Triplet Interaction Improves Graph Transformers

Accurate Molecular Graph Learning with Triplet Graph Transformers

By Md Shamim Hussain, Mohammed J. Zaki

(Rensselaer Polytechnic Institute)

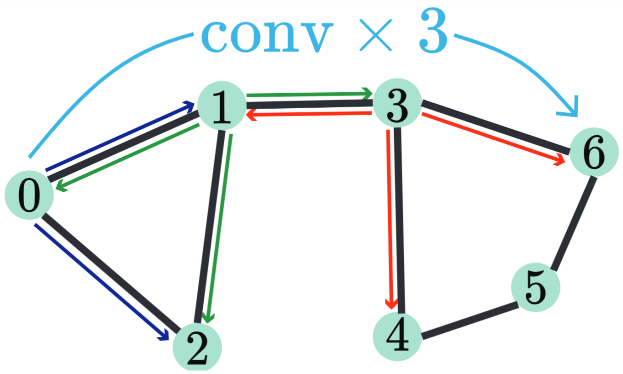
and

Dharmashankar Subramanian

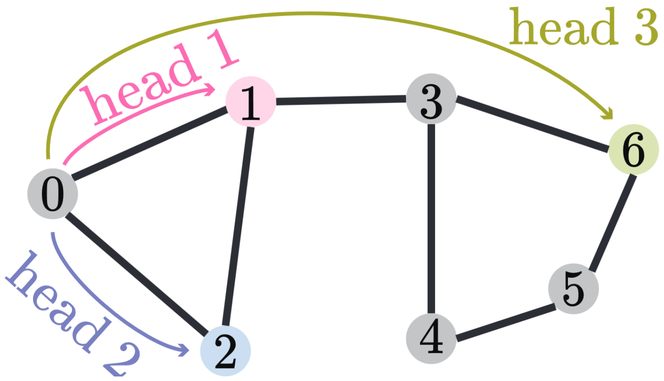
(IBM Thomas J. Watson Research Center)

Graph Transformers vs. GCNs

- Long-distance, dynamic interaction
  - Not limited to neighbors
  - Attention weights are determined by the network
- Limitations of hand-crafted encodings/features
  - Positional-encoding based GTs (e.g., Graphormer)
    - Structural understanding ≈ as good as the used positional encoding
  - Geometric GTs (e.g., Equiformer)
    - Geometric understanding ≈ as good as the used geometric features
- **Goal: let the network form its <u>own</u> geometry / structure**



