Enhancing Graph Transformers with Hierarchical Distance Structural Encoding

Yuankai Luo, Hongkang Li, Lei Shi, Xiao-Ming Wu

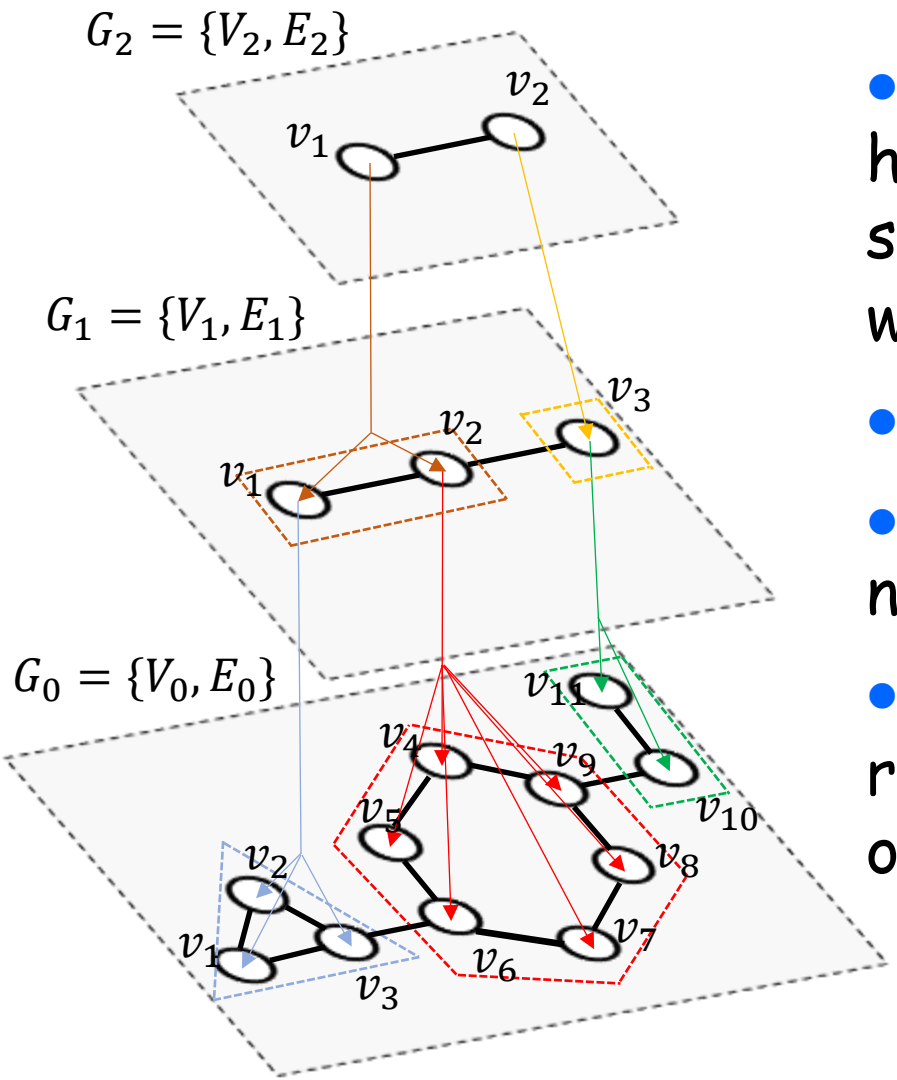
北京航空航天大學
BEIHANG UNIVERSITY

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Rensselaer

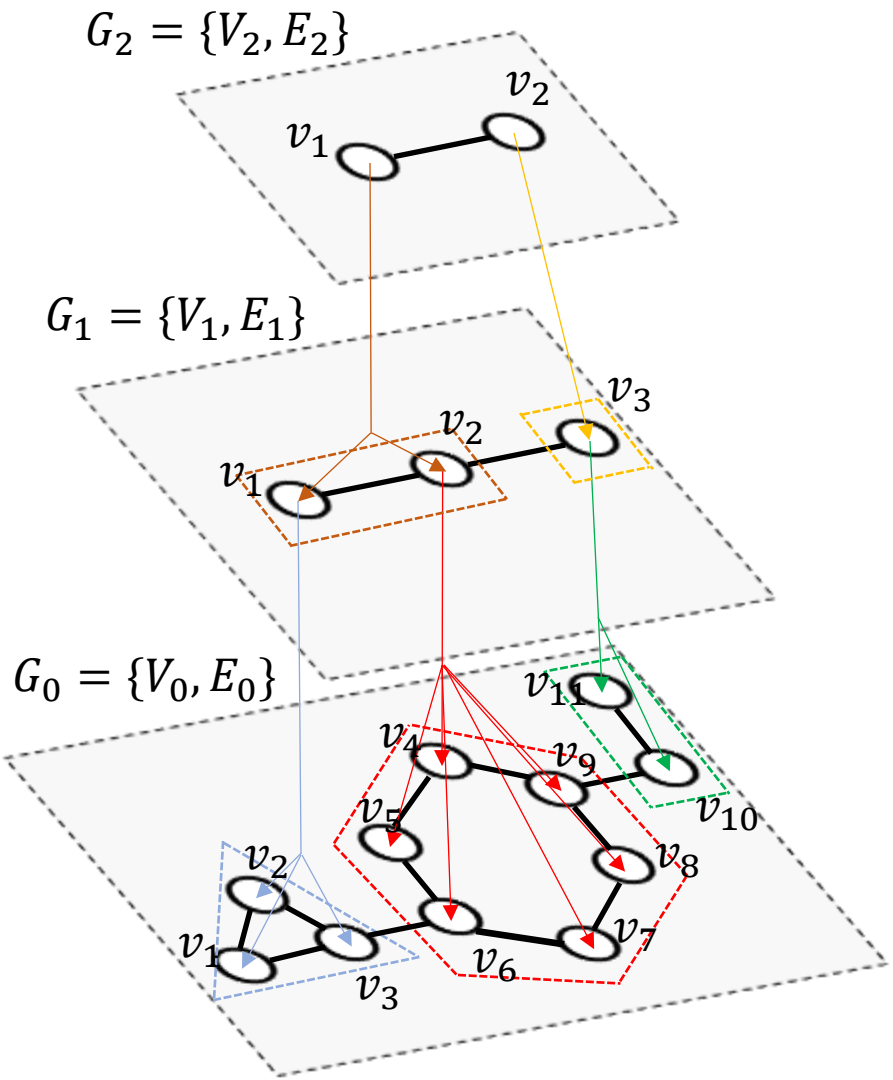
NeurIPS 2024

Graph Hierarchies



- Given an input graph G, a graph hierarchy of G consists of a sequence of graphs $(G_k, \varphi_k)_{\{k \geq 0\}}$, where:
- G_0 denotes G.
- φ_k: $V_k \rightarrow V_{k+1}$ are surjective node mapping functions.
- Each node $v_{k+1,j} \in V_{k+1}$ represents a cluster of a subset of nodes $\{v_{k,i}\} \in V_k$.

