We thank all reviewers (denoted as R1, R2 and R3) for constructive feedback and questions. We provide answers below.

(R1, R2, R3) Training weakly-supervised models. Assume we want to train model **A** with "weak-sup" attention on a dataset w/o ground truth attention. We first need to train model **B** that has the same architecture as **A**, but does not have any attention/pooling between graph conv. layers. So, model **B** has only global pooling. After training **B** with the $\mathcal{L}_{MSE/CE}$ loss, we need to evaluate training graphs on **B** as follows. For each training graph $\mathcal{G}$ with N nodes, we first make a prediction y for the entire $\mathcal{G}$. Then, for each $i \in [1, N]$, we remove node i from $\mathcal{G}$, and feed this reduced graph with $N-1$ nodes to model **B** and record the model's prediction y_i. We then use Eq. 6 to compute α^{WS} based on y and y_i. Now, we can train **A** and use α^{WS} instead of ground truth α^{GT} in Eq. 5 to optimize both *MSE/CE* and *KL* losses.

(R1) Focus and a title of the paper. Instead of "*Understanding Attention...*" we propose the new title "*On Initialization and Supervision of Attention...*". The main purpose of incorporating other topics was to support our conclusions about *attention*, which is our central focus. It can often be relatively easy to identify a phenomenon in (graph) neural networks, but it is hard to resolve it. Thus, our weak-sup method complements our more analysis rather than methods-driven focus.

(R1) Limited number of models. We evaluate and draw conclusions based on three models of different strength: GIN, ChebyNet, ChebyGIN. We will also add results of GCN supporting our conclusions (Table 1 and Figure 1). Many other GNNs proposed in the literature are slight modifications of these. Generally, stronger models are required to achieve higher attention accuracy, which would lead to the exponential gains in classification accuracy that we observe. With weaker models, this phenomenon can be observed on rather simple datasets (COLORS). Note that in Table 1 of the submitted paper, for COLORS and MNIST-75sp, ChebyGINs are equivalent to ChebyNets as described in Table 1 of the Supplementary material and elaborated on following that table (see footnote 3). We will update Table 1 in the paper to make it clear.

Table 1: **(R1, R3) About Table 2.** In addition to the results of ChebyNet in the submitted paper, we report results of GCN. For more reliable comparison, we repeat experiments for 100 random seeds instead of 10. "init tune" denotes tuning σ and choosing between $\mathcal{N}$ or U (see Figure 1 at the bottom); tuning is done in the same way as for other hyperparameters.

Model	**Proteins**$_{25}$
GCN + Global max	$74.4_{\pm 1.0}$
GCN + Unsup	$75.6_{\pm 1.4}$
GCN + Weaksup	$76.2_{\pm 0.7}$
GCN + Unsup+init tune	$75.5_{\pm 1.7}$
GCN + Weaksup+init tune	$\mathbf{76.4}_{\pm \mathbf{0.7}}$

(R2) Hard or soft attention scores? In our model, the features are *weighted* by attention scores according to Eq. 3, so it is soft. In this case, the features indeed reduce their scale. But we haven't found this problematic in our tasks. For really large graphs, we found that using a constant multiplier $c > 1$ can help considerably to keep the scale of features in a reasonable range, while still permitting weighting of node features: $c(\alpha_i X_i)$. This is more an implementation trick and we are releasing code to support such cases. **(R2) Why GIN is better than ChebyGIN in some cases?** ChebyGIN has larger capacity and, thus, can overfit easily to such a simple training distribution as in COLORS. Moreover, the difference between GIN and ChebyGIN in COLORS is not significant (see the standard deviations). **(R2) Computing AUC vs Rank correlation.** AUC is computed between binarized ground truth attention scores (i.e. $\alpha^{GT} > 0$) and predicted α. We indeed have considered other metrics. While they can be good for evaluation, the choice of AUC is more natural, since we can directly choose a pooling threshold $\tilde{\alpha}$ by looking at the ROC curve and finding a good balance between false-positives (pooling unimportant nodes) and false-negatives (dropping important nodes). So, AUC provides a comprehensive picture for different $\tilde{\alpha}$. Also, the problem with rank correlation metrics is that it is unclear whether to include α of dropped nodes during calculation of the metric, and that can lead to very different results, which complicates comparison.

(R3) Tuning threshold $\tilde{\alpha}$ and other hyperparams. The best $\tilde{\alpha}$ is typically around $1/4N$ to $2/N$, where N is the max number of nodes in the training graphs. We chose $\tilde{\alpha}$ from the range $[0.0001, 0.1]$. We tune $\tilde{\alpha}$ and a few other hyper-parameters (r in top-k and β in Eq. 5) on the val. sets generated in the same way as TEST-ORIG for synthetic datasets; part of the training set for MNIST-75sp; using 10-fold cross-val. on the training set for COLLAB, PROTEINS and D&D.

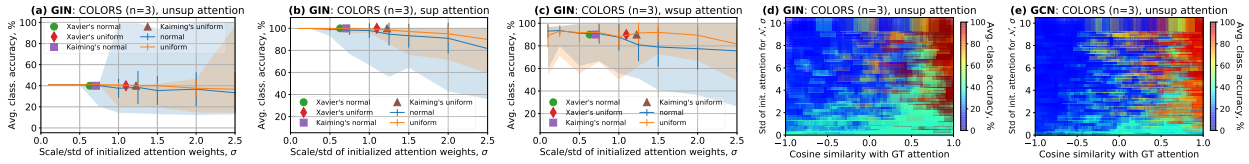

Figure 1: **(R1) Initialization methods.** In these experiments, we evaluate **GIN** *(a-d)* and **GCN** *(e)* on a wide range of random distributions $\mathcal{N}(0, \sigma)$ and $U(-\sigma, \sigma)$ by varying σ. In the submitted version of the paper, we used $\mathcal{N}(0, 1)$ by default. We show points corresponding to the commonly used initialization strategies of (Xavier Glorot & Bengio, 2010) and (Kaiming He et al., 2015). We see that for unsupervised training *(a)*, larger initial values and the Normal distribution should be used to make it possible to converge to an optimal solution, which is still unlikely and greatly depends on cosine similarity with GT attention *(d,e)*. For supervised and "weak-sup" attention, we should use smaller initial weights and either the Normal or Uniform distribution *(b,c)*. We have similar plots for COLORS with $n = 16$ dimensional features to be added to the camera-ready version. *(a-c)* Shaded areas show range, bars show ± 1 std.