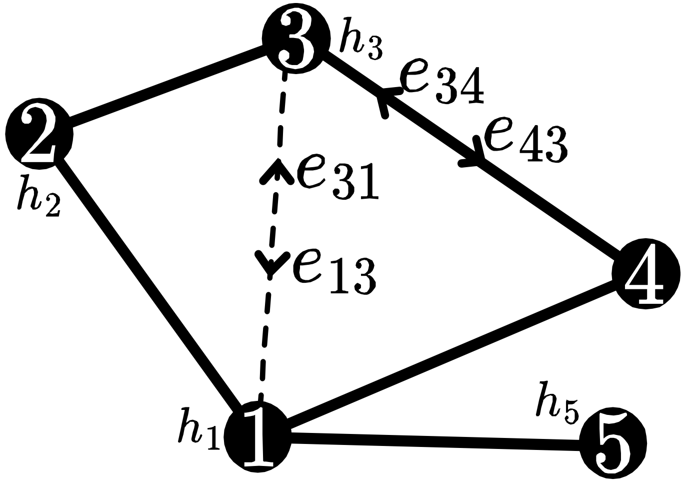
(a) Convolution



(b) Self-attention

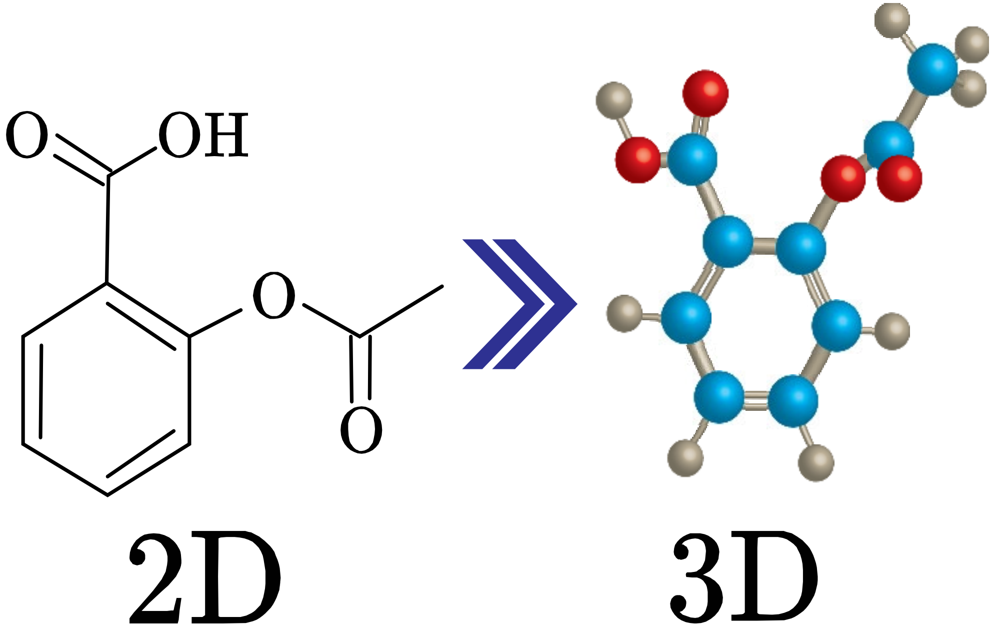
Graph Structure and Pair Representations

- Pairs → directed existing/non-existing edges
  - e.g., 3→4, 4→3, 3→1, 1→3
- For graphs pair representations (e_{ij}) can be as important as node representations (h_i)
  - Allow the structure of the graph to evolve over layers
  - Refine structure/topology internally in case of inaccuracies
  - Directly perform pair related task
    - Link prediction
    - Edge classification
    - Distance Prediction
- EGT (Edge-augmented Graph Transformer)
  - Make pairs (2-tuple) first class citizen, just like nodes
  - Break free of the input graph topology
  - **Limitation: only 2nd-order interaction**



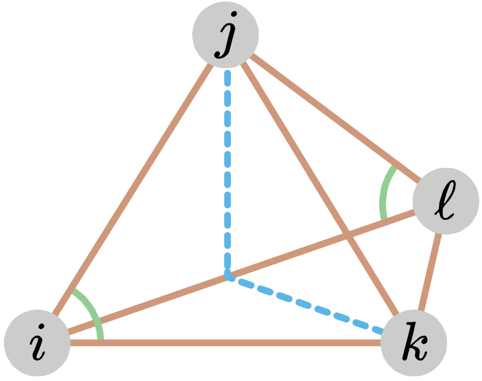
3D Molecular Geometry

- 2D
  - Bonds + Atoms (i.e., chemical formula)
- 3D
  - Coordinates
  - Often interatomic distances is enough
- 3D shape directly dictates molecular property
  - But costly to compute (QM simulation required)
- **Train a model: 2D→3D**
  - **i.e., predict the molecular geometry**



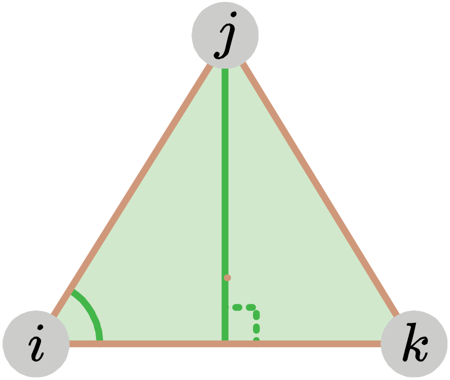
K-order interaction vs. K-order features

- **2nd Order (i, j): Pairwise distances**
- **3rd Order (i, j, k): Angles, area of triangles, etc.**
- **4th Order (i, j, k, ℓ): Dihedral angles, volume of tetrahedrons, etc.**



We need either higher order interactions or higher order features for full geometric understanding

- **Crucial for 3D geometry prediction**
- **Our contribution: Third order interaction**
  - Pairs $(i, j), (j, k), (i, k)$ within the 3-tuple (i, j, k)



Why higher order interactions?

