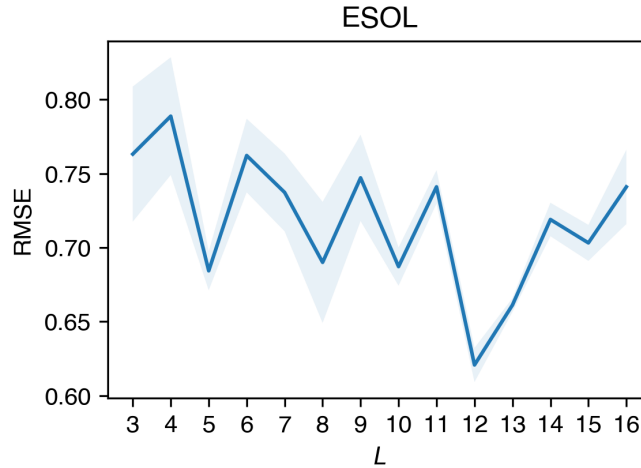


	Texas	Wisc.	Cornell
GCN	55.1 $\pm$ 4.2	51.8 $\pm$ 3.3	60.5 $\pm$ 4.8
GAT	52.2 $\pm$ 6.6	51.8 $\pm$ 3.1	60.5 $\pm$ 5.3
GCNII	77.6 $\pm$ 3.8	80.4 $\pm$ 3.4	77.9 $\pm$ 3.8
Geom-GCN	66.8 $\pm$ 2.7	64.5 $\pm$ 3.7	60.5 $\pm$ 3.7
PairNorm	60.3 $\pm$ 4.3	48.4 $\pm$ 6.1	58.9 $\pm$ 3.2
GPS	75.4 $\pm$ 1.5	78.0 $\pm$ 2.9	65.4 $\pm$ 5.7
Transformer	77.8 $\pm$ 1.1	76.1 $\pm$ 1.9	71.9 $\pm$ 2.5
Graphomer	76.8 $\pm$ 1.8	77.7 $\pm$ 2.0	68.4 $\pm$ 1.7
RUM	80.0 $\pm$ 7.0	85.8 $\pm$ 4.1	71.1 $\pm$ 5.6

Expanded Table 4: Node classification test set accuracy $\uparrow$ and standard deviation on heterophilic datasets, compared with convolutional GNNs and graph transformers.

Method	#param.	AUC (%)
GCN-GraphNorm	526K	78.83 $\pm$ 1.00
PNA	326K	79.05 $\pm$ 1.32
PHC-GNN	111K	79.34 $\pm$ 1.16
DeeperGCN-FLAG	532K	79.42 $\pm$ 1.20
DGN	114K	79.70 $\pm$ 0.97
GIN-VN (fine-tune)	3.3M	77.80 $\pm$ 1.82
Graphormer-FLAG	47.0M	80.51 $\pm$ 0.53
RUM	87K	80.01 $\pm$ 1.20

New Table: Test set performance $\downarrow$ on MolHIV dataset.



New Appendix Figure: Graph regression RMSE $\downarrow$ plotted against L , the length of the random walk. Shown here is the mean over 5 experiments with standard deviation marked in shades.