Graph Hierarchies

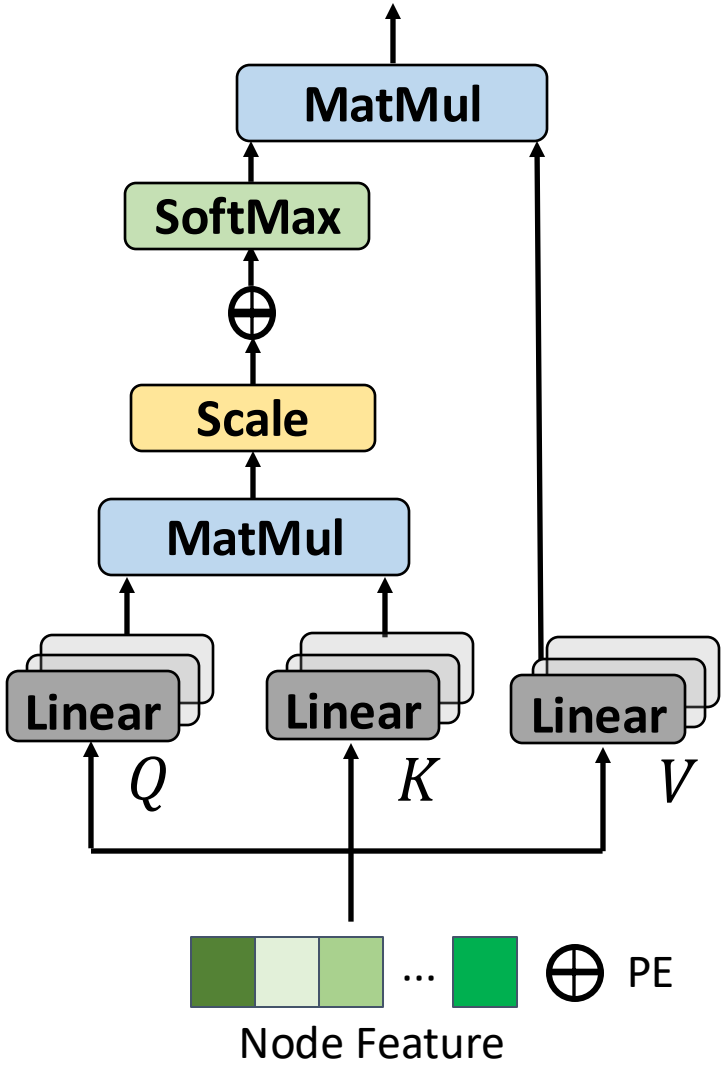


- Graph hierarchies can be constructed by repeatedly applying graph coarsening algorithms:
- METIS, Spectral clustering, Loukas methods, Newman methods, Louvain methods

- These algorithms take a graph, G_k, and generate a mapping function φ_k: $V_k \rightarrow V_{k+1}$, which maps the nodes in G_k to the nodes in the coarser graph G_{k+1}.

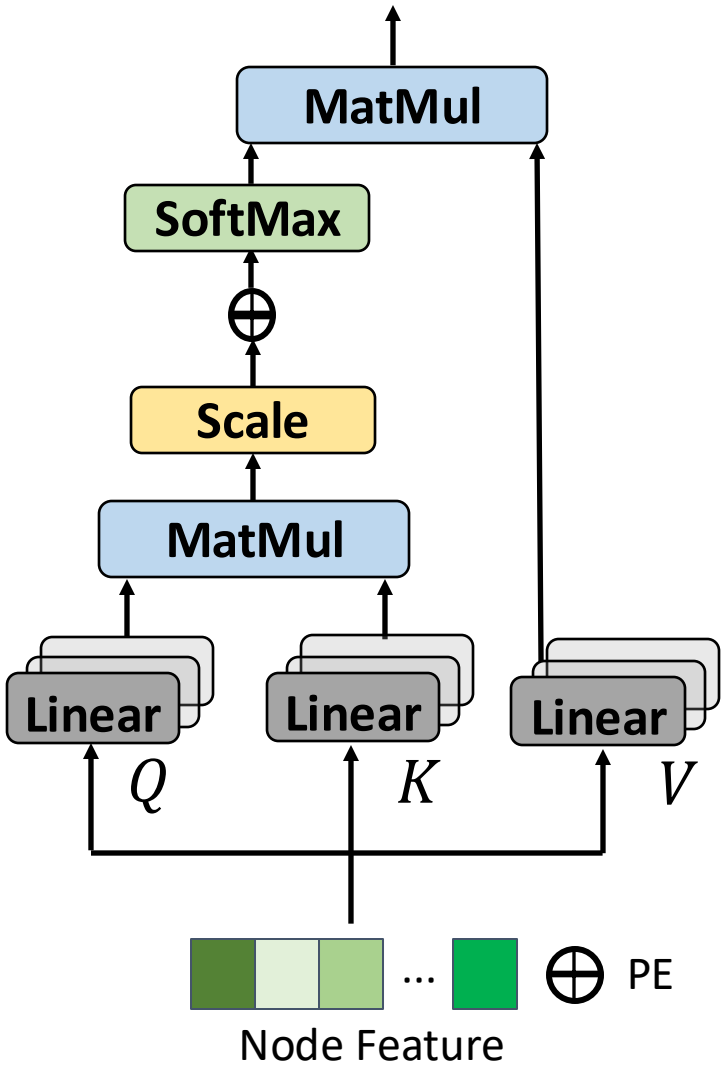
Transformers on Graphs

- Transformers have revolutionized deep learning, in particular sequence learning, and yield promising performance over graphs.
- Graph Transformers usually apply the regular attention quadratic attention across all graph nodes and encode the graph connectivity using specific positional encodings (PEs).



