



**ICLR**  
International Conference On  
Learning Representations

# AutoGT: Automated Graph Transformer Architecture Search

---

Zizhao Zhang, Xin Wang, Chaoyu Guan, Ziwei Zhang,  
Haoyang Li, Wenwu Zhu

Tsinghua University



# Graph Transformer

- Graph Transformer uses **graph encoding** to introduce graph information to Transformer

# Graph Transformer

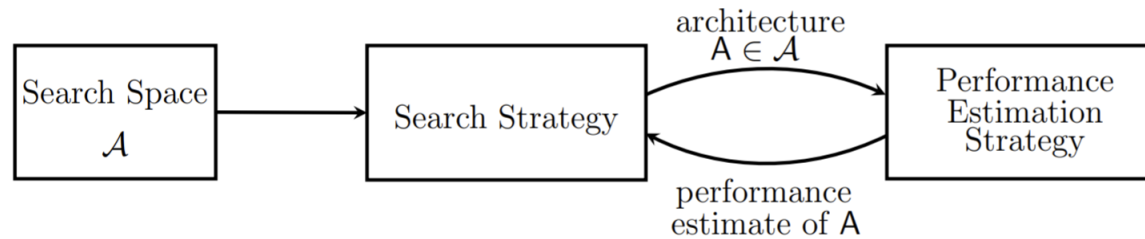
- Graph Transformer uses **graph encoding** to introduce graph information to Transformer
- It has attracted intensive research interests
  - UniMP achieves new state-of-the-art results on OGB
  - Graphormer wins the first place in OGB-LSC 2021
  - GPS++ wins the first place in OGB-LSC 2022

# Graph Transformer

- Graph Transformer uses **graph encoding** to introduce graph information to Transformer
- It has attracted intensive research interests
  - UniMP achieves new state-of-the-art results on OGB
  - Graphormer wins the first place in OGB-LSC 2021
  - GPS++ wins the first place in OGB-LSC 2022
- Limitation
  - Rely on human labor and expertise knowledge to design
  - May introduce bias, which leads to sub-optimal solution

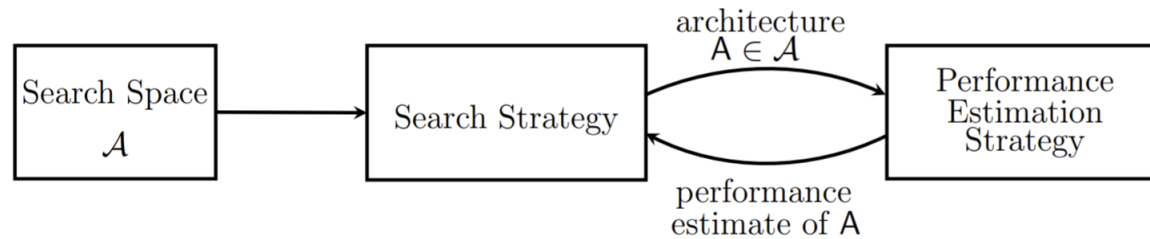
# Neural Architecture Search (NAS)

- Goal: Automatically find the best neural architecture



# Neural Architecture Search (NAS)

- Goal: Automatically find the best neural architecture



- Two challenges:
- How to design a unified search space for graph Transformer?
- How to tackle the coupling relations between Transformer architectures and graph encoding?

# Neural Architecture Search (NAS)

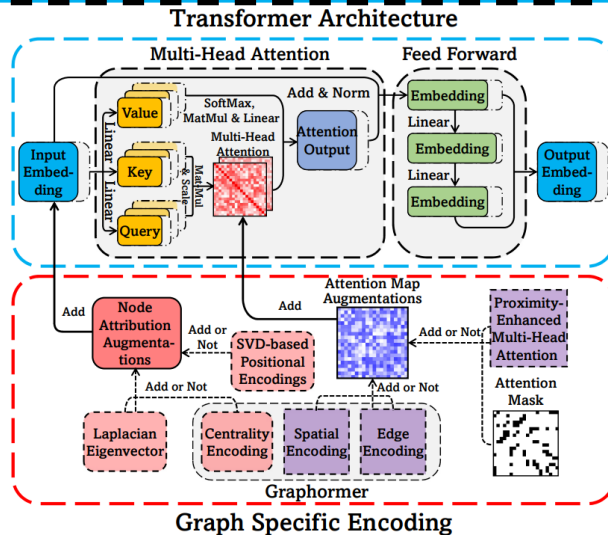
- Two challenges:
- How to design a unified search space for graph Transformer?
- How to tackle the coupling relations between Transformer architectures and graph encoding?



# Neural Architecture Search (NAS)

- Two challenges:
- How to design a unified search space for graph Transformer?
- How to tackle the coupling relations between Transformer architectures and graph encoding?

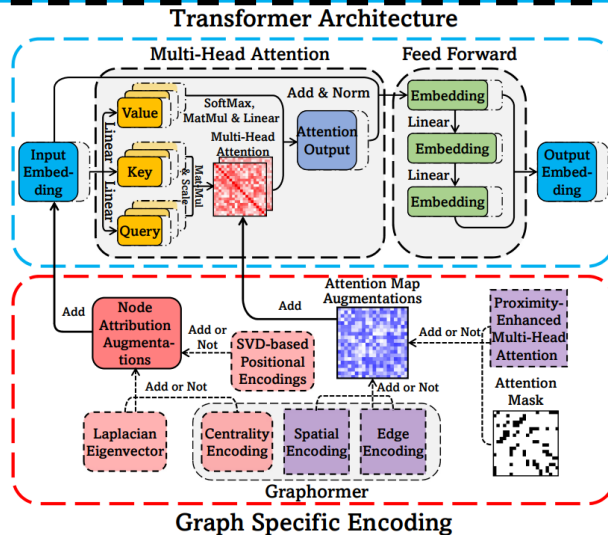
## Graph Transformer Search Space



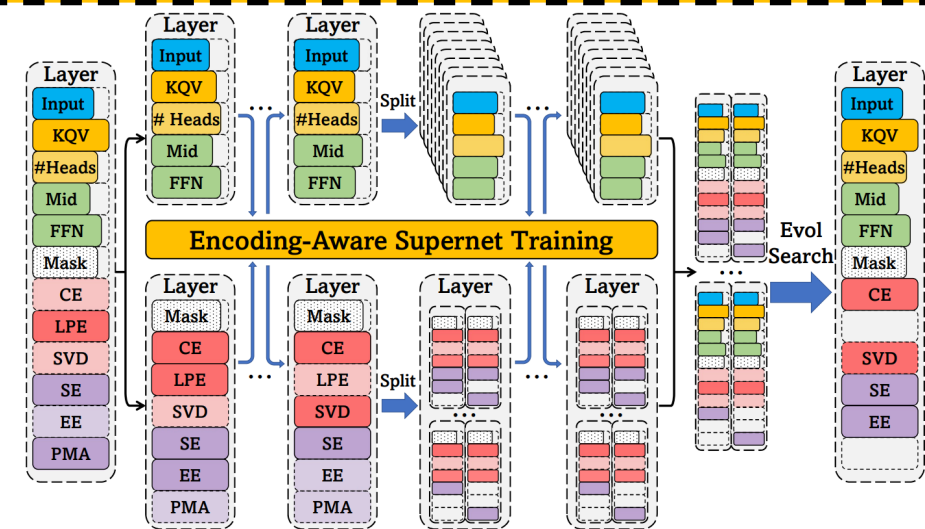
# Neural Architecture Search (NAS)

- Two challenges:
- How to design a unified search space for graph Transformer?
- How to tackle the coupling relations between Transformer architectures and graph encoding?

