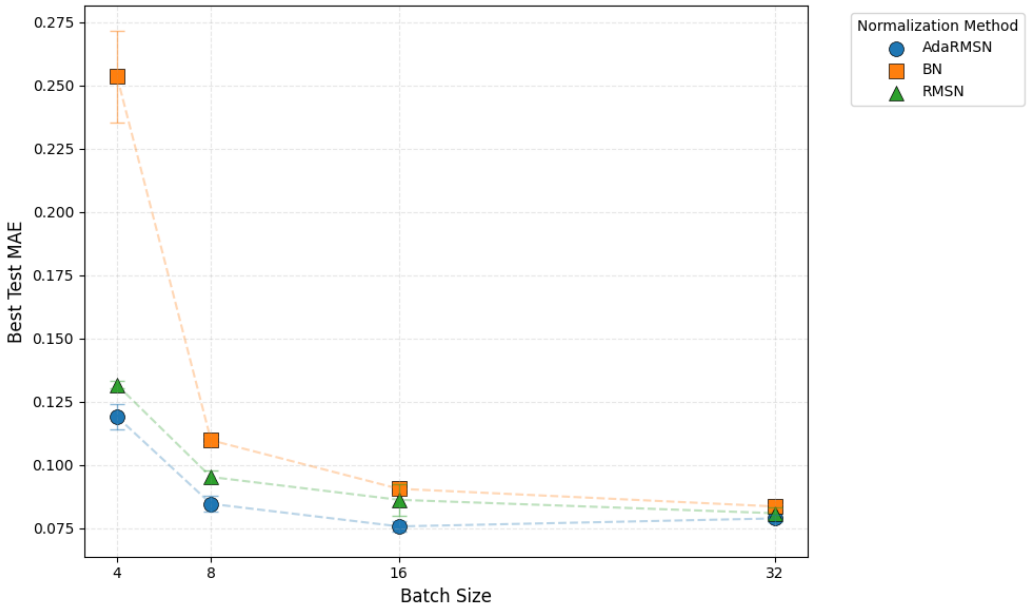

Figure 1. SGT on ZINC: Test MAE v.s. Batch Size (BS). # Training epochs are adjusted per batch-size for the same total update steps: $400 * \text{BS}/32$. The first 10% epochs are in the warmup stage. AdaRMSN and RMSN demonstrate better stability and less sensitivity to varying batch-sizes compared to BN.

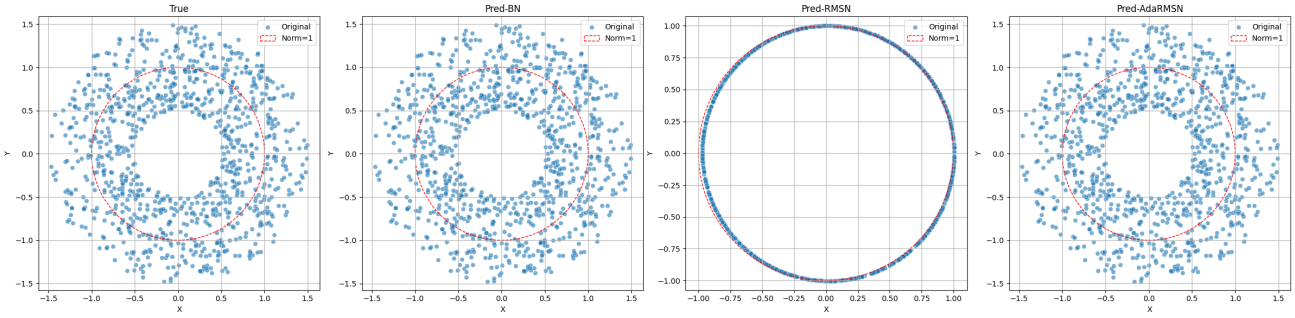

Figure 2. (Case Study of AdaRMSN) Visualization of Input and Pred data points [(1). Input; (2) Predictions w/ BN; (3) Predictions w/ RMSN; (4) Predictions w/ AdaRMSN]. We perform an overfitting test on Auto-encoders of 2-dim (**Linear** $\rightarrow$ **BN/RMSN/AdaRMSN** $\rightarrow$ **Linear**): each model is trained 5000 epochs via AdamW without regularization. (together with Figure. 3). RMSN is ineffective in preserving magnitude information, whereas both BN and AdaRMSN successfully maintain the crucial magnitude information of the data points.

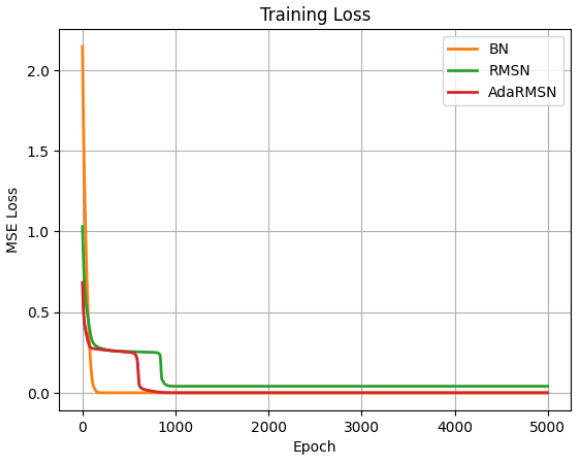

Figure 3. (Case Study of AdaRMSN) Training curves of the overfitting test. (together with Figure. 2). AdaRMSN and BN achieve a lower MSE loss compared to RMSN, demonstrating their superior ability to preserve crucial magnitude information.

