Deprecation Notice

This technical report has been deprecated.

Please refer to our **paper on arXiv:**
`https://arxiv.org/abs/2402.04538`
for the most recent version titled –

**"Triplet Interaction Improves Graph Transformers:
Accurate Molecular Graph Learning with Triplet Graph Transformers"**

IMPROVED QUANTUM CHEMICAL PROPERTY PREDICTION BY TRIANGULAR ATTENTION AND INCORPORATION OF 3D STRUCTURE

Md Shamim Hussain & Mohammed J. Zaki
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
hussam4@rpi.edu, zaki@cs.rpi.edu

Dharmashankar Subramanian
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA
dharmash@us.ibm.com

ABSTRACT

Quantum chemical property prediction is a challenging task due to the intricate nature of the underlying physics. In this study, we present a series of key insights that enhance the efficacy of transformer architectures, specifically the Edge-augmented Graph Transformer (EGT), in the task of predicting the HOMO-LUMO gap on the PCQM4Mv2 large-scale molecular graph dataset by OGB. Our investigation reveals that integrating a precise estimation of the three-dimensional (3D) molecular structure, particularly inter-atomic distances, is pivotal for accurate HOMO-LUMO gap estimation. To address this, we employ a two-stage architecture comprising a distance predictor and a gap predictor. Additionally, we identify the substantial benefits of third-order interactions among atoms, referred to as triangular attention, in achieving precise structure prediction and, consequently, accurate HOMO-LUMO gap estimation. Furthermore, we introduce several innovations in training and inference techniques, which help us achieve state-of-the-art results on the PCQM4Mv2 dataset.

1 INTRODUCTION

In recent times, there has been notable progress in the realm of deep learning for molecular property prediction, primarily driven by the advent of graph transformers Ying et al. (2021); Hussain et al. (2022); Park et al. (2022); Liu et al. (2022). Graph transformers utilize global self-attention to facilitate dynamic information exchange among atomic representations, commonly referred to as node embeddings. The Edge-augmented Graph Transformer (Hussain et al., 2022) first introduced edge channels, which pairwise embeddings to evolve over layers, alongside node embeddings. These channels are especially useful for pairwise predictions, such as inter-atomic distance prediction, which are traditionally performed based on node embeddings only. GEM-2 (Liu et al., 2022) further extended upon EGT to include higher-order interactions among atoms, which showed promising results at the cost of increased computational complexity.

GEM-2 was the first to incorporate 3D structural information in the form of RDKit coordinates – which are cheap to compute. However, RDKit coordinates are much less accurate than the DFT coordinates which consider the full underlying physics and can be considered the gold standard when it comes to molecular structure. Despite the superior accuracy of DFT coordinates, their computation involves a resource-intensive DFT simulation, which contradicts the objective of data-driven molecular property prediction to bypass extensive simulations. Consequently, the PCQM4Mv2 dataset exclusively incorporates DFT coordinates in the training split, necessitating inference without these coordinates. Transformer-M (Luo et al., 2022) first incorporated DFT coordinates in the training stage in a multi-task setting where the transformer simultaneously learns to make predictions in the presence and absence of DFT coordinates. The multi-task training regularizes the model to effectively make predictions in the absence of structural information. Later works like GPS++ (Masters et al., 2022) also followed this training method. UniMol+ (Lu et al., 2023) first proposed a prediction scheme where the RDKit coordinates are iteratively refined by the network before making the final HOMO-LUMO gap prediction based on a more accurate structure.

1.1 Our Contribution

In this work, we make some important observations. Firstly, achieving precise predictions of the HOMO-LUMO gap requires providing the gap predictor network with an accurate structural approximation. Our experiments reveal a strong correlation between the accuracy of predicting the structure, specifically inter-atomic distances, and the accuracy of predicting the HOMO-LUMO gap. A near-perfect estimation of the former would, in turn, render the estimation of the latter trivial. However, lacking such a precise estimation of the structure (i.e., DFT coordinates) during inference, we turn to a distance predictor network. This network learns to produce a reliable estimation of inter-atomic distances, which the gap predictor then utilizes. This differs from UniMol+ (Lu et al., 2023), which employs the same network for both structural prediction (refinement) and HOMO-LUMO gap prediction.