Transformers on Graphs

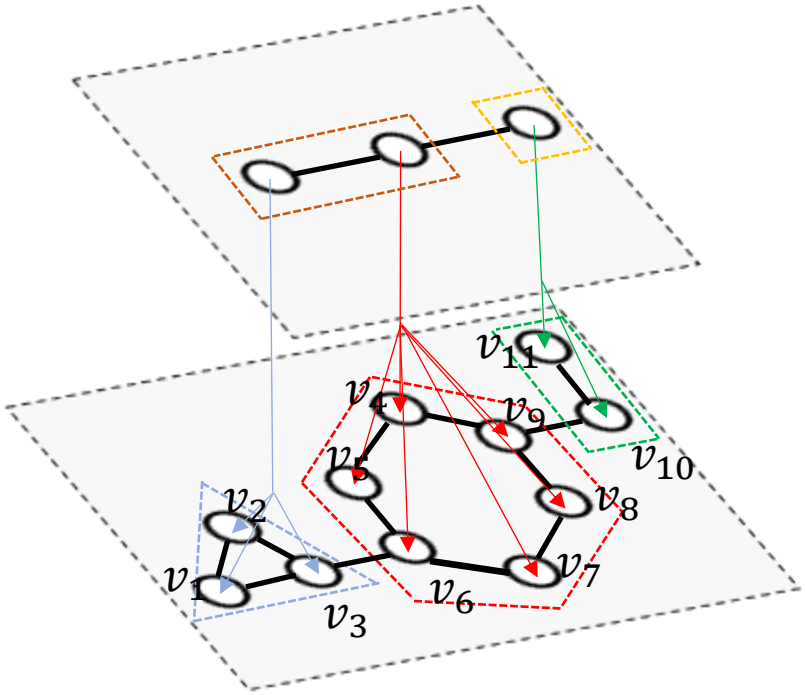
However, Graph Transformers struggle with learning hierarchical structures, limiting their performance, for example, on complex molecular graphs like polymers and proteins.



Graph Hierarchy Distance

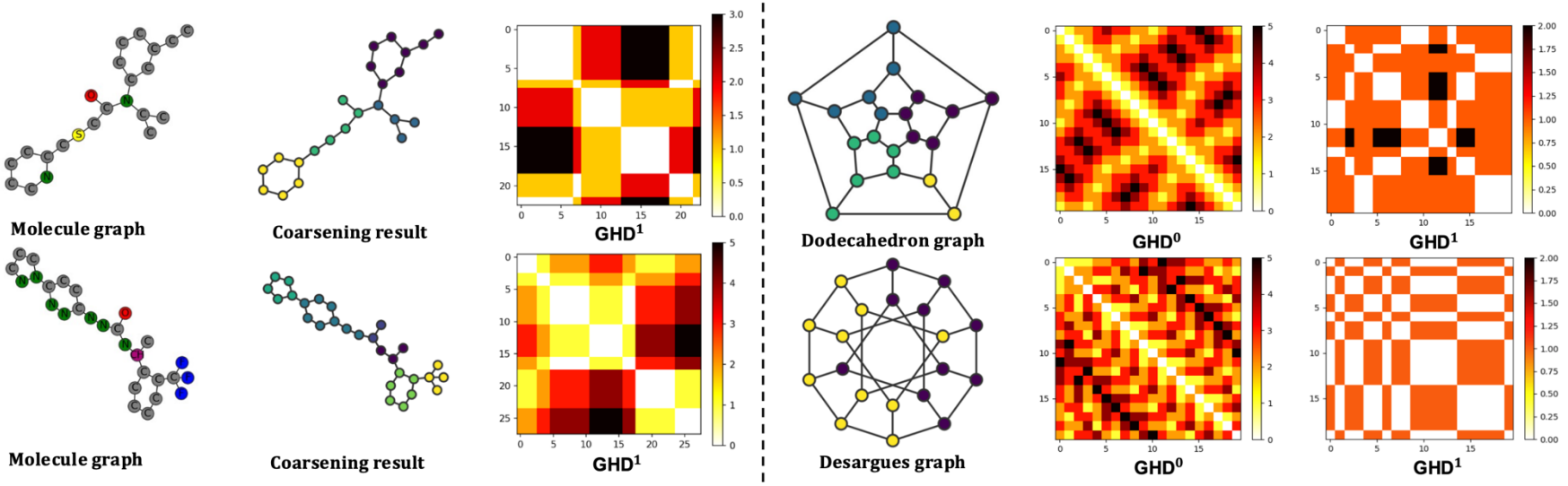
- We introduce a novel distance called graph hierarchy distance (GHD):

$$\mathrm{GHD}^0(u,v) = \mathrm{SPD}(u,v),$$
$$\mathrm{GHD}^k(u,v) = \mathrm{SPD}(\phi_{k-1}\ldots\phi_0(u), \phi_{k-1}\ldots\phi_0(v))$$



