

# ICLR25\_Rebuttal

November 2024

## 1 Reviewer 5kwJ

The authors thank Reviewer 5kwJ for all the interesting questions.

### Weakness #1

As shown in previous studies [1], we experiment with augmenting self-attention with additional bias terms with the idea to model such attention biases explicitly, introducing additional learnable parameters for each head:

$$b_e = Qke' \tag{1}$$

$$b_v = \text{softmax}(A_e)v', \tag{2}$$

where  $k, e, v \in \mathbb{R}^d$  are the key, edge, and node bias terms (one per each attention head),  $A_e$  is the edge attention output, and  $d$  the corresponding hidden dimension. Equations 1 and 2 were chosen based on the architectures of the models considered and following validation experiments.

### Weakness #2

Our motivation was to mathematically highlight MAs as an anomalous behavior that deviated from a standard one. Therefore, we defined a priori what the standard behavior was, i.e., the distribution of activations for an untrained base model with Xavier randomly initialized weights (Fig.3a on the main paper), on which the best approximation was precisely given by a gamma family distribution. Consequently, the same distribution had to be reused in all the other datasets/models/layers precisely to highlight anomalous behaviors such as MAs. We would like to specify that the method by which we chose the best base model distribution relies on an empirical approach, where we tried various families of distributions (e.g., normal, lognormal, exponential, Weibull) whose experiments showed that the gamma family was the best approximation. The fit is always done independently for each layer, model, dataset.

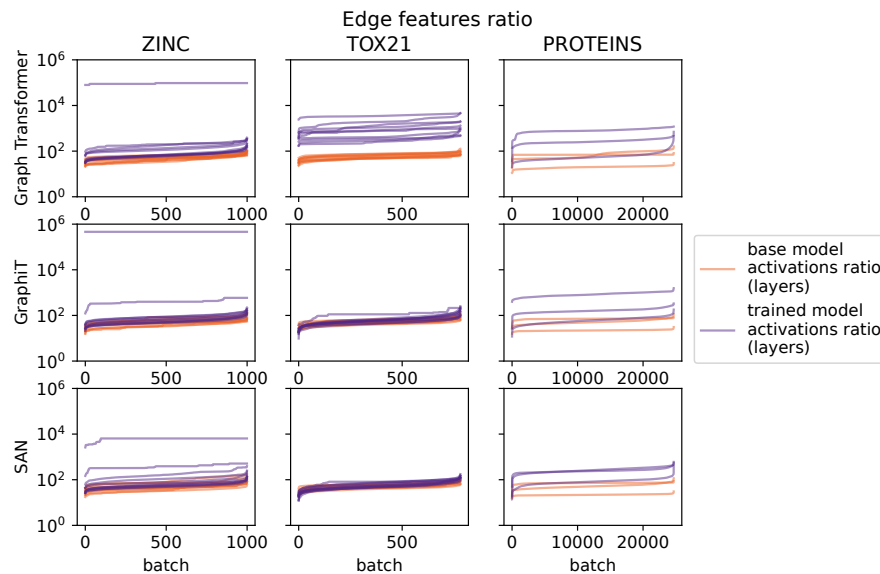


Figure 2: (corresponding to Figure 2 in paper) - Comparison of MAs on trained against base models, without the use of Explicit Bias Term. Represented ratios have been sorted increasingly for each layer independently.

### Weakness #3

Initially, we plotted figures displaying lines for each layer of the base and trained models. However, the authors found that displaying 20 lines in each of the nine figures was visually confusing (mostly with ZINC). To improve clarity, the layer values for the base model were transformed into a min-max range, which helped highlight the MAs in the layers of the trained model. Nevertheless, one of the original figures is included in the ‘supplementary\_material/ICLR25\_Rebuttal.pdf’ as Figure 2.

### Weakness #4

We appreciated the suggestion and we will modify tables and illustrations to meet the standards.

### Question #1

Our motivation was to mathematically highlight MAs as an anomalous behavior that deviated from a standard one. Therefore, we defined a priori what the standard behavior was, i.e., the distribution of activations for an untrained base model with Xavier randomly initialized weights (Fig.3a on the main paper), on which the best approximation was precisely given by a gamma family

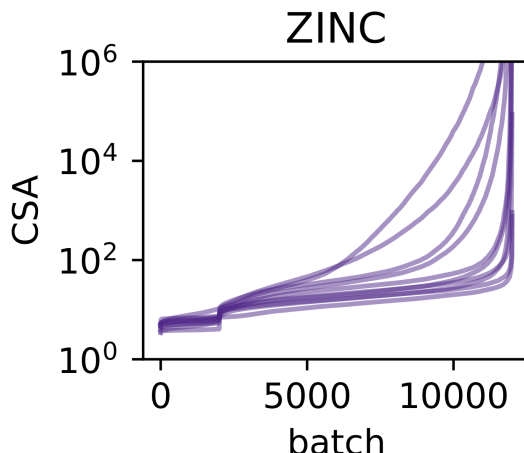


Figure 3: Activation ratios for CSA on ZINC dataset. y-axis scale is the same as all main paper’s figures. Incredibly high values show the presence of MAs.

distribution. Consequently, the same distribution had to be reused in all the other datasets/models/layers precisely to highlight anomalous behaviors such as MAs. We would like to specify that the method by which we chose the best base model distribution relies on an empirical approach, where we tried various families of distributions (e.g., normal, lognormal, exponential, Weibull) whose experiments showed that the gamma family was the best approximation. The fit is always done independently for each layer, model, dataset.