Our second key observation underscores the necessity of incorporating DFT structure during the training phase of the gap predictor. However, given the absence of a perfect structural estimate at inference, it is imperative to ensure the network's adaptability to imperfect structural input. Our proposed solution involves a two-stage training process for the gap predictor. In the first stage, the predictor is fed distances derived from noisy DFT coordinates, followed by a fine-tuning stage where it receives distance estimates from the distance predictor network. Throughout both stages, the gap predictor is tasked with denoising imperfect distances in the input. The inclusion of noisy DFT coordinates and the denoising objective proves to be an effective regularization method during the pretraining stage, akin to Transformer-M (Luo et al., 2022). However, unlike Transformer-M, we do not require the gap predictor to make predictions in the absence of any structural information. Instead, we leverage the distance predictor's output to fine-tune the gap predictor and make predictions during inference.

Finally, we make an important innovation in the network architecture. Our approach predominantly embraces the Edge-augmented Graph Transformer (EGT) architecture, as outlined in (Hussain et al., 2022). Notably, we leverage the edge channels to directly update pairwise embeddings, ultimately predicting distances in both the distance predictor and the denoising head of the gap predictor. This departure from conventional practices, where structural predictions rely on node embeddings, sets our work apart. Additionally, we introduce multihead triangular attention to empower the network with the ability to capture third-order interactions among atoms. This enhancement proves particularly valuable in distance prediction, as the distance between a pair of atoms is intricately influenced by the presence of other atoms in the molecule. Our approach represents an advancement over earlier methodologies such as GEM-2 (Liu et al., 2022), which relies on **axial** attention, and UniMol+ (Lu et al., 2023), incorporating a less expressive triangular **update**. Our work combines the expressiveness of both third-order axial attention and triangular update to propose triangular attention, which is more expressive than both.

2 Methodologies

2.1 Multihead Triangular Attention

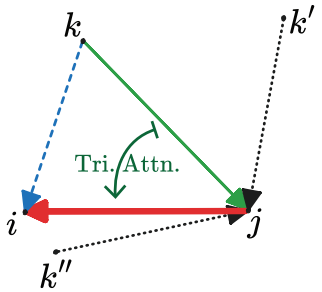

Figure 1: Triangular attention allows for direct communication between two adjacent pairs of nodes, i.e., (i, j) and (j, k), alleviating the bottleneck at the junction node j. This interaction also takes into account the third arm of the triangle, i.e., the pair (i, k).

While the Edge-augmented Graph Transformer (EGT) facilitates unconstrained long-range pairwise interactions and maintains pairwise embeddings e_{ij} for every node pair (i, j), it lacks an inherent mechanism for modeling direct interactions between two pairs. This limitation arises from updating the pairwise embeddings e_{ij} solely based on the node embeddings h_i and h_j. While this choice ensures a computational complexity of $O(N^2)$, similar to a typical transformer, it constrains the model's expressiveness, particularly in capturing three-dimensional geometry. Enabling completely unconstrained pair interactions would escalate computational complexity to $O(N^4)$, rendering it impractical for large molecules.

A judicious compromise involves identifying a crucial bottleneck in adjacent pairs, namely those sharing a common node. Illustrated in Figure 1, this bottleneck is exemplified by the pair (i, j) only being able to interact with the pair (j, k) through the intermediary node j. However, the pair (j, k) contends with all other pairs (e.g., (j, k'), (j, k''), etc.) that share the node j in transmitting information to neighboring pairs, creating a bottleneck. This bottleneck is mitigated by allowing direct interaction between pairs (i, j) and (j, k) without necessitating involvement of the node j. The linear arrows between a pair of nodes in Figure 1 denote the flow of information in the node channels, whereas the curved arrow represents the flow of information between pairwise embeddings due to triangular attention. This interaction also considers the third arm of the triangle, i.e., the pair (i, k). This encapsulates the essence of multihead triangular attention, with a computational complexity of $O(N^3)$. While still resource-intensive, this represents a significant improvement over the $O(N^4)$ complexity of unconstrained interaction between pairs and, at the same time, significantly boosts the expressivity of the model.