## Graph Transformer Search Space

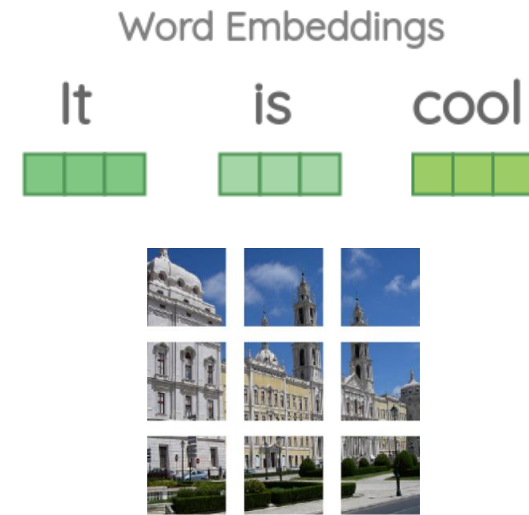


## Encoding-Aware Supernet Training



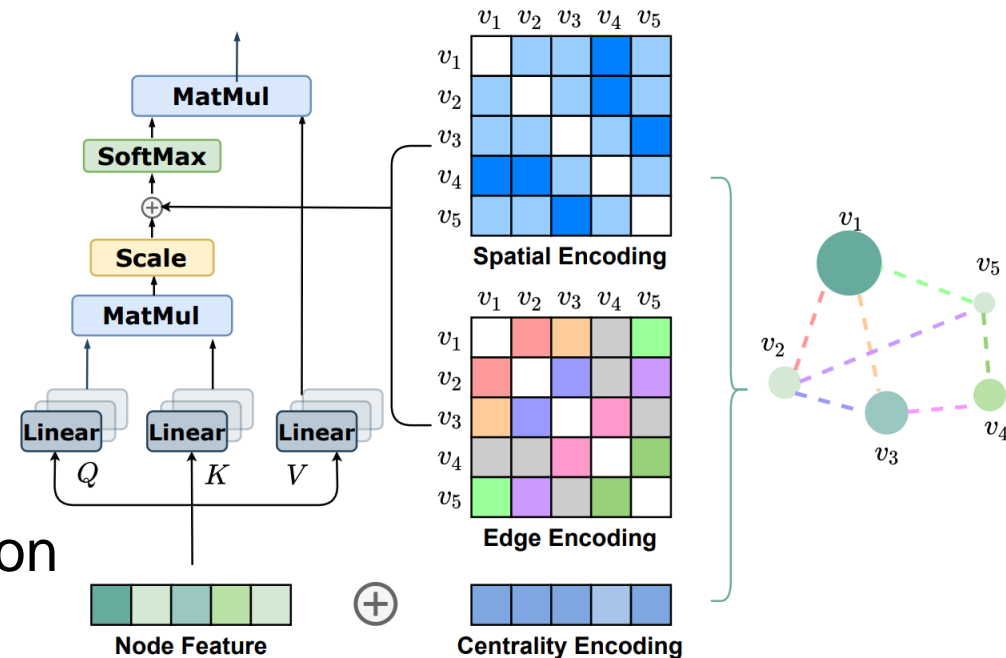
# Graph Encoding

- Transformer Input: Sequential Data
- Language Data: Words in Sentence
- Visual Data: Image Patches in Image
- Graph Data: Nodes in Graph



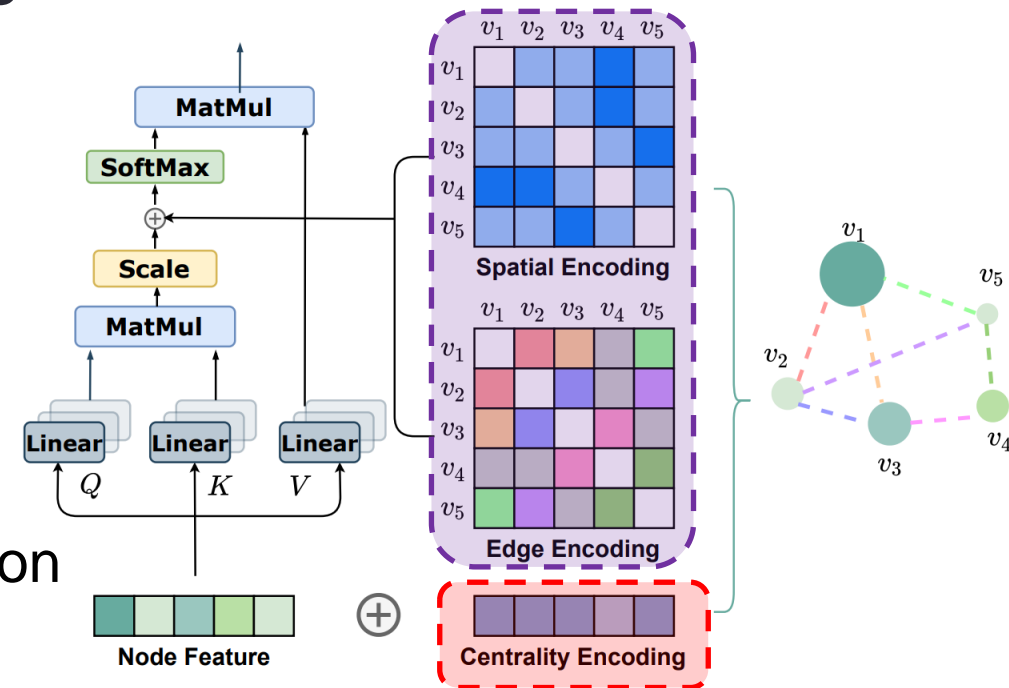
# Graph Encoding

- There are two kinds of graph encoding
- **Node Attribution Augmentations**
  - Pre-calculate node positional encoding
  - Added to the node feature
- **Attention Map Augmentations**
  - Manually designed graph structural information
  - Added to the attention matrix