- It can be observed that GHD^0 (v1, v11) = 7, whereas GHD^1 (v1, v11) = 2

Graph Hierarchy Distance



- GHD can capture chemical motifs such as CF3 and aromatic rings on molecule graphs.
- GHD can distinguish the Dodecahedron and Desargues graphs. The Dodecahedral graph has GHD^1 of length 2 (indicated by the dark color), while the Desargues graph doesn't.

Hierarchical Distance Structural Encoding

- Based on GHD, we propose hierarchical distance structural encoding (HDSE):

$$\mathrm{D}_{i,j} = \left[\mathrm{GHD}^0, \mathrm{GHD}^1, ..., \mathrm{GHD}^K\right]_{i,j} \in \mathbb{R}^{K+1}$$

where K controls the maximum level of hierarchy.

- [Expressiveness of HDSE]:

GD-WL with HDSE is strictly more expressive than GD-WL with the shortest path distance SPD.

Integrating HDSE in Graph Transformers

- We integrate HDSE into the attention mechanism of each graph transformer layer to bias each node update:

$$\mathrm{H}_{i,j} = \mathrm{MLP}\left(\left[\mathbf{e}^0_{\mathrm{clip}^0_{i,j}}, \cdots, \mathbf{e}^K_{\mathrm{clip}^K_{i,j}}\right]\right) \in \mathbb{R},$$

$$\mathrm{clip}^k_{i,j} = \min\left(L, \mathrm{GHD}^k_{i,j}\right), 0 \leq k \leq K,$$

$$\mathrm{Attention}\left(\mathbf{X}\right) = \mathrm{softmax}\left(\mathrm{A} + \mathrm{H}\right)\mathbf{V}, \mathrm{A} = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d'}}$$

This module is backbone-agnostic and can be seamlessly integrated into the self-attention mechanism of existing graph transformer architectures.

Integrating HDSE in Graph Transformers

- [Expressiveness of Graph Transformers with HDSE]:

There exists a graph transformer using HDSE (with fixed parameters), denoted as M, such that M is more expressive than graph transformers with the same architecture using SPD or using no relative positional encoding, regardless of their parameters.

It demonstrate the superior expressiveness of HDSE over SPD or no RPE in graph transformers.

Integrating HDSE in Graph Transformers

- [Generalization of Graph Transformers with HDSE]:

For a semi-supervised binary node classification problem, suppose the label of each node $i \in V$ is determined by node features in the "hierarchical core neighborhood" $S_i = \{j : D = D^*\}$ for a certain D^*, where D is HDSE. Then, a properly initialized one-layer graph transformer equipped with HDSE can learn such graphs with a desired generalization error, while using SPD or using no relative positional encoding cannot guarantee satisfactory generalization.

It indicates that learning with HDSE can capture the labeling function characterized by the hierarchical core neighborhood, which is more general and comprehensive than the core neighborhood for SPD or no RPE.

Evaluation

Table 2: Test performance in five benchmarks from [20]. The results are presented as the mean ± standard deviation from 5 runs using different random seeds. Baseline results were obtained from their respective original papers. * indicates a statistically significant difference against the baseline w/o HDSE from the one-tailed t-test. Highlighted are the top **first**, **second** and **third** results.

