# Towards Precise Prediction Uncertainty in GNNs: Refining GNNs with Topology-grouping Strategy

**Hyunjin Seo[1,3], Kyusung Seo[1*], Joonhyung Park[1*], Eunho Yang[1,2†]**
[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]AITRICS
[3]Polymerize
{bella72, seo3650, deepjoon, eunhoy}@kaist.ac.kr

## Abstract

Recent advancements in graph neural networks (GNNs) have highlighted the critical need of calibrating model predictions, with neighborhood prediction similarity recognized as a pivotal component. Existing studies suggest that nodes with analogous neighborhood prediction similarity often exhibit similar calibration characteristics. Building on this insight, recent approaches incorporate neighborhood similarity into node-wise temperature scaling techniques. However, our analysis reveals that this assumption does not hold universally. Calibration errors can differ significantly even among nodes with comparable neighborhood similarity, depending on their confidence levels. This necessitates a re-evaluation of existing GNN calibration methods, as a single, unified approach may lead to sub-optimal calibration. In response, we introduce a novel approach that categorizes nodes by both neighborhood similarity and their own confidence, irrespective of proximity or connectivity. Our method allows fine-grained calibration by employing group-specific temperature scaling, with each temperature tailored to address the specific miscalibration level of affiliated nodes, rather than adhering to a uniform trend based on neighborhood similarity. Extensive experiments demonstrate the effectiveness of our framework across diverse datasets on different GNN architectures, achieving up to 13.79% error reduction compared to uncalibrated GNN predictions.

## Introduction

Graph neural networks (GNNs) have demonstrated remarkable performance in modeling graph data and addressing diverse graph-based tasks, such as node classification (Kipf and Welling 2016; Hamilton, Ying, and Leskovec 2017; Xu et al. 2018; Park, Song, and Yang 2021), link prediction (Zhang and Chen 2018; Yun et al. 2021; Ahn and Kim 2021; Zhu et al. 2021), and graph classification (Lee, Rossi, and Kong 2018; Sui et al. 2022; Hou et al. 2022). Beyond achieving correct prediction, precisely quantifying prediction uncertainty is nontrivial for the reliable utilization of neural networks in downstream decision-making process. Recognizing such need, numerous calibration studies have been actively proposed in vision and language domains (Guo et al. 2017; Mukhoti et al. 2020; Zhang, Kailkhura, and Han 2020; Xing et al. 2019; Jiang et al. 2021; Minderer et al. 2021).

Recently, network calibration has also drawn attention in the field of GNNs (Wang et al. 2021; Hsu et al. 2022; Hsu, Shen, and Cremers 2022; Shi et al. 2022; Wang, Yang, and Cheng 2022; Liu et al. 2022), highlighting neighborhood prediction similarity as a crucial factor for calibration. Contemporary studies in GNN calibration, CaGCN (Wang et al. 2021) and GATS (Hsu et al. 2022), suggest that nodes with similar neighborhood prediction similarity tend to exhibit analogous calibration characteristics. Specifically, CaGCN asserts that nodes with disparate neighbors should ideally have lower confidence levels, as the local message propagation in GNNs makes accurately classifying such instances more challenging. Conversely, GATS elucidates the correlation between neighborhood prediction similarity and calibration errors, indicating the highest errors for nodes with conflicting neighbors. To account for these trends, they incorporate neighborhood similarity into node-wise temperature scaling, facilitating confidence propagation between adjacent nodes.

However, our analysis reveals that calibration cannot be effectively addressed by applying a single, unified trend. Specifically, we observe that calibration errors vary significantly among nodes with comparable neighborhood similarity, depending on their individual confidence levels. More critically, both over-confidence and under-confidence can occur in nodes with similar neighborhood similarity but differing confidence levels. This phenomenon has not been effectively captured in previous studies, as they do not fully account for both factors. Consequently, their assumptions may lead to sub-optimal calibration, as they are not universally applicable.

To address this, we introduce SIMI-MAILBOX, a novel post-hoc calibration method designed to overcome these limitations. Our method categorizes nodes based on both neighborhood representational similarity and confidence, irrespective of proximity or connectivity. This grouping strategy is grounded on our observation that nodes with comparable levels of neighborhood similarity and confidence exhibit similar calibration errors. SIMI-MAILBOX then assigns *group-specific* temperatures to adjust the predictions of nodes within each group. This fine-grained approach en-


†Corresponding Author.
*Equal Contribution.

sures that each group-wise temperature is tailored to address the specific miscalibration of affiliated nodes, instead of relying on a uniform tendency.