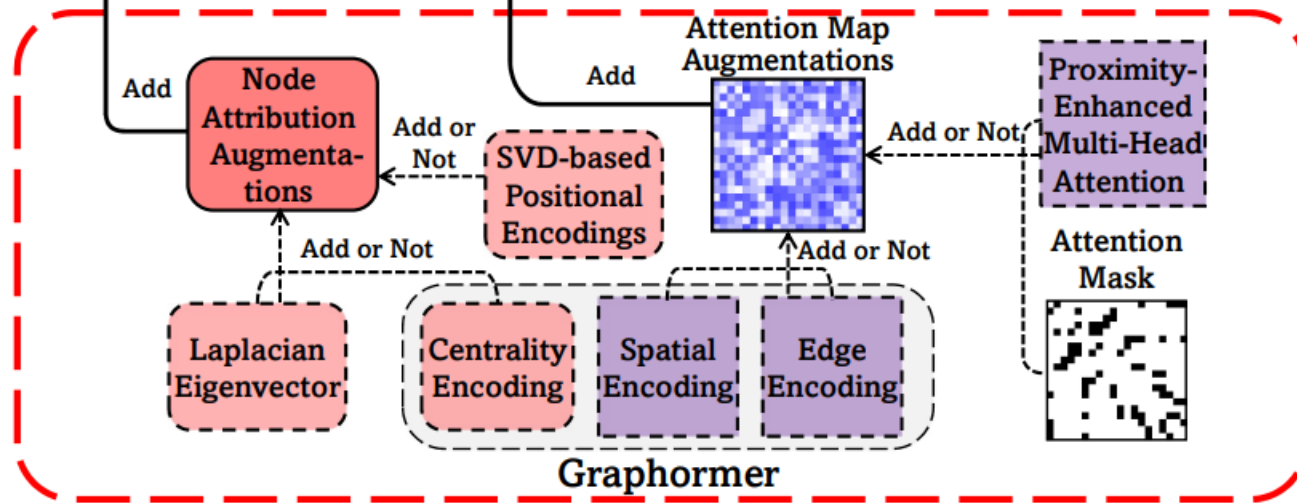
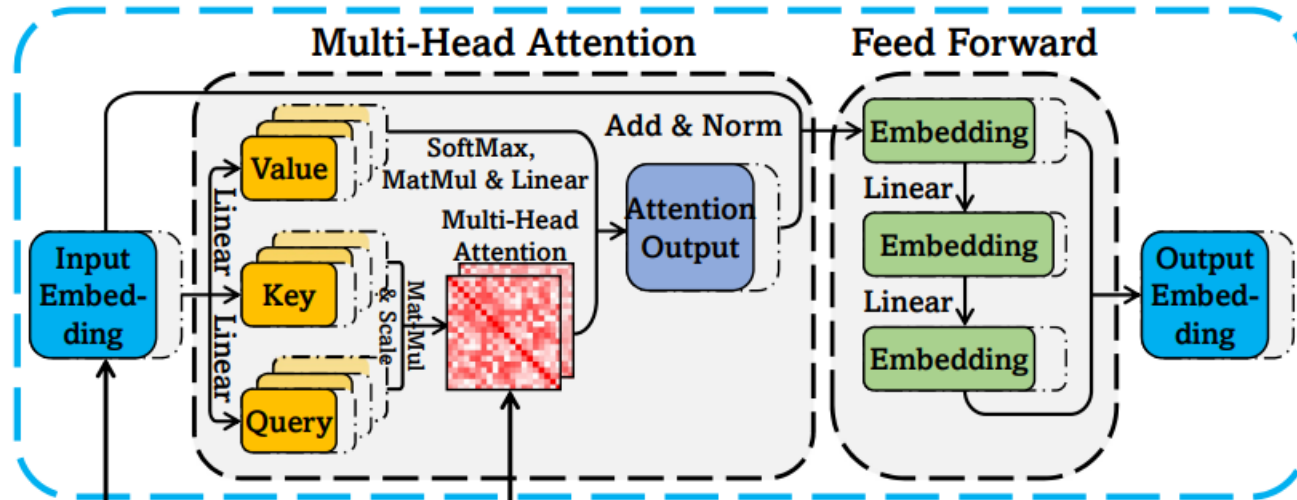
# Graph Encoding

- There are two kinds of graph encoding
- **Node Attribution Augmentations**
  - Pre-calculate node positional encoding
  - Added to the node feature
- **Attention Map Augmentations**
  - Manually designed graph structural information
  - Added to the attention matrix