Model	ZINC MAE ↓	MNIST Accuracy ↑	CIFAR10 Accuracy ↑	PATTERN Accuracy ↑	CLUSTER Accuracy ↑
GCN	0.367 ± 0.011	90.705 ± 0.218	55.710 ± 0.381	71.892 ± 0.334	68.498 ± 0.976
GIN	0.526 ± 0.051	96.485 ± 0.252	55.255 ± 1.527	85.387 ± 0.136	64.716 ± 1.553
GatedGCN	0.282 ± 0.015	97.340 ± 0.143	67.312 ± 0.311	85.568 ± 0.088	73.840 ± 0.326
PNA	0.188 ± 0.004	97.940 ± 0.120	70.350 ± 0.630	–	–
CIN	0.079 ± 0.006	–	–	–	–
GIN-AK+	0.080 ± 0.001	–	72.190 ± 0.130	86.850 ± 0.057	–
SGFormer	0.306 ± 0.023	–	–	85.287 ± 0.097	69.972 ± 0.634
SAN	0.139 ± 0.006	–	–	86.581 ± 0.037	76.691 ± 0.650
Graphormer-GD	0.081 ± 0.009	–	–	–	–
Specformer	**0.066 ± 0.003**	–	–	–	–
EGT	0.108 ± 0.009	98.173 ± 0.087	68.702 ± 0.409	86.821 ± 0.020	**79.232 ± 0.348**
Graph ViT/MLP-Mixer	0.073 ± 0.001	97.422 ± 0.110	73.961 ± 0.330	–	–
Exphormer	-	**98.550 ± 0.039**	74.696 ± 0.125	86.742 ± 0.015	78.071 ± 0.037
GT	0.226 ± 0.014	90.831 ± 0.161	59.753 ± 0.293	84.808 ± 0.068	73.169 ± 0.622
GT + HDSE	0.159 ± 0.006*	94.394 ± 0.177*	64.651 ± 0.591*	86.713 ± 0.049*	74.223 ± 0.573*
SAT	0.094 ± 0.008	–	–	86.848 ± 0.037	77.856 ± 0.104
SAT + HDSE	0.084 ± 0.003*	–	–	**86.933 ± 0.039***	78.513 ± 0.097*
GraphGPS	0.070 ± 0.004	98.051 ± 0.126	72.298 ± 0.356	86.685 ± 0.059	78.016 ± 0.180
GraphGPS + HDSE	**0.062 ± 0.003***	**98.367 ± 0.106***	**76.180 ± 0.277***	86.737 ± 0.055	78.498 ± 0.121*
GRIT	**0.059 ± 0.002**	98.108 ± 0.111	**76.468 ± 0.881**	**87.196 ± 0.076**	**80.026 ± 0.277**
GRIT + HDSE	**0.059 ± 0.004**	**98.424 ± 0.124***	**76.473 ± 0.429**	**87.281 ± 0.064**	**79.965 ± 0.203**

Evaluation

Table 3: Test performance on two peptide datasets from Long-Range Graph Benchmarks (LRGB) [23].

Model	Peptides-func AP ↑	Peptides-struct MAE ↓
GCN	0.5930 ± 0.0023	0.3496 ± 0.0013
GINE	0.5498 ± 0.0079	0.3547 ± 0.0045
GatedGCN	0.5864 ± 0.0035	0.3420 ± 0.0013
GatedGCN+RWSE	0.6069 ± 0.0035	0.3357 ± 0.0006
GT	0.6326 ± 0.0126	0.2529 ± 0.0016
SAN+RWSE	0.6439 ± 0.0075	0.2545 ± 0.0012
MGT+WavePE	0.6817 ± 0.0064	**0.2453 ± 0.0025**
GRIT	**0.6988 ± 0.0082**	**0.2460 ± 0.0012**
Exphormer	0.6527 ± 0.0043	0.2481 ± 0.0007
Graph ViT/MLP-Mixer	0.6970 ± 0.0080	0.2475 ± 0.0015
DRew	**0.7150 ± 0.0044**	0.2536 ± 0.0015
GraphGPS	0.6535 ± 0.0041	0.2500 ± 0.0012
GraphGPS + HDSE	**0.7156 ± 0.0058***	**0.2457 ± 0.0013***

Table 4: Ablation experiments of coarsening algorithms on ZINC.

Model	Coarsening algorithm	ZINC MAE ↓
SAT	w/o	0.094 ± 0.008
	METIS	0.089 ± 0.005
	Spectral	0.088 ± 0.004
	Loukas	**0.084 ± 0.003**
	Newman	**0.087 ± 0.002**
	Louvain	0.088 ± 0.003
GraphGPS	w/o	0.070 ± 0.004
	METIS	0.069 ± 0.002
	Spectral	**0.063 ± 0.003**
	Loukas	0.067 ± 0.002
	Newman	**0.062 ± 0.003**
	Louvain	0.064 ± 0.002

- Over all datasets, our HDSE makes the transformers outperform the original transformers.
- Different graph coarsening algorithms result in distinct multi-level graph structures. The Newman algorithm exhibits optimal performance on small molecular graphs.