For a particular pair of nodes (i, j), the triangular attention is computed as follows

$$\boldsymbol{o}_{ij} = \sum_{k=1}^{N} w_{ijk} \boldsymbol{v}_{jk} \tag{1}$$

Where $\boldsymbol{v}_{jk}$ is derived from a learnable projection of the pairwise embedding e_{jk} and w_{ijk} is the attention weight assigned to the pair (j, k) by the pair (i, j). $\boldsymbol{o}_{ij}$ is used to update the pairwise embedding of the pair (i, j). This is done for multiple attention heads, similar to the traditional multi-head attention mechanism. At this point, we can point out the difference between triangular attention and the triangular update mechanism used in UniMol+ (Lu et al., 2023). In UniMol+, weights are unnormalized and the values being aggregated are scalers (which effectively results in no distinction between weight and value), whereas in triangular attention, the weights are normalized and the values are vectors. However, the most important distinction is that the weights are dependent on the junction node j in triangular attention; whereas, in UniMol+, the weights are only dependent on the pair (i, k) and independent of j. This can limit the expressiveness of the model since it does not let the pairs (i, j) decide how much importance to assign to the pairs arriving at junction node j. While this does cost some memory and compute because we need to fully materialize a $O(N^3)$ weight matrix, we found the gains in expressiveness to be worth the additional cost. Note that, the cost of UniMol+ is also close to $O(N^3)$ due to multiplication between two $O(N^2)$ matrices.

The weights w_{ijk} are computed as follows

$$\tilde{w}_{ijk} = \exp\left(\frac{1}{\sqrt{d}} \boldsymbol{q}_{ij}^T \boldsymbol{p}_{jk} + b_{ik}\right) \tag{2}$$

$$w_{ijk} = \frac{\tilde{w}_{ijk}}{\sum_{k'=1}^{N} \tilde{w}_{ijk'}} \times \sigma(g_{ik}) \tag{3}$$

Here, q_{ij} is the query vector derived from the pairwise embedding e_{ij}, p_{jk} are the key vectors formed from the pairwise embeddings e_{jk}, and b_{ik}, g_{ik} are scalers derived from the pairwise embeddings e_{ik}. The third arm of the triangle (i, k) participates in the process by providing these bias terms b_{ij} and the gating term g_{ik}. The gating term, which is added in the same spirit as the attention process in the node channels in EGT, is not mandatory but improves the performance of the model. Finally, σ is the sigmoid function, and d is the dimension of the query and key vectors.

The triangular attention is computed for all pairs of nodes (i, j) in each attention head. Another parallel update is performed by reversing/transposing the direction of the pairs, i.e., (j, i) and (k, j) and third arm (k, i). This allows for information to flow in the opposite direction, which is important for modeling three-dimensional geometry. The final update is a sum of the two updates.

Note that multihead triangular attention is an improvement upon the third-order axial attention in GEM-2 (Liu et al., 2022) because it takes the third arm of the triangle, i.e., the pair (i, k), into account. At the same time, it is significantly more expressive than the triangular update in UniMol+ (Lu et al., 2023) because the assigned weights come from the scaled-dot product-based attention mechanism and consider each junction node j separately.

2.2 Network Architecture

As mentioned before, we have two networks - a distance predictor and a gap predictor. We follow the same network architecture networks. We adopt the EGT architecture with some modifications. The most important addition we make is the multihead triangular attention block which is added to the edge channels, after the pairwise attention block and before the edge Feed Forward Network (FFN) block. The resulting architecture is shown in figure 2.

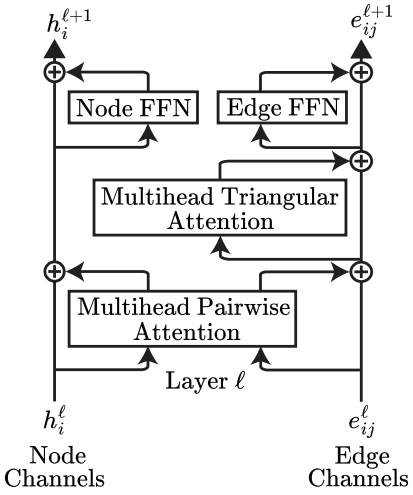

Figure 2: The network architecture.

Firstly, we removed the dot product clipping in the original work, which was found to be beneficial for training convergence. However, in our training setting, clipping was not found to be useful. Secondly, we used the GELU Hendrycks & Gimpel (2016) activation instead of ELU Clevert et al. (2015) because it leads to improved performance and better convergence.

We also changed the dropout patterns in the network. Before the residual connections, instead of performing traditional dropout, we used path dropout Huang et al. (2016). We also applied activation dropout to the Feed Forward Network blocks. We also introduce a new dropout pattern in the multihead (pairwise) attention block called source dropout.

