2 dataset, the metric is MAE, the lower the better.

Datasets	Pre-training	Test	Valid
PCQM4Mv2	✗	N/A	0.1086
	NTP	N/A	0.0875
	SMTP	N/A	**0.086**
OGBG-MOLPCBA	✗	12.8	13.31
	NTP	23.85	27.77
	SMTP	27.56	28.74
	SMTP*	27.2	28.49
	SMTP* + FT†	**28.07**	**29.01**
OGBL-PPA	✗	41.28	40.14
	NTP	55.56	54.87
	SMTP	**55.68**	**54.93**
OGBN-PROTEINS	✗	57.52	61.19
	NTP	75.61	80.47
	SMTP	**83.02**	**86.41**

- Pre-training brings substantial improvements.
- SMTP > NTP in most cases.
- Strong in-domain transferability.

Github, paper, ckps

Node-reindex

Table 7. Ablation study of node re-indexing on the ogbg-molpcba dataset with two model sizes. PT means pre-training with NTP.

Params	Re-index	PT Loss	Test	Valid
4.48M	✗	**0.0844**	0.2310	0.2525
	✓	0.0874	**0.2385**	**0.2777**
114.12M	✗	**0.0689**	0.2270	0.2621
	✓	0.0750	**0.2517**	**0.2857**

Node-Identity Coding

Table 8. Ablation study of node identity encoding on the ogbl-ppa and ogbn-proteins datasets using NTP pre-training. NIE stands for Node identity encoding.

Datasets	Params	NIE	Test	Valid
OGBL-PPA	14.75M	✗	44.38	45.08
		✓	**55.56**	**54.87**
OGBN-PROTEINS	10.76M	✗	60.22	65.66
		✓	**75.61**	**80.47**

Limitations

- Model size is large, high computational resource is required
- A larger graph dataset is required to demonstrate superiority.
- Transferability: Pre-training is currently limited to same-domain datasets, making generalization to other graph data domains challenging.
  - The transferability of graph structure understanding is evident.

Outlook

- General Graph Structure Understanding [Graph] Foundation Model (GFM)
- Specialized Domain Understanding GFM (e.g., molecule)
- Combined with LLM, similar to Llava

Synergy with diffusion LLM (dLLM)

- *Speed and Performance*: dLLM has shown superior generation speed and comparable performance compared to AR (autoregressive) LLM: Mercury and Gemini Diffusion.
- *Same Pre-training*: Masked dLLM like LLaDa-8B, Dream-7B share almost the same pre-training objectives as SMTP employed by GraphGPT.
- *SMTP > NTP*: GraphGPT shows dLLM-like pre-training SMTP is much better than AR-like pre-training NTP in most graph datasets.
- *Multi-modality*: GraphGPT processes graph data in a way closely aligned with dLLM: using sequential tokens, a transformer encoder, and a masked token prediction objective. It implies graph data can be naturally incorporated in the dLLM.
- *AI for Science*: Some scientific data is naturally represented as graphs—for example, molecules and integrated circuits. Other scientific data, such as proteins and DNA/RNA, is represented as sequences. Unlike language, these data types lack the autoregressive (AR) inductive bias, making them better suited for modeling with dLLM.

References

[1] Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, Aditya Grover, Volodymyr Kuleshov. Mercury: Ultra-Fast Language Models Based on Diffusion. arXiv preprint arXiv:2506.17298, 2025.

[2] Google DeepMind, Gemini Diffusion. url: https://deepmind.google/models/gemini-diffusion/.

[3] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, Jure Leskovec. Open Graph Benchmark: Datasets for Machine Learning on Graphs. arXiv preprint arXiv: 2005.00687, 2020.

[4] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, Chongxuan Li. Large Language Diffusion Models. arXiv preprint arXiv: 2502.09992, 2025. (LLaDA-8B)

[5] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7B. url: https://hkunlp.github.io/blog/2025/dream/