In summary, our contributions are three-fold:

- We elucidate the limitations inherent in current calibration methods, particularly concerning neighborhood prediction similarity - a recognized key component for GNN calibration.
- Given these limitations, we propose SIMI-MAILBOX, a novel calibration method that rectifies miscalibration by introducing group-specific temperatures. Each group-wise temperature is focused on adjusting the predictions of affiliated nodes, rather than scaling all nodes according to a unified trend.
- We validate the efficacy of SIMI-MAILBOX through comprehensive experiments, incorporating both quantitative and qualitative evaluations.

## Related Works

**Uncertainty Quantification and Post-hoc Calibration.** Network calibration, while sharing the root in uncertainty quantification with conformal prediction and bayesian methods, concentrates on aligning model predictions with empirical event frequencies, differing from providing intervals around prediction or modeling uncertainty under probability distribution. Conformal prediction aims to generate tight prediction sets that encompass the true outcome with a pre-specified coverage. Its foundational concept was presented in (Vovk, Gammerman, and Shafer 2005), and the study for providing good coverage has been consistently explored, evolving through (Romano, Sesia, and Candes 2020; Cauchois, Gupta, and Duchi 2021; Angelopoulos et al. 2020). Bayesian approaches, on the other hand, use probabilistic modeling to interpret the uncertainty via posterior distribution. Their representative techniques include ensembles (Lakshminarayanan, Pritzel, and Blundell 2017; Wen, Tran, and Ba 2020), dropout (Gal and Ghahramani 2016) and Bayesian Neural Networks (BNNs) for applying Bayesian inference in neural networks (Depeweg et al. 2018; Maddox et al. 2019; Dusenberry et al. 2020).

Distinct from the aforementioned approaches, calibration is focused on refining the trustworthiness of the model prediction. Their goal is focused on adjusting the model's confidence to match the ground-truth probability. Among diverse calibration techniques, post-hoc calibration methods have found widespread adoption, owing to their computationally efficiency compared to traditional Bayesian approaches and model regularization-based methods (Ma and Blaschko 2021; Jung et al. 2023). Moreover, they imposes no constraints during the pretraining phase of main models, thereby enhancing its versatility across diverse architectures. Techniques such as Platt scaling (Platt et al. 1999), Temperature scaling (TS) (Guo et al. 2017), and Ensemble temperature scaling (ETS) (Zhang, Kailkhura, and Han 2020) have been developed for this purpose, with TS being notably effective for its simplicity and effectiveness in multi-class calibration.

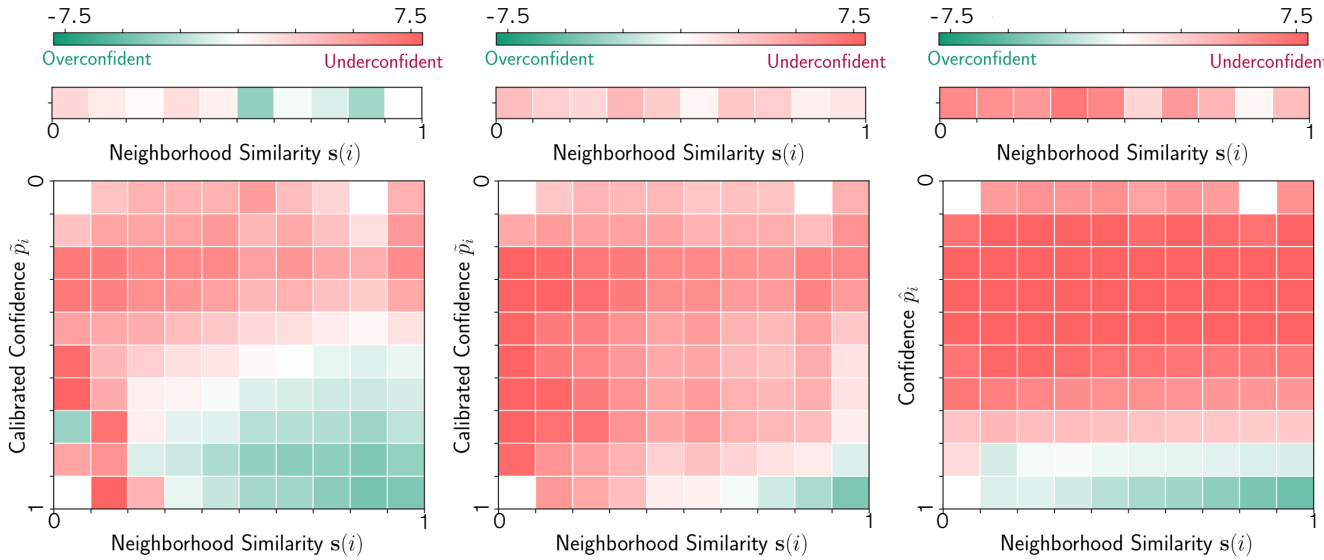
**Grouping-based Calibration.** Addressing miscalibrations in a group-wise manner has been studied in (Hébert-Johnson et al. 2018; Perez-Lebel, Morvan, and Varoquaux 2022; Yang, Zhan, and Gan 2023). (Hébert-Johnson et al. 2018) introduced multicalibration strategy, aiming to achieve calibration within diverse, overlapping subgroups to enhance both fairness and accuracy in machine learning models. Meanwhile, (Perez-Lebel, Morvan, and Varoquaux 2022) presented the concept of grouping loss as a novel metric to assess the variance in true probabilities sharing the same confidence score, challenging existing calibration approaches. (Yang, Zhan, and Gan 2023) proposed a new semantic partitioning approach for neural network calibration and utilized learnable grouping function to refine calibration beyond traditional methods. Nevertheless, these studies do not provide the **specific principles** for effective categorization, which highlights the distinction of our work from preceding ones. More unique aspects of our approach in comparison to prior works are discussed in the Appendix.