## Question #2

We experimented with newer models such as GraphGPS [2] and CSA [3], which showed us the presence of MAs on both ZINC (as depicted in Figure 3) and PCQM4Q-full v2 datasets. In addition, we will enrich the benchmark by introducing new models and datasets to further show the presence of MAs, e.g we will investigate [2] and [3] over:

- ogbg-molhiv;
- ogbg-molpcba;
- ogbg-ppa;
- ogbg-code2.

The choice of models and datasets, in any way, always reflects our original idea: that is, models that incorporate extra information about edges from their datasets, via edge features, and thus focus attention on both nodes and edges during their message passing.

## 2 Reviewer Hp9k

The authors thank Reviewer Hp9k for all the interesting questions.

### Weakness/Question #1

The authors are committed to further specify in the Introduction Section that the main motivation of the article lies in the identification of MAs in graph transformers models capable of using dataset’s edge features in their attention mechanisms, so that additional information is considered in the message-passing. Therefore, conventional attention-based GNNs (e.g., GATs and corresponding variants) lacking edge features attention are not analyzed.

### Weakness/Question #2

Edge features in these contexts refer to additional attributes or properties associated with the edges of the graphs. They represent information about the relationship or interaction between the connected nodes. Typically, edge features are specific to a domain, e.g., bond types in molecular graphs like ZINC and TOX21, or spatial/interaction properties between amino acids or secondary structure elements like PROTEINS.

### Weakness/Question #3

Further experiments will be conducted extensively on other models and datasets, for example, we attach the results obtained for CSA-GraphGPS on PCQM4Mv2-full (from OGB) and ZINC (see Figure 3). In addition, we will also consider other OGBN datasets, as already done in the paper with PROTEINS (which belongs precisely to OGBN for node prediction properties <https://ogb.stanford.edu/docs/nodeprop/#ogbn-proteins>).

### Weakness/Question #4

Accordingly to the answer #1 and #3 we attach results of more recent edge-features attention-based graph transformers on PCQM4Mv2-full and ZINC datasets.

### Weakness/Question #5

To have insights on the correlation between MAs and models’ robustness, adversarial attack was carried out only on the most susceptible model-dataset configuration which is the one that exhibits MAs within each layer. Furthermore, authors want to specify that adversarial attack is proposed as possible future directions to be analyzed, and not meant to be the main focus at this stage, but rather to be deeper studied in follow-up works.

### 3 Reviewer u2zz

The authors thank Reviewer u2zz for all the interesting questions.

#### Weakness #1

Due to similar questions, we report hereafter the same answer as Reviewer 5kwJ.

We experimented with newer models such as GraphGPS [2] and CSA [3], which showed us the presence of MAs on both ZINC (as depicted in Figure 3) and PCQM4Q-full v2 datasets. In addition, we will enrich the benchmark by introducing new models and datasets to further show the presence of MAs, e.g. we will investigate [2] and [3] over:

- ogbg-molhiv;
- ogbg-molpcba;
- ogbg-ppa;
- ogbg-code2.

The choice of models and datasets, in any way, always reflects our original idea: that is, models that incorporate extra information about edges from their datasets, via edge features, and thus focus attention on both nodes and edges during their message passing.

#### Weaknesses #2 - #3

As suggested in “Conclusion and Future Work”, we acknowledge the importance of exploring the impact of MAs on various downstream tasks (e.g., customized adversarial MAs, downstream-driven MAs). However, being the first work to study them on graphs we have chosen to explore deeper the characterization of MAs and whether their presence can be spotted using statistical tests (KS-test have been used indeed) and rigorous empirical analysis. In relation to the latter, we are committed to investigate more datasets and configurations as stated in point #1.

#### Question #1

Indeed, the proposed models compute attention between pairs of adjacent nodes only. However, recent experiments with newer models such as GraphGPS and CSA (see Figure 3), which show the presence of MAs on ZINC and PCQM4Q-full-v2, compute attention also for non-adjacent node pairs.

### Reviewer o79Q

The authors thank Reviewer o79Q for all the interesting questions.

### **Weakness/Question #1**

Although [1] explored MAs for LLMs, our work differs in:

- being the first work to study them on attention-based GNN;
- employing statistical tool, such as KS test and statistics, and gamma distribution, to characterize them;
- treating them as anomalies with respect the base distribution;
- defining the base model as the untrained model with Xavier randomly initialized weights;
- overviewing MAs implication on models’ robustness (Eq.4 - Table 2).

### **Weakness #2**

We thank the reviewer for raising this points, however could you please point out what “unimportant features” are?

### **Weakness #3**

We are committed to integrate into Introduction section a deeper explanation of the methodology of spotting MAs. We precise that we used the base (untrained) model to define what a standard behavior is for activation values, and that we used the median activation value to normalize the activation values in a given layer. Therefore, the median normalization is done for layers of both the trained and base models, whose distributions are then compared between each other.

### **Weakness #4 -#5**

We appreciated the suggestion and we will modify tables and illustrations to meet the standards.

### **Question #2**

Thank you again for raising this point. To provide an adequate response, we would like to know, according to the reviewer’s opinion, what confusion the statement could generate.

### **Question #3**

Indeed, the behavior of MAs changes depending on the choice of dataset, however the values of activations in the trained model generally exceed the range of the base model (without MAs).

## References

- [1] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.
- [2] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.
- [3] Romain Menegaux, Emmanuel Jehanno, Margot Selosse, and Julien Mairal. Self-attention in colors: Another take on encoding graph structure in transformers. *arXiv preprint arXiv:2304.10933*, 2023.