- Using higher order features such as angles implies
  - An initial estimate of geometry is required
  - Features are only as accurate as the estimate
  - Specialized for geometric graphs only
- Using higher order interactions implies
  - No estimate of geometry is required, a simple graph is enough
  - The network can form representations that are more refined than the initial estimate
  - Applicable to both geometric and non-geometric graphs

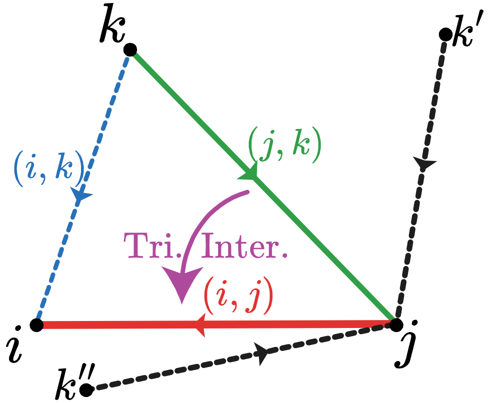
- **Our work improves geometric expressivity without losing generality**
- **We break free from the inaccuracies of the initial geometry (when given)**

Our Contribution

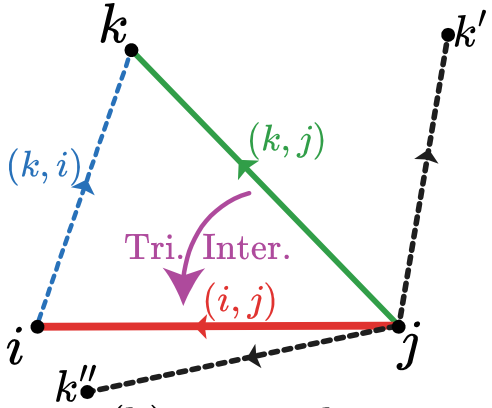
Triplet Graph Transformer

Triplet Interaction within a 3-tuple (i, j, k)

- Participants (inward): $(i, j), (j, k), (i, k)$
- (i, j) gathers information from (j, k)
- Without triplet interaction
  - $(j, k) \rightarrow j \rightarrow (i, j)$
  - Bottleneck at node j
- With triplet interaction
  - $(j, k) \xrightarrow{(i,k)} (i, j)$
- (i, k) also participates in this process
- Similarly: outward update



(a) inward



(b) outward

Triplet Interaction within a 3-tuple (i, j, k)

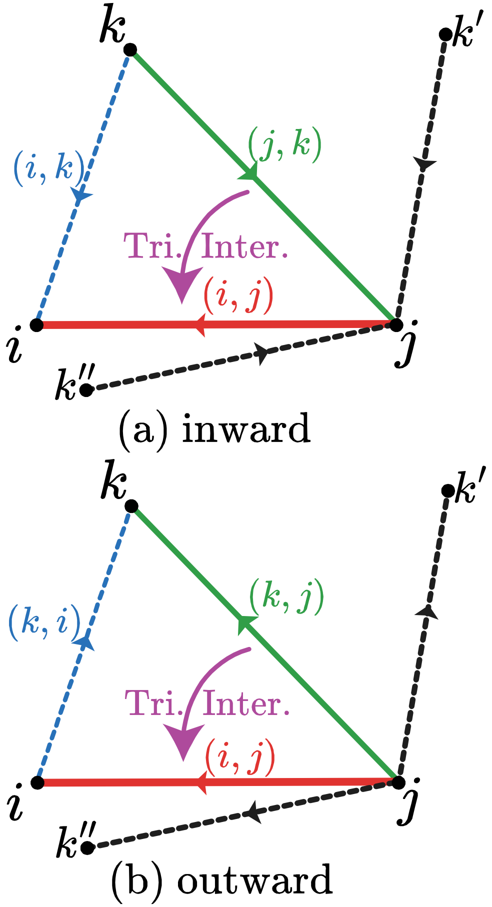
- Two mechanisms
  - Triplet Attention (TGT-At)