Table 1. Performance on ZINC. GPS+sL_2: integrating sL_2 attention into GPS without changing other parts. (run 3 trials). sL_2 attention introduces performance improvements to GPS.

ZINC	**GPS**	**GPS+sL_2**	**SGT**
MAE (↓)	0.070 ± 0.004	0.0693 ± 0.0023	0.0566 ± 0.002

Table 2. Comparison of peak GPU memory usage and per-epoch training time for GRIT and SGT. Dataset: Peptides-Structure (15K graphs); Model config.: 5 transformer layers, 96 channels, batch size 32. Hardware: a single Nvidia V100 GPU with 32GB memory, supported by 80 Intel Xeon Gold 6140 CPUs running at 2.30GHz

Model	**GPU Memory (GB)**	**Training Time (Sec/Epoch)**
GRIT	29.16	141.60
SGT	25.07	100.68
Improv.	∼14.03%	∼28.9%

Table 3. Performance on PCMQM4Mv2 (over 3.7M graphs). The eval. pipeline follows Rampášek et al. (2022); no 3D-info included. SGT outperforms other GTs.

PCMQM4Mv2	**Val MAE (↓)**	**# Param.**
Graphormer	0.0864	48.3M
GPS	0.0858	19.4M
GRIT	0.0859	16.6M
SGT	**0.0856**	17.6M

Table 4. Performance comparison across different models on various datasets. Best results are highlighted in bold. * indicates the difference to the best is not statistically significant (by two-tail T-test)

Model	ZINC MAE (↓)	SP-CIFAR Acc. (↑)	SP-MNIST Acc. (↑)	PATTERN W.Acc. (↑)	CLUSTER W.Acc. (↑)	Peptides-Struct MAE (↓)	Peptides-Func AP (↑)
Exphormer	-	74.69±0.125	98.55±0.037	86.74±0.015	78.07±0.037	0.2481±0.0007	0.6527±0.0043
GEAET	-	76.634±0.427	98.513±0.086	86.993±0.026	-	**0.2445±0.0013**	-
GEANet	0.193±0.001	73.857±0.306	98.315±0.097	85.607±0.038	77.013±0.224	0.2512±0.0003	0.6722±0.0065
SGT	**0.0566±0.002**	**78.560±0.700**	**98.614±0.096**	**89.752±0.030**	**80.027±0.114**	0.2450±0.0017*	**0.6961±0.0062**

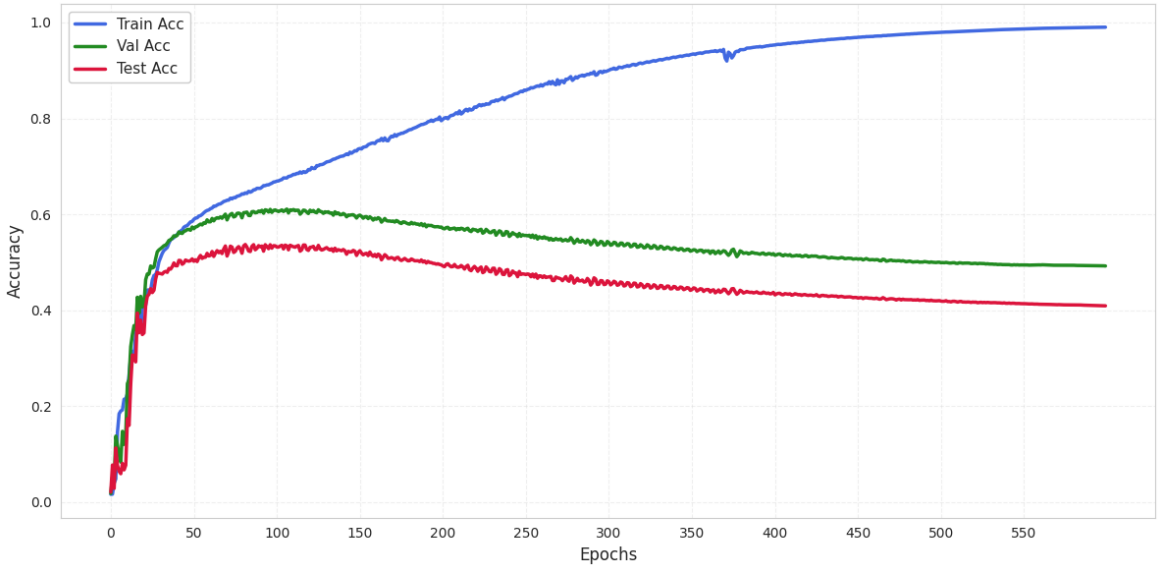

Figure 4. Sanity check of Exp-SGT on a large-scale graph in OGBN-ArXiv (169,343 nodes). Use the same configuration as Exphormer and remove all regularizations to validate the trainability via an overfitting test. Exp-SGT can be successfully trained on OGBN-ArXiv, achieving close to 100% training accuracy.