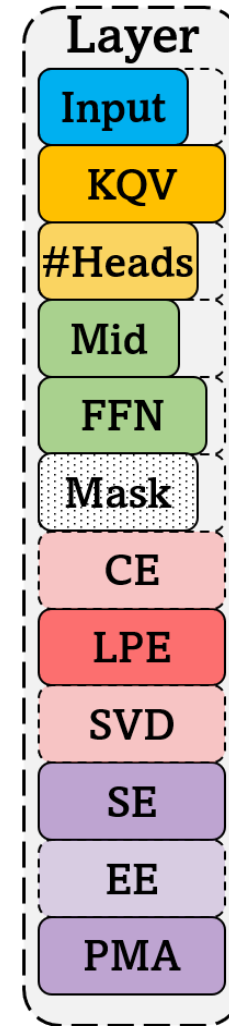


# Graph Transformer Search Space

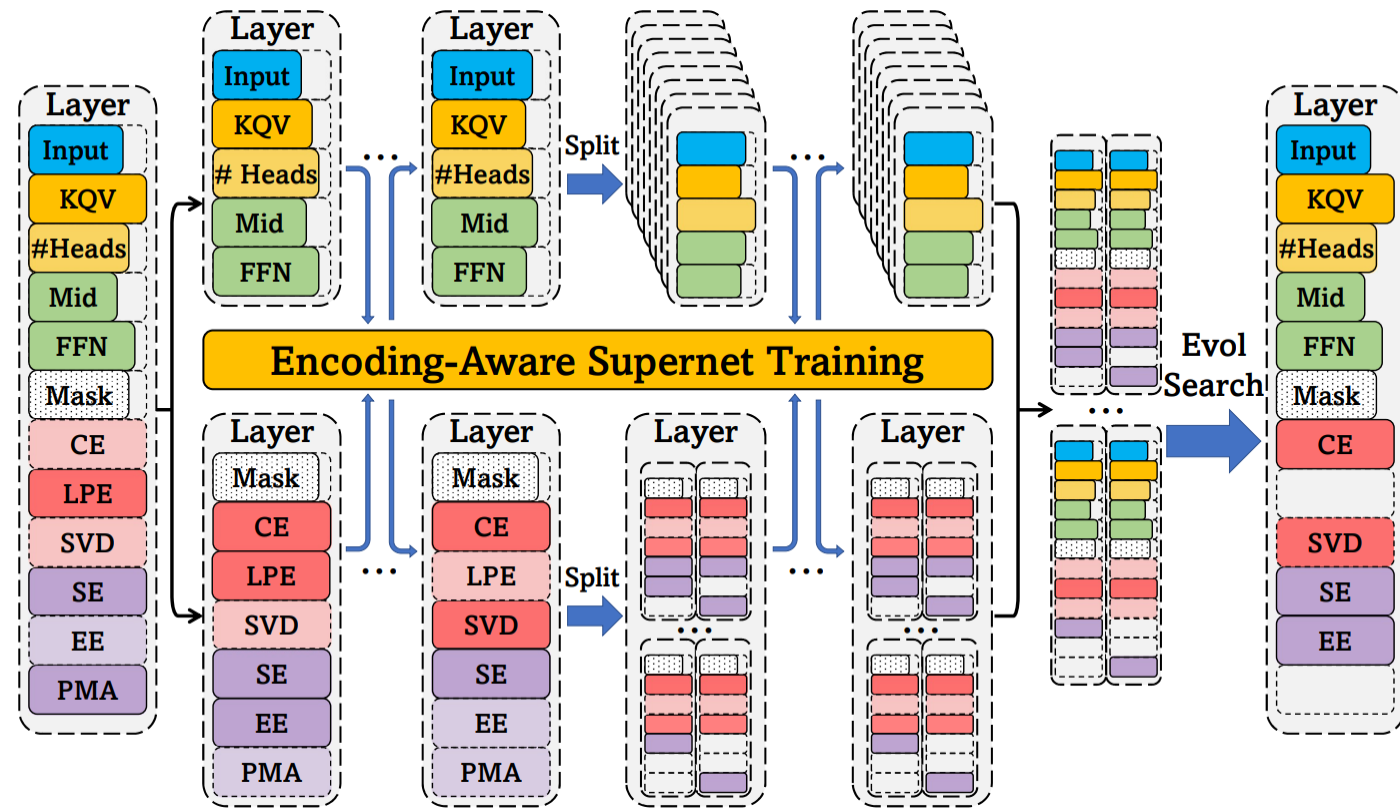
## Transformer Architecture



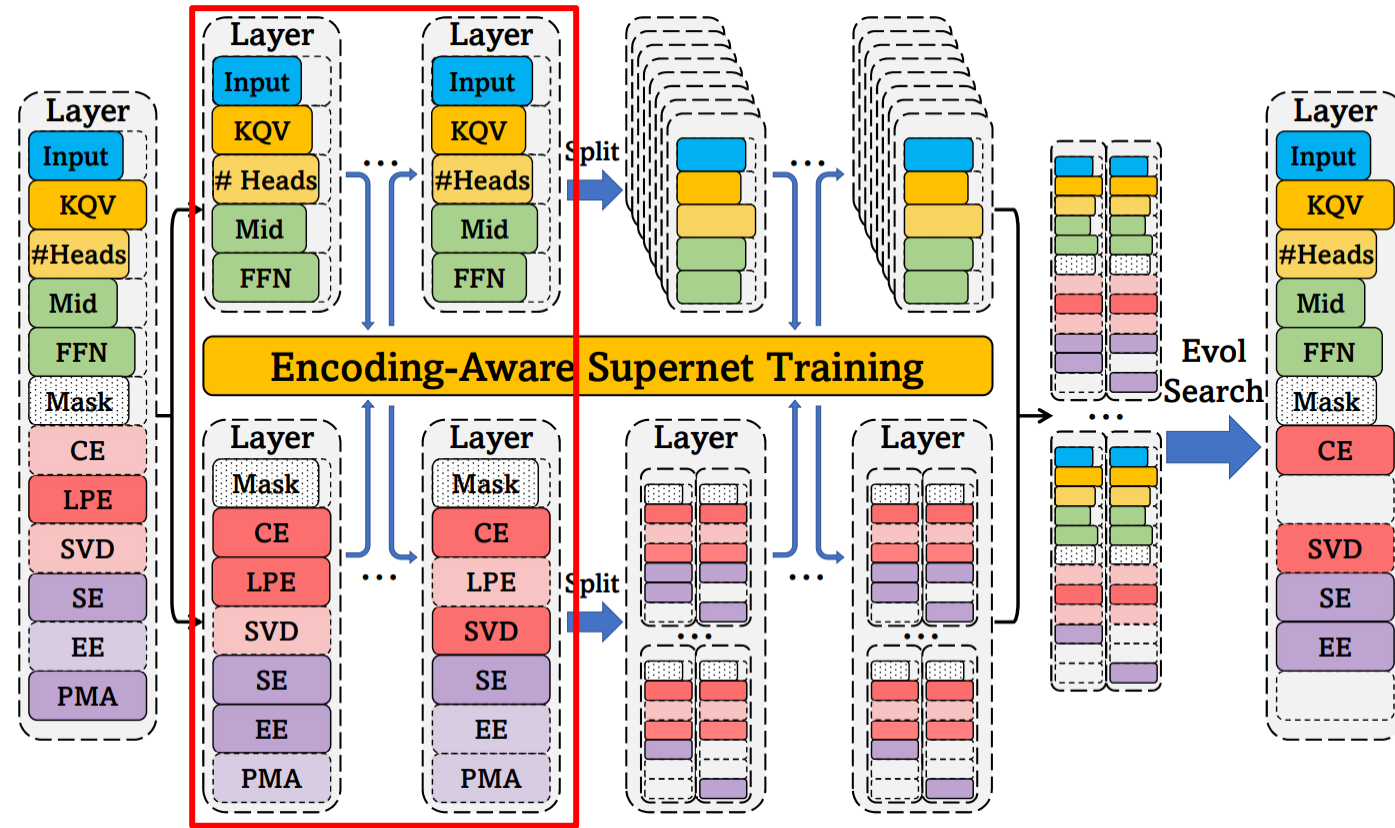
## Graph Specific Encoding