Evaluation

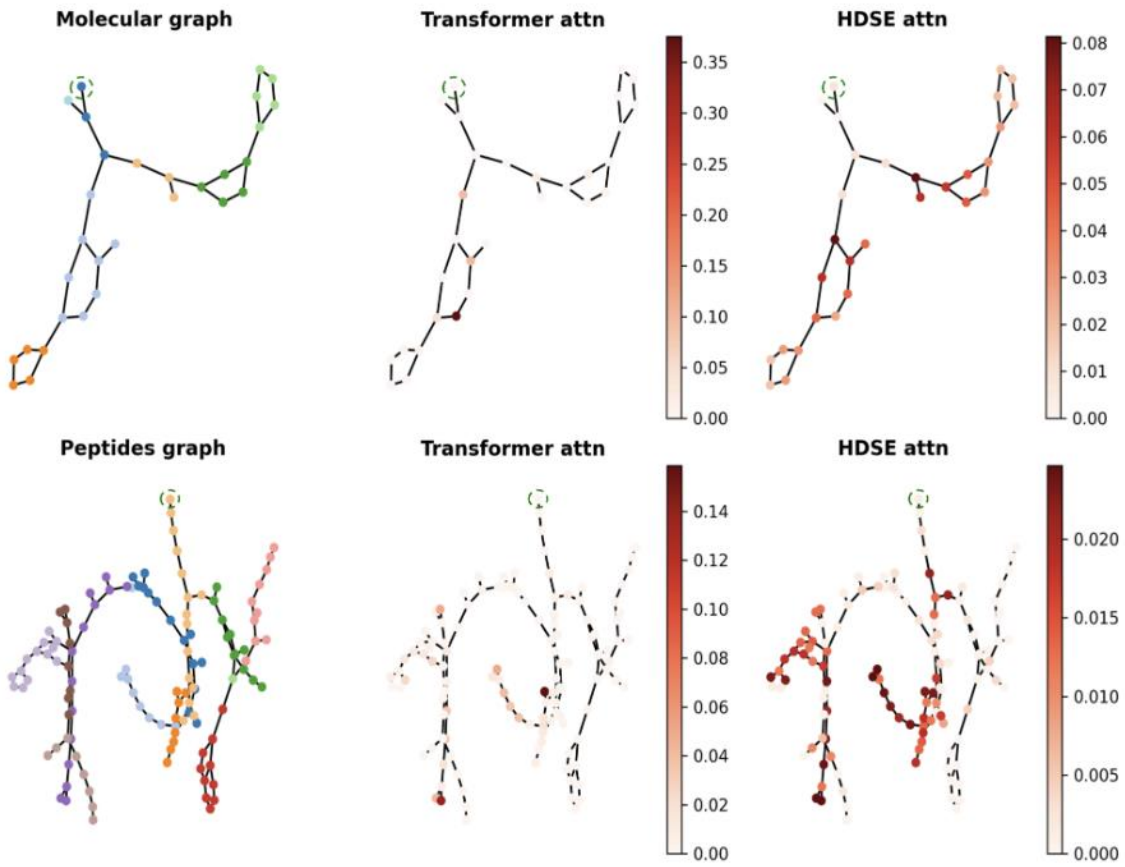


Figure 3: Visualization of attention weights for the transformer attention and HDSE attention. The left side illustrates the graph coarsening result. The center column displays the attention weights of a sample node learned by the classic GT [19], while the right column showcases the attention weights learned by the HDSE attention.

- HDSE successfully leverages hierarchical structure.

Conclusions

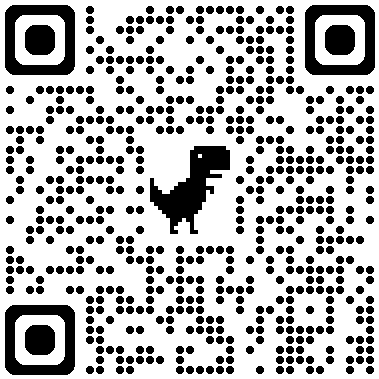
- Our HDSE improves SOTA graph transformer performance on graphs which exhibit community structures.
- We theoretically prove the superiority of HDSE in terms of expressivity and generalization

https://github.com/LUOyk1999/HDSE

Conclusions

- Our HDSE improves SOTA graph transformer performance on graphs which exhibit community structures.
- We theoretically prove the superiority of HDSE in terms of expressivity and generalization

Thanks for listening!



https://github.com/LUOyk1999/HDSE