$$\mathbf{o}_{ij}^{\text{in}} = \sum_{k=1}^{N} a_{ijk}^{\text{in}} \mathbf{v}_{jk}^{\text{in}} \quad ; \quad a_{ijk}^{\text{in}} = \text{softmax}_k\left(\frac{1}{\sqrt{d}} \mathbf{q}_{ij}^{\text{in}} \cdot \mathbf{p}_{jk}^{\text{in}} + b_{ik}^{\text{in}}\right) \times \sigma(g_{ik}^{\text{in}})$$

    - More expressive, $O(N^3)$
  - Triplet Aggregation (TGT-Ag)

$$\mathbf{o}_{ij}^{\text{in}} = \sum_{k=1}^{N} a_{ik}^{\text{in}} \mathbf{v}_{jk}^{\text{in}} \; ; \qquad a_{ik}^{\text{in}} = \text{softmax}_k(b_{ik}^{\text{in}}) \times \sigma(g_{ik}^{\text{in}})$$

    - More efficient, $O(N^{2.37})$

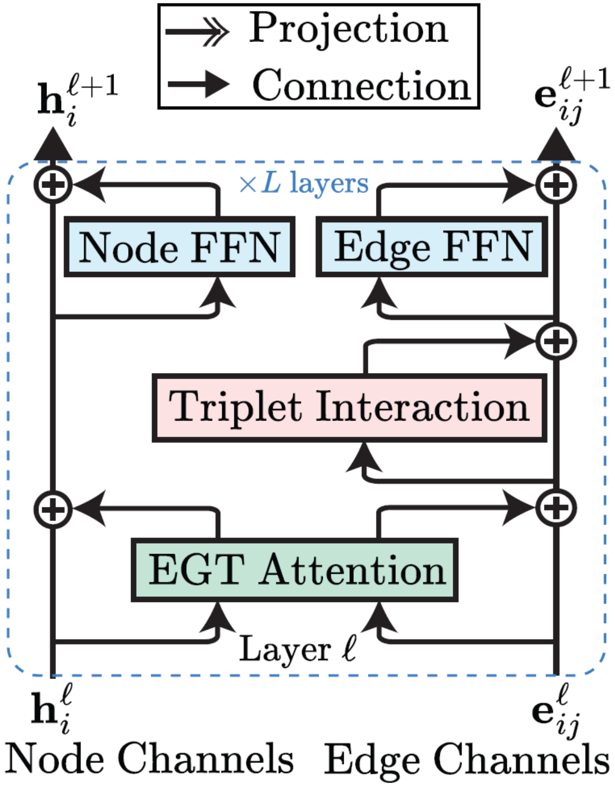


(a) inward

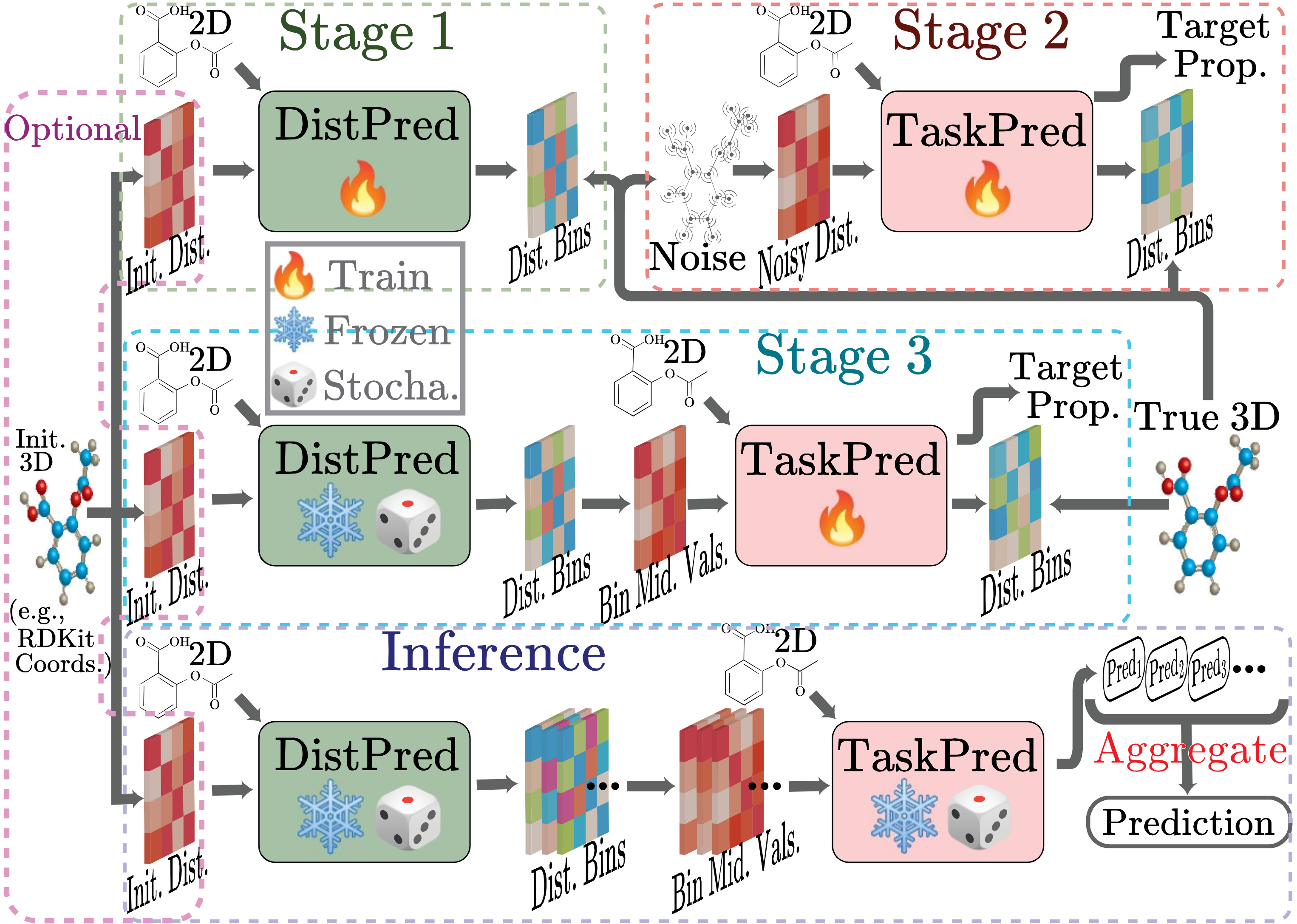
(b) outward

Network Architecture

- EGT (our previous work)
  - Node representations → Node channels
  - Pair representations → Edge channels
  - Only 2nd order interactions
- Triplet interaction (TGT)
  - Update pair representations based on each other
  - 3rd order interactions
- Pair embeddings are directly used for predicting (binned) pairwise atomic distances
- Useful for other geometric tasks as well (e.g., Traveling Salesman Problem)