# Encoding-Aware Supernet Training



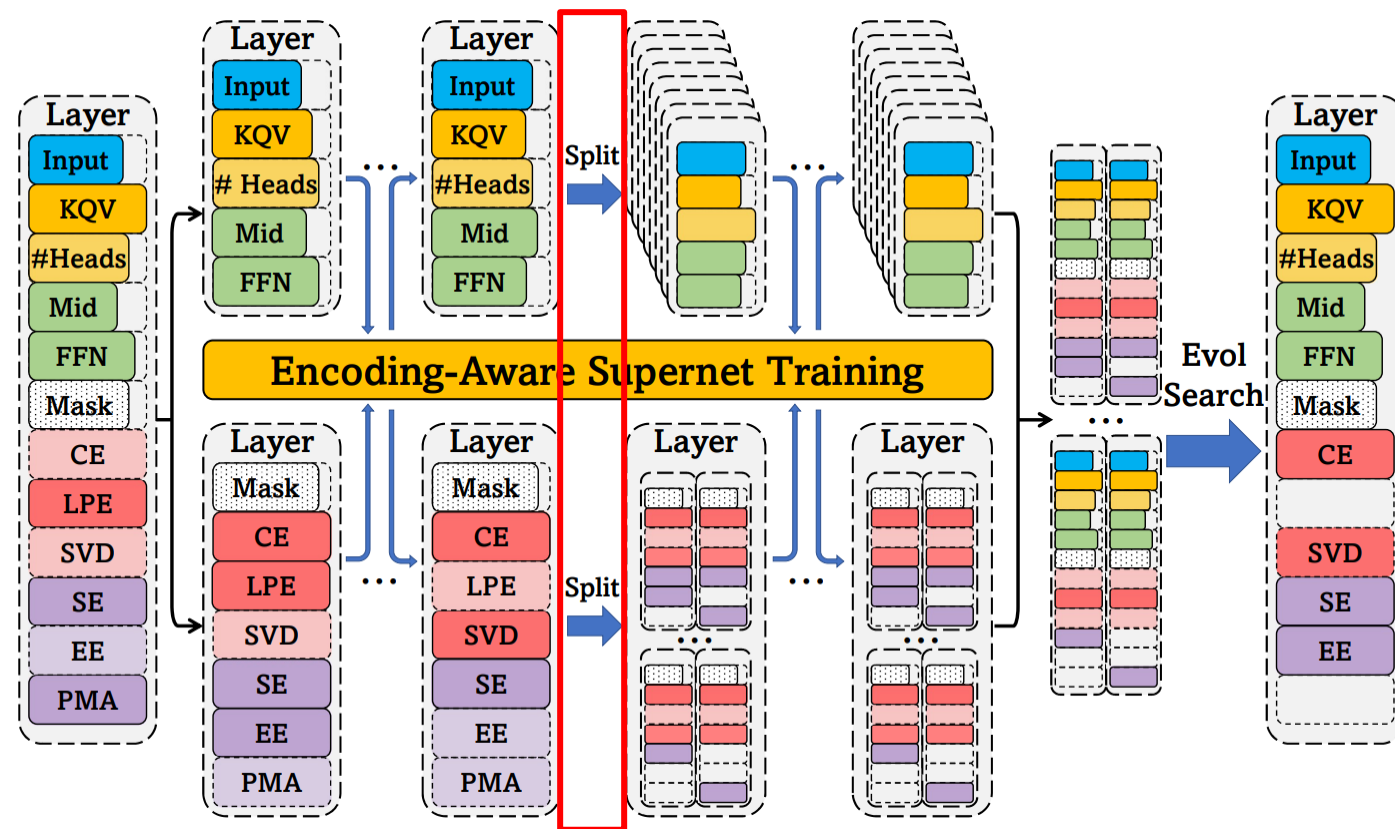
# Encoding-Aware Supernet Training



① Train Supernet

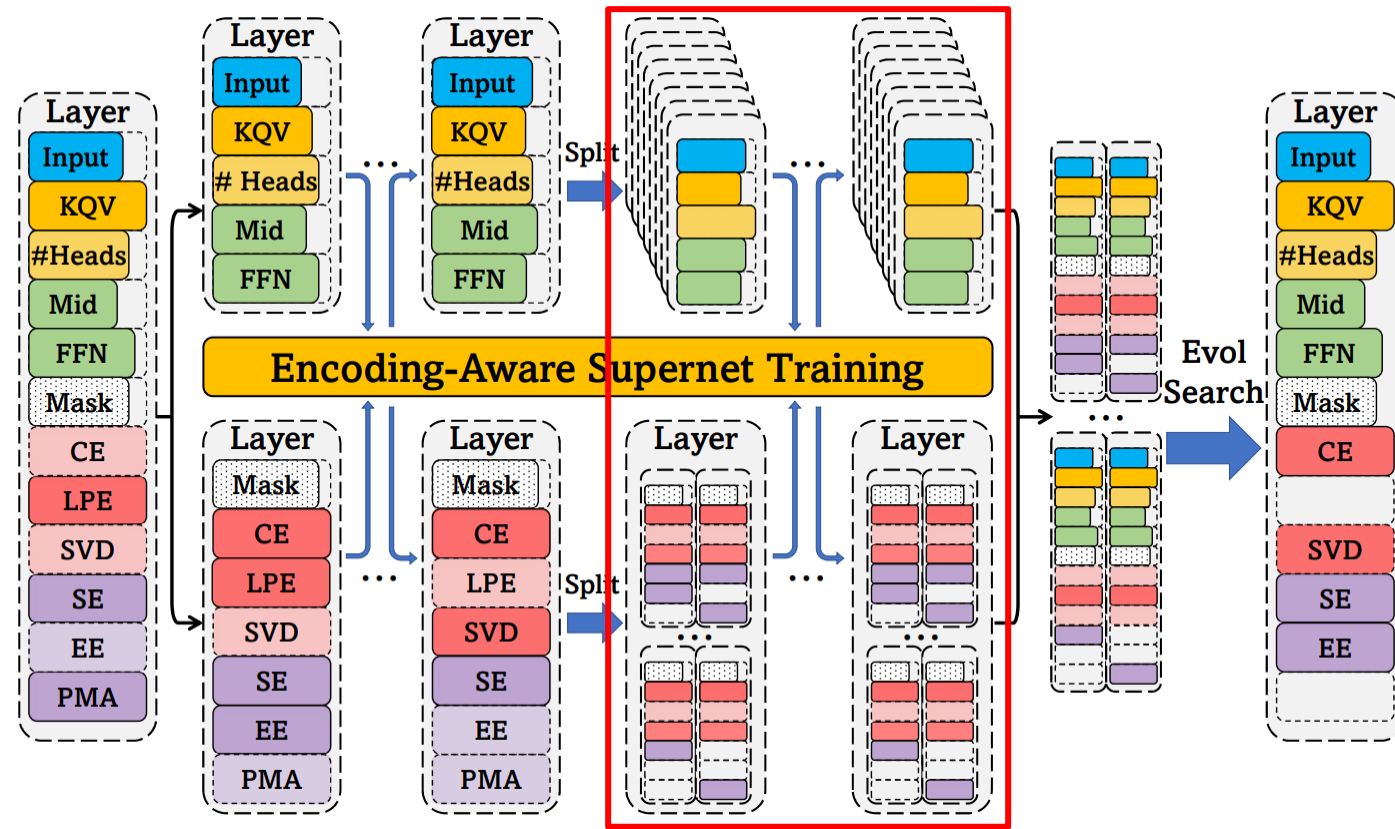


# Encoding-Aware Supernet Training



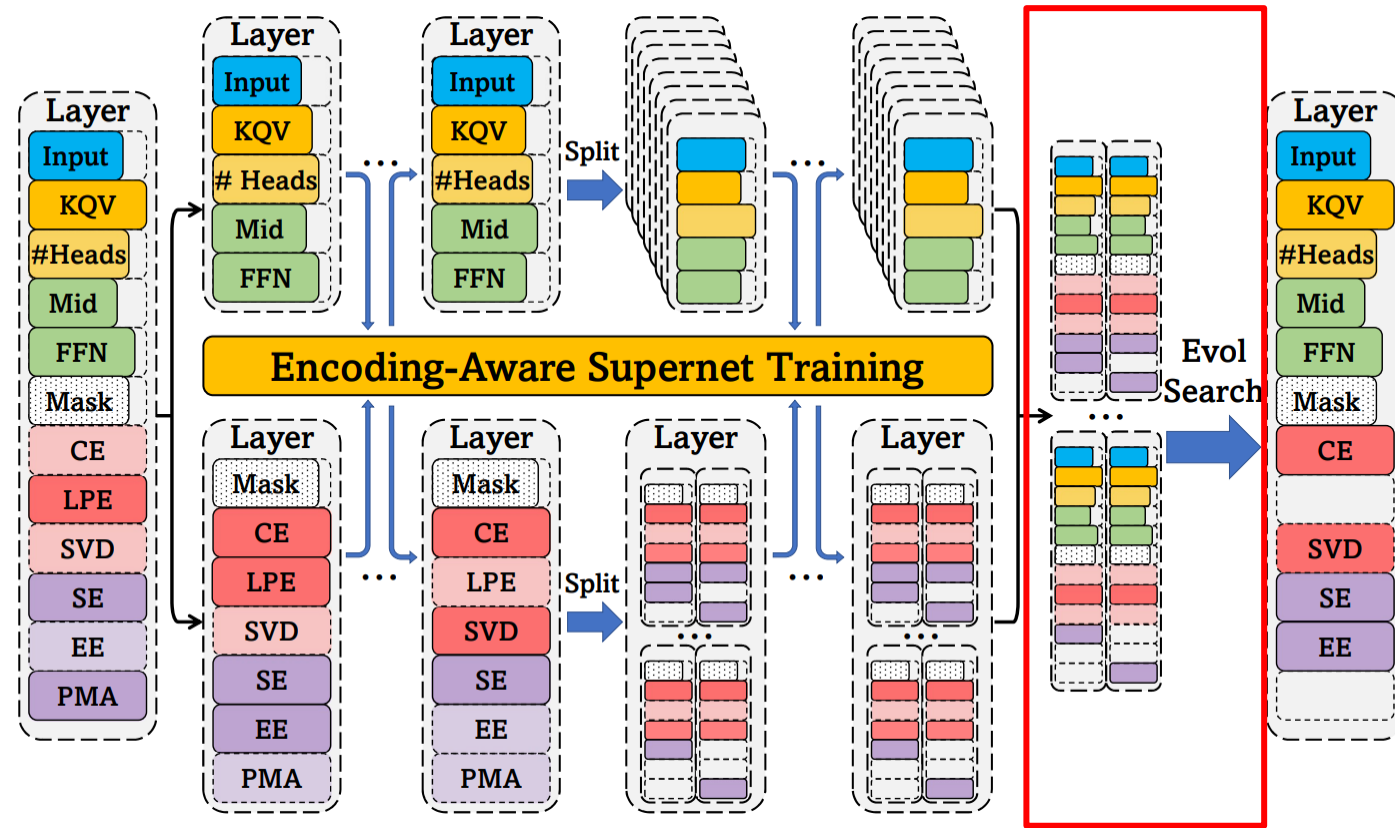