Source Dropout: In the pairwise attention block, in the absence of any dropout all the nodes in the molecule are allowed to attend to all other nodes. To regularize these processes we introduce source dropout. In this process, we update all the nodes but they are allowed to attend to only a subset of the nodes. This differs from attention dropout Zehui et al. (2019) in that, this subset is kept fixed for all the nodes in all attention heads in a layer. That is, all the nodes serve as targets, but only a subset serves as sources of information. This is similar to SSA, introduced in (Hussain et al., 2023), but is more focused on regularization rather than efficiency. We achieve this by adding $-\infty$ to the input to the softmax function for the nodes with a given probability, which is the source dropout rate. The distinction between source dropout and attention dropout is illustrated in figure 3.

Input Distance Encoding and Distance Prediction Head: We used the same Gaussian kernel-based distance encoding as in (Luo et al., 2022) in the input. For distance prediction, we used a binned categorical predictor, where the distance is binned into a fixed number of equal-sized bins. The bins start at 0 Angstrom and end at 8 Angstrom. Any distance beyond 8 Angstrom is clipped. We found that this range of distances is sufficient for the HOMO-LUMO gap prediction. The distance predictor is not inherently symmetric, and we do not impose it in any way during training. Rather

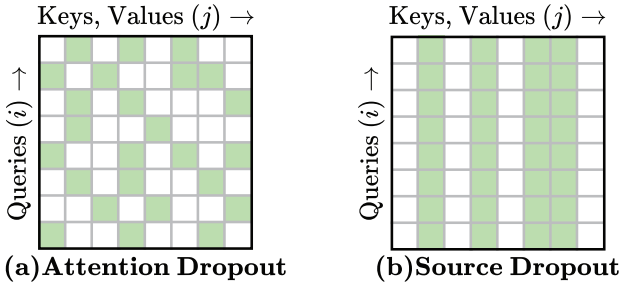

Figure 3: The distinction between source dropout and attention dropout.

during inference, the probabilities over the bins are averaged for (i, j) and (j, i) to ensure that the predicted distances are symmetric.

We pick the bin with the highest probability during inference. In order to convert this discrete distance into a continuous value, we choose the central value of the bin. We also ensure that the distances on the diagonal of the distance matrix (distance from an atom to itself) are zero. However, we do not in any way impose the distances to represent a three-dimensional structure.

REFERENCES

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pp. 646–661. Springer, 2016.

Md Shamim Hussain, Mohammed J Zaki, and Dharmashankar Subramanian. Global self-attention as a replacement for graph convolution. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 655–665, 2022.

Md Shamim Hussain, Mohammed J Zaki, and Dharmashankar Subramanian. The information pathways hypothesis: Transformers are dynamic self-ensembles. *arXiv preprint arXiv:2306.01705*, 2023.

Lihang Liu, Donglong He, Xiaomin Fang, Shanzhuo Zhang, Fan Wang, Jingzhou He, and Hua Wu. Gem-2: Next generation molecular property prediction network with many-body and full-range interaction modeling. *arXiv preprint arXiv:2208.05863*, 2022.

Shuqi Lu, Zhifeng Gao, Di He, Linfeng Zhang, and Guolin Ke. Highly accurate quantum chemical property prediction with uni-mol+. *arXiv preprint arXiv:2303.16982*, 2023.

Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. One transformer can understand both 2d & 3d molecular data. *arXiv preprint arXiv:2210.01765*, 2022.

Dominic Masters, Josef Dean, Kerstin Klaser, Zhiyi Li, Sam Maddrell-Mander, Adam Sanders, Hatem Helal, Deniz Beker, Ladislav Rampášek, and Dominique Beaini. Gps++: An optimised hybrid mpnn/transformer for molecular property prediction. *arXiv preprint arXiv:2212.02229*, 2022.

Wonpyo Park, Woong-Gi Chang, Donggeon Lee, Juntae Kim, et al. Grpe: Relative positional encoding for graph transformer. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation? *arXiv preprint arXiv:2106.05234*, 2021.

Lin Zehui, Pengfei Liu, Luyao Huang, Junkun Chen, Xipeng Qiu, and Xuanjing Huang. Dropattention: a regularization method for fully-connected self-attention networks. *arXiv preprint arXiv:1907.11065*, 2019.