Training and Inference

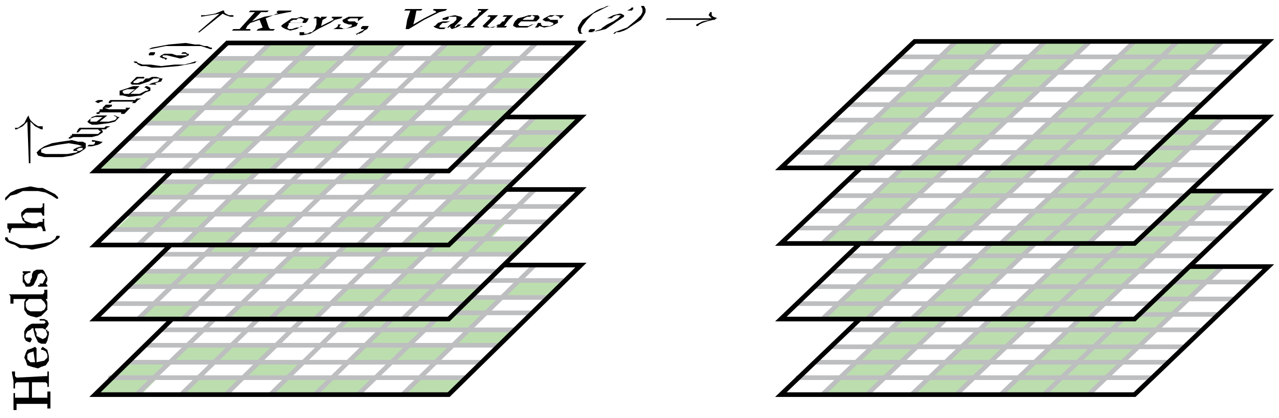


Other Contributions

- Locally smoothed 3D noise for pretraining the task predictor

$$\mathbf{r}'_i = \mathbf{r}_i + \sum_{j=1}^{N} e^{-\frac{\|\mathbf{r}_i - \mathbf{r}_j\|}{\nu}} \mathbf{u}_j; \text{ where } \mathbf{u}_j \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$$

- Source dropout: Stronger regularization for graph transformers



(a)Attention Dropout **(b)Source Dropout**

Key Contributions

- Triplet Graph Transformer (TGT)
  - Novel triplet interaction mechanisms for direct pair-to-pair communication
  - Accurately models geometry in graphs
- Two-stage model
  - Separate distance and task predictors
  - Eliminates need for initial 3D coordinates
- Three-stage training procedure and stochastic inference
  - Significantly improves training efficiency and predictive performance
- TGT for graph learning
  - Traveling Salesman Problem

Quantum Chemistry (Large-scale)

- Predict in absence of ground truth 3D

(a) PCQM4Mv2 (Molecules)

Model	MAE↓ (meV)
EGT	86.2
Transformer-M	78.2
Uni-Mol+ (+RDKit)	70.5
TGT-At	69.8
TGT-At (+RDKit)	**68.3**

(b) OC20 IS2RE (Crystals)

	MAE↓	EwT↑
Model	**(meV)**	**(%)**
SphereNet	618.8	3.32
EquiFormer	466	5.66
Uni-Mol+	**414.3**	8.23
TGT-At	**414.7**	**8.3**

Quantum Chemistry (Transfer Learning)

- Ground truth 3D is provided
- Fine-tuned from PCQM4Mv2

(a) QM9 (Molecules)

Method	μ	α	ϵ_H	ϵ_L	$\Delta\epsilon$	**ZPVE**	C_v
3D Infomax	0.034	0.075	29.8	25.7	48.8	1.67	0.033
SphereNet	0.025	0.053	22.8	18.9	31.1	**1.12**	0.024
Equiformer	**0.011**	0.046	15	14	30	1.26	0.023
Transformer-M	0.037	0.041	17.5	16.2	27.4	1.18	0.022
TGT-Ag	0.025	**0.040**	**9.9**	**9.7**	**17.4**	1.18	**0.020**

Molecular Property (Transfer Learning)

- Non-quantum property prediction and drug discovery
- Use frozen (not finetuned) distance predictor from PCQM4Mv2
  - Provides more accurate 3D information than RDKIT

MOLPCBA (Property)

Model	AP↑ (%)
PHC-GNN	29.47
Graphormer	31.40
TGT-Ag+RDKit	31.44
TGT-Ag+DP	**31.67**

LITPCBA (Drugs)