- ① Train Supernet
- ② Split Supernet into Subnets

# Encoding-Aware Supernet Training



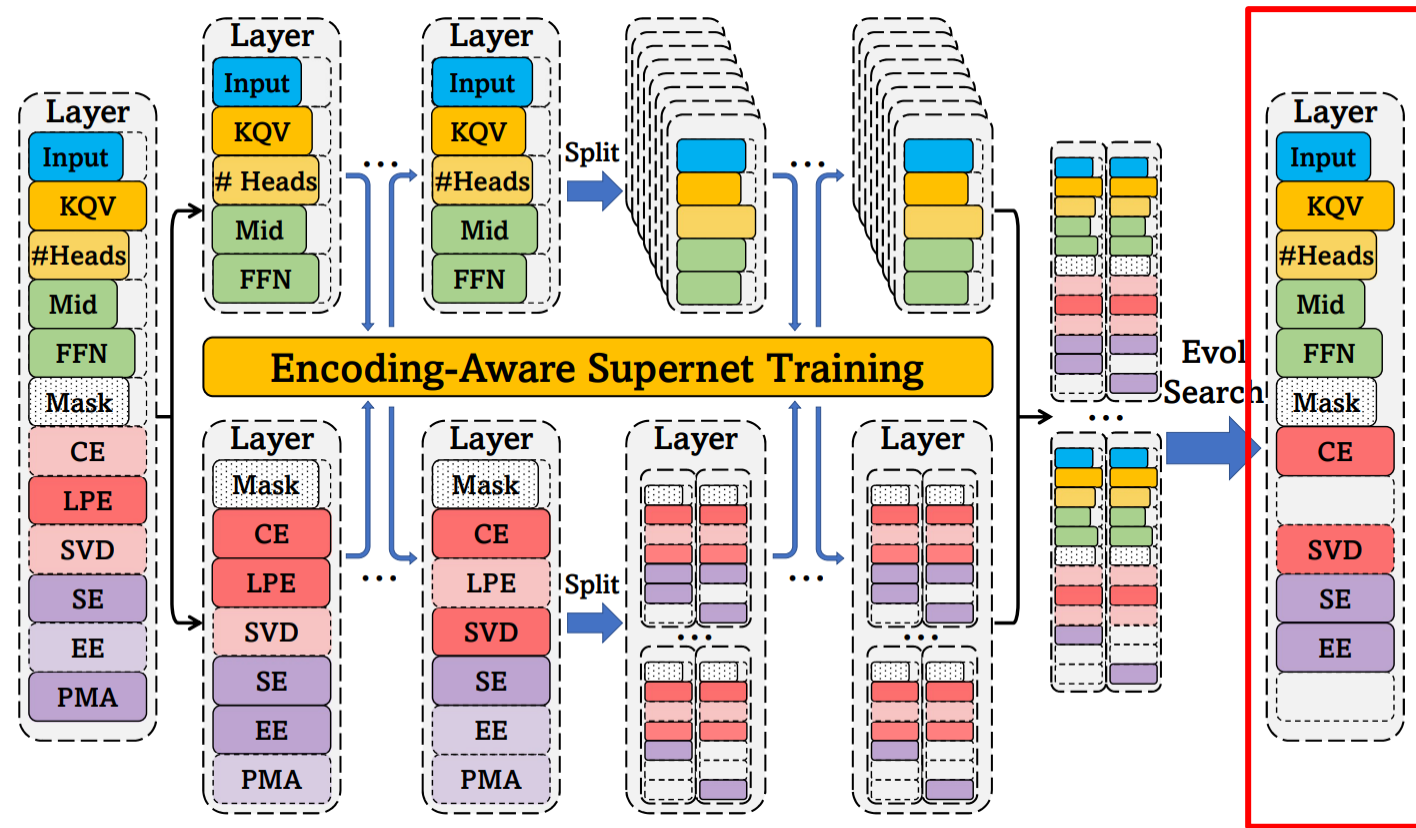
- ① Train Supernet
- ② Split Supernet into Subnets
- ③ Train Subnets