**Uncertainty Quantification for GNNs.** Recent literature has increasingly focused on quantifying uncertainty in GNNs, with methods ranging from conformal prediction using local topologies (Huang et al. 2024; Zargarbashi, Antonelli, and Bojchevski 2023) to Bayesian approaches (Stadler et al. 2021; Rong et al. 2019; Hasanzadeh et al. 2020; Elinas, Bonilla, and Tiao 2020; Pal, Regol, and Coates 2019; Zhao et al. 2020) that concentrate on the interdependent graph data and GNNs. The literature also highlights post-processing calibration strategies (Wang et al. 2021; Hsu et al. 2022; Hsu, Shen, and Cremers 2022; Wang, Yang, and Cheng 2022; Shi et al. 2022; Liu et al. 2022), with (Wang et al. 2021) pioneering in revealing unexpected underconfidence in GNN predictions. They introduced CaGCN, which employs GCN for node-specific calibration through adjacent predictions. Expanding this, GATS (Hsu et al. 2022) explored factors leading to GNN calibration errors and designed GAT-based node-wise calibration function considering these factors. They further introduced an edge-wise calibration error metric to capture the non-iid nature of graphs in (Hsu, Shen, and Cremers 2022). In a different approach, GCL (Wang, Yang, and Cheng 2022) addressed the underconfidence of GNNs by integrating a minimal-entropy regularization with the cross-entropy loss, up-weighting the loss on highly confident nodes.

## Preliminaries

**Problem Setup.** We focus on calibrating the prediction uncertainty of GNNs for semi-supervised node classification in a post-hoc setting. In this context, uncertainty denotes the model's confidence level in its predictions, while calibration aims to align this uncertainty with the true accuracy, enhancing the model's reliability. Thus, our objective is to minimize the gap between the predicted probability and the actual accuracy of given data. During the post-hoc calibration phase, the validation set is used for training to enhance generalization to unseen data, avoiding the overfitting risk associated with reusing the original training set.

Let an undirected graph be denoted as $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$




(a) CoraFull (CaGCN) (b) CoraFull (GATS) (c) CoraFull (UnCal.)

Figure 1: Analysis of uncalibrated and calibrated logits via prior works, CaGCN and GATS. The $x$-axis divides nodes into sub-intervals based on neighborhood similarity, while the $y$-axis represents corresponding confidence intervals. Each cell in the heatmap represents the subtraction of the average confidence from the accuracy, with color intensity indicating the magnitude of this discrepancy. Contrary to the uniform assumptions in prior works on neighborhood similarity, the results demonstrate that calibration errors can significantly differ among nodes with comparable neighborhood similarity but different confidence levels. Moreover, prior approaches exhibit sub-optimal calibration across varying neighborhood similarity levels when predictions are extended across confidence intervals.