Model	ROC-AUC↑ (%)
GEM	78.4
GEM-2+RDKit	**81.5**
EGT+RDKit	81.2
EGT+DP	**81.5**

Traveling Salesman Problem(Graph Learning)

- Points on a 2D plane
- Edge classification

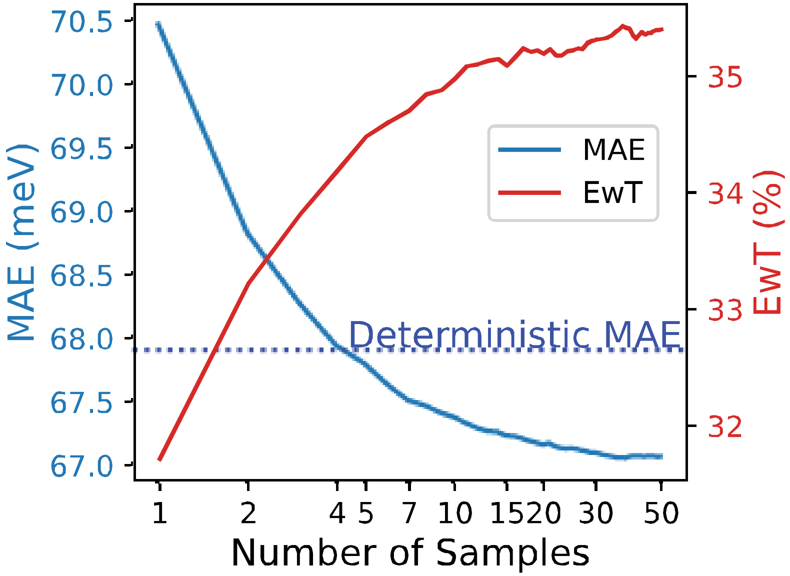
TSP

Model	**F1↑ (%)**
GatedGCN	83.8
ARGNP	85.5
EGT	85.3
TGT-Agx4	**87.1**

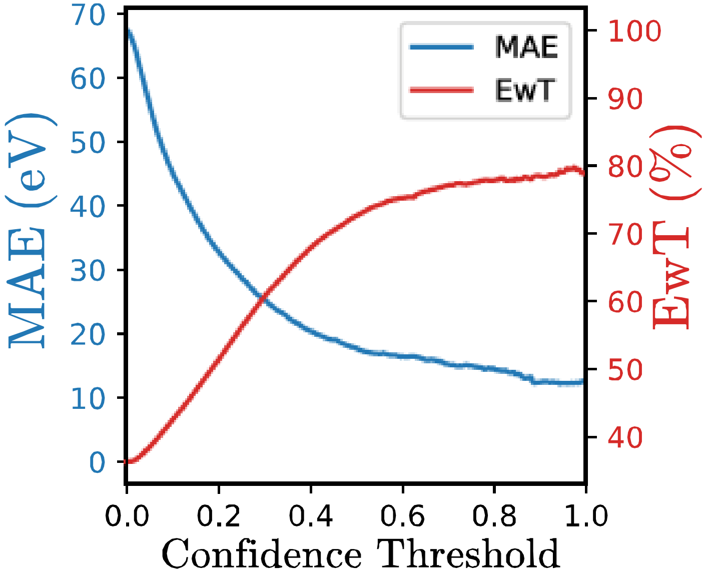
- Demonstrate the generality of our model

Merits of Stochastic Inference

- Outperforms deterministic inference with only ~4 samples
- Higher confidence implies higher accuracy



#Samples vs Performance



Confidence vs Performance

Future Work

- Explore use of triplet interaction for other graph learning tasks
    - Molecule and conformation generation
    - Link prediction
    - Combinatorial Problem
    - Self-supervised/semi-supervised/generative graph learning
- Improve compute and memory efficiency of triplet interaction
    - Sparsity
    - Linearity

Thank You

Please check out our paper for more details.

Paper: https://arxiv.org/abs/2402.04538

Implementation: https://github.com/shamim-hussain/tgt