# Encoding-Aware Supernet Training



- ① Train Supernet
- ② Split Supernet into Subnets
- ③ Train Subnets
- ④ Evaluate performance using Subnets

# Encoding-Aware Supernet Training



- ① Train Supernet
- ② Split Supernet into Subnets
- ③ Train Subnets
- ④ Evaluate performance using Subnets
- ⑤ Find the best architecture

# Dataset Statistics

| Dataset  | #Graph | #Class | #Avg. Nodes | #Avg. Edges | # Node Feature | # Edge Feature |
|----------|--------|--------|-------------|-------------|----------------|----------------|
| COX2_MD  | 303    | 2      | 26.28       | 335.12      | 7              | 5              |
| BZR_MD   | 306    | 2      | 21.3        | 225.06      | 8              | 5              |
| PTC_FM   | 349    | 2      | 14.11       | 14.48       | 18             | 4              |
| DHFR_MD  | 393    | 2      | 23.87       | 283.01      | 7              | 5              |
| PROTEINS | 1,133  | 2      | 39.06       | 72.82       | 3              | 0              |
| DBLP     | 19,456 | 2      | 10.48       | 19.65       | 41,325         | 3              |

| Dataset      | #Graph | #Class | #Avg. Nodes | #Avg. Edges | # Node Feature | # Edge Feature |
|--------------|--------|--------|-------------|-------------|----------------|----------------|
| OGBG-MolBACE | 1,513  | 2      | 25.51       | 27.47       | 9              | 3              |
| OGBG-MolBBBP | 2,039  | 2      | 34.09       | 36.86       | 9              | 3              |
| OGBG-MolHIV  | 41,127 | 2      | 24.06       | 25.95       | 9              | 3              |

# Experimental Results

- Our method shows the best performance on various datasets

| Dataset      | COX2_MD                       | BZR_MD                        | PTC_FM                       | DHFR_MD                      | PROTEINS                     | DBLP                         |
|--------------|-------------------------------|-------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| GIN          | 45.82 <sub>14.35</sub>        | 59.68 <sub>14.65</sub>        | 57.87 <sub>8.86</sub>        | 62.88 <sub>8.26</sub>        | 73.76 <sub>4.61</sub>        | 91.18 <sub>0.42</sub>        |
| DGCNN        | 54.81 <sub>18.51</sub>        | 62.74 <sub>20.59</sub>        | 62.17 <sub>3.62</sub>        | 63.89 <sub>5.91</sub>        | 72.68 <sub>3.75</sub>        | 91.57 <sub>0.54</sub>        |
| DiffPool     | 51.45 <sub>14.28</sub>        | 65.01 <sub>14.74</sub>        | 60.16 <sub>5.87</sub>        | 61.06 <sub>9.42</sub>        | 73.31 <sub>3.75</sub>        | OOT                          |
| GraphSAGE    | 49.59 <sub>12.80</sub>        | 57.43 <sub>13.50</sub>        | 64.17 <sub>3.28</sub>        | 66.92 <sub>2.35</sub>        | 67.19 <sub>6.97</sub>        | 51.01 <sub>0.02</sub>        |
| Graphormer   | 56.39 <sub>15.03</sub>        | 63.94 <sub>12.58</sub>        | 64.88 <sub>7.58</sub>        | 64.88 <sub>7.58</sub>        | 75.29 <sub>3.10</sub>        | 89.36 <sub>2.31</sub>        |
| GT(ours)     | 54.44 <sub>16.84</sub>        | 63.33 <sub>11.67</sub>        | 64.18 <sub>2.60</sub>        | 65.68 <sub>5.64</sub>        | 73.94 <sub>3.78</sub>        | 90.67 <sub>1.01</sub>        |
| AutoGT(ours) | <b>59.72</b> <sub>23.26</sub> | <b>65.92</b> <sub>10.00</sub> | <b>65.60</b> <sub>3.71</sub> | <b>68.22</b> <sub>5.02</sub> | <b>77.17</b> <sub>3.40</sub> | <b>91.66</b> <sub>0.79</sub> |

# Experimental Results

- Our method shows the best performance on various datasets
- Our encoding-aware supernet training strategy indeed improves performance

| Dataset             | OGBG-MolHIV                  | OGBG-MolBACE                 | OGBG-MolBBBP                 |
|---------------------|------------------------------|------------------------------|------------------------------|
| GIN                 | 71.11 <sub>2.57</sub>        | 70.42 <sub>4.78</sub>        | 63.37 <sub>1.81</sub>        |
| DGCNN               | 69.97 <sub>2.16</sub>        | 75.62 <sub>2.64</sub>        | 60.92 <sub>1.78</sub>        |
| DiffPool            | 74.58 <sub>1.71</sub>        | 73.87 <sub>4.50</sub>        | 66.68 <sub>6.08</sub>        |
| GraphSAGE           | 67.82 <sub>3.67</sub>        | 72.91 <sub>1.24</sub>        | 64.19 <sub>3.50</sub>        |
| Graphormer          | 71.89 <sub>2.66</sub>        | 76.42 <sub>1.67</sub>        | 66.52 <sub>0.74</sub>        |
| <b>AutoGT(ours)</b> | <b>74.95</b> <sub>1.02</sub> | <b>76.70</b> <sub>1.42</sub> | <b>67.29</b> <sub>1.46</sub> |

| Method                               | Accuracy                                       |
|--------------------------------------|--|
| <b>One-Shot<br/>Positional-Aware</b> | 75.92 <sub>3.10</sub><br>76.19 <sub>3.42</sub> |
| <b>AutoGT</b>                        | 77.17 <sub>3.40</sub>                          |

# Conclusions

- We propose AutoGT, the first neural architecture search framework for graph Transformer
- We design a unified search space for graph Transformer, and propose a novel encoding-aware supernet training strategy
- Extensive experiments demonstrate its effectiveness and wide applicability




# Poster

- Time: 4:30 p.m. ~ 6:30 p.m.
- Place: MH1-2-3-4 #122

## AutoGT: Automated Graph Transformer Architecture Search

Zizhao Zhang, Xin Wang, Chaoyu Guan, Ziwei Zhang, Haoyang Li, Wenwu Zhu  
Tsinghua University

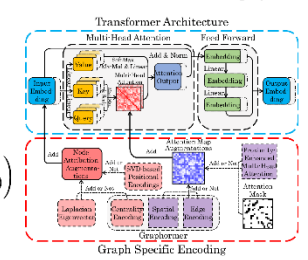


### ➤ Motivation

- Graph Transformer gained success.
- *Human labor and expert knowledge* is needed for designing proper *Transformer architecture* and *graph specific encoding*.
- However, current Transformer automatic design works focus on non-graph data *without considering graph encoding*.
- How to design graph Transformer automatically?

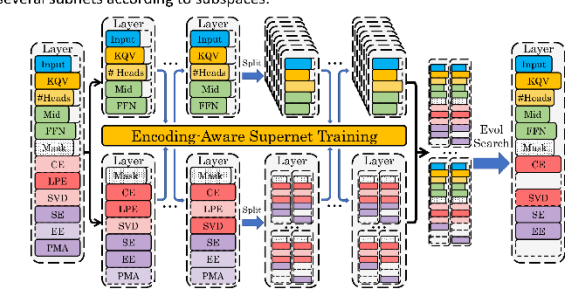
### ➤ Method

- We propose **Automated Graph Transformer (AutoGT)** for *Graph Transformer Neural Architecture Search*.
- 1) **Graph Transformer Search Space**: based on a unified graph Transformer formulation, including both candidate Transformer architectures and various graph encodings



- Unified graph Transformer formulation:
- **Node Attribution Augmentation**
- $\mathbf{H}_{aug}^{(l)} = \mathbf{H}^{(l)} + Enc_{node}(G)$
- **Attention Map Augmentation**
- $A_{h,aug}^{(l)} = \text{softmax}\left(\frac{Q_h^{(l)} K_h^{(l)T}}{\sqrt{d}} + Enc_{map}(G)\right)$

- 2) **Encoding-Aware Supernet Training**: train a single supernet first and split into several subnets according to subspaces.



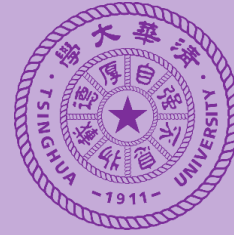
### ➤ Experiment

| Dataset    | COX2_MD                       | BZR_MD                        | PTC_FM                        | DHFR_MD                       | PROTEINS                      | DBLP                          |
|------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| GIN        | 45.82 <sub>±4.35</sub>        | 59.68 <sub>±4.05</sub>        | 57.87 <sub>±8.86</sub>        | 62.88 <sub>±2.26</sub>        | 73.76 <sub>±4.61</sub>        | 91.18 <sub>±2.42</sub>        |
| DGCNN      | 54.81 <sub>±4.51</sub>        | 62.74 <sub>±2.59</sub>        | 62.17 <sub>±3.27</sub>        | 63.89 <sub>±3.01</sub>        | 72.68 <sub>±3.75</sub>        | 91.57 <sub>±3.51</sub>        |
| DHPPool    | 51.45 <sub>±2.28</sub>        | 65.01 <sub>±1.74</sub>        | 60.16 <sub>±5.87</sub>        | 61.10 <sub>±6.42</sub>        | 73.31 <sub>±3.75</sub>        | OOPT                          |
| GraphSAGE  | 49.59 <sub>±2.80</sub>        | 57.43 <sub>±1.50</sub>        | 64.17 <sub>±3.28</sub>        | 66.92 <sub>±3.25</sub>        | 67.19 <sub>±3.47</sub>        | 51.01 <sub>±0.02</sub>        |
| Graphormer | 56.39 <sub>±5.01</sub>        | 63.94 <sub>±2.58</sub>        | 64.88 <sub>±2.58</sub>        | 64.88 <sub>±2.58</sub>        | 75.29 <sub>±3.10</sub>        | 89.36 <sub>±1.31</sub>        |
| GTours     | 54.44 <sub>±8.81</sub>        | 63.33 <sub>±1.67</sub>        | 64.18 <sub>±3.60</sub>        | 65.68 <sub>±3.84</sub>        | 73.94 <sub>±3.78</sub>        | 90.67 <sub>±1.01</sub>        |
| AutoGTours | <b>59.72</b> <sub>±3.26</sub> | <b>65.92</b> <sub>±2.00</sub> | <b>65.60</b> <sub>±2.71</sub> | <b>68.22</b> <sub>±2.02</sub> | <b>77.17</b> <sub>±3.09</sub> | <b>91.66</b> <sub>±0.79</sub> |

| Dataset    | OGBG-MolHIV                   | OGBG-MolBACE                  | OGBG-MolBBBP                  |
|------------|-------------------------------|-------------------------------|-------------------------------|
| GIN        | 71.11 <sub>±2.57</sub>        | 70.42 <sub>±2.18</sub>        | 63.37 <sub>±1.81</sub>        |
| DGCNN      | 69.97 <sub>±2.10</sub>        | 75.62 <sub>±2.64</sub>        | 60.92 <sub>±1.78</sub>        |
| DHPPool    | 74.58 <sub>±1.71</sub>        | 73.87 <sub>±1.00</sub>        | 66.68 <sub>±0.88</sub>        |
| GraphSAGE  | 67.82 <sub>±2.87</sub>        | 72.91 <sub>±2.14</sub>        | 64.19 <sub>±3.50</sub>        |
| Graphormer | 71.89 <sub>±3.86</sub>        | 76.42 <sub>±1.67</sub>        | 66.52 <sub>±2.74</sub>        |
| AutoGTours | <b>74.95</b> <sub>±3.22</sub> | <b>76.70</b> <sub>±1.42</sub> | <b>67.29</b> <sub>±1.46</sub> |

# Thanks!



Zizhao Zhang, Tsinghua University  
zzz22@mails.tsinghua.edu.cn

---

AutoGT: Automated Graph Transformer  
Architecture Search

Poster Location: MH1-2-3-4 **#122**



**ICLR**  
International Conference On  
Learning Representations