and $\mathcal{E}$ indicate the sets of vertices and edges respectively. The vertex set $\mathcal{V}$ is represented by a feature matrix $\boldsymbol{X} = [\mathbf{x}_1^\mathsf{T}, ..., \mathbf{x}_{|\mathcal{V}|}^\mathsf{T}] \in \mathbb{R}^{|\mathcal{V}| \times D}$ and the edge set $\mathcal{E}$ is denoted by an adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$. Given the node-wise predictions $\hat{y} = [\hat{y}_1, ..., \hat{y}_{|\mathcal{V}|}]^\mathsf{T}$ and output confidence $\hat{p} = [\hat{p}_1, ..., \hat{p}_{|\mathcal{V}|}]^\mathsf{T} \in \mathbb{R}^{|\mathcal{V}|}$ from a trained GNN, the GNN $f_\theta$ is *well-calibrated* if $\hat{p}_i$ for each node $i$ accurately serves the ground-truth probability $p_{\text{true}}$, formulated as below:

$$\mathbb{P}(\hat{y}_i = y_i | \hat{p}_i = p_{\text{true}}) = p_{\text{true}}, \quad \forall p_{\text{true}} \in [0, 1]. \tag{1}$$

The expected calibration error (ECE) (Naeini, Cooper, and Hauskrecht 2015) has been recognized as the de facto metric to evaluate the calibration quality of network predictions. ECE groups nodes according to their confidences into $M$ equally partitioned confidence intervals $\{B_1, ..., B_M\}$ and assesses the expected discrepancy between accuracy and average confidence within individual bins:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{|\mathcal{V}|} \Big| \text{acc}(B_m) - \text{conf}(B_m) \Big|, \tag{2}$$

where $|B_m|$ refers to the number of nodes within the $m$-th interval. Here, the accuracy and average confidence for the $m$-th bin are defined as $\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}[y_i = \hat{y}_i]$ and $\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$, respectively.

**Neighborhood Similarity in Prior Studies.** The concept of neighborhood similarity has been recognized as a primary element in the field of GNN calibration (Wang et al. 2021; Hsu et al. 2022; Hsu, Shen, and Cremers 2022; Liu et al. 2022). Among them, CaGCN (Wang et al. 2021) advocates that given the challenges GNNs encounter in accurately classifying nodes with conflicting neighbors, the confidence levels in such cases should ideally remain still or decrease. Conversely, confidence for nodes linked to agreeing nodes should elevate, addressing the prevalent underconfidence in GNNs. Stemmed from this insight, they employ GCN (Kipf and Welling 2016) as a node-wise calibration function to propagate the confidence to neighboring counterparts. In parallel, GATS (Hsu et al. 2022) underscores the correlation between neighborhood prediction similarity and calibration error, demonstrating an increment in error with a decrement in similarity. This relationship is incorporated into the normalized attention coefficients within their GAT (Veličković et al. 2017)-founded node-level temperature function.

## In-depth Analysis on Neighborhood Similarity

In this section, we provide a comprehensive analysis of both uncalibrated and calibrated predictions from existing studies, CaGCN and GATS, using the CoraFull dataset (Bojchevski and Günnemann 2017). Leveraging GCN as the backbone architecture, we first partition nodes into 10 equal intervals $\{B^{(1)}, ..., B^{(10)}\}$ based on the proportion of neighbors sharing the same predicted labels, denoted as neighborhood prediction similarity $\boldsymbol{s}(i)$:

$$\boldsymbol{s}(i) = \frac{\sum_{j \in \mathcal{N}_i} \mathbf{1}[\hat{y}_i = \hat{y}_j]}{|\mathcal{N}_i|}, \tag{3}$$

where $\mathcal{N}_i$ represents the set of neighbors associated with node $i$. For each subgroup $B^{(l)}$, we calculate the calibration error as the discrepancy between their average confidence and the accuracy, *i.e.*, $\text{acc}(B^{(l)}) - \text{conf}(B^{(l)})$. These discrepancies are depicted as heatmap bars in the second row of Figure 1.

Furthermore, we also analyze predictions by considering both neighborhood similarity and confidence. We begin by grouping confidence into 10 equal intervals $\{B_1, ..., B_{10}\}$, and then further categorize nodes within each confidence interval into 10 equal-width intervals based on $\boldsymbol{s}(i)$. The subgroup within the $l$-th similarity interval and $m$-th confidence interval is denoted as $B_m^{(l)}$. For each subgroup $B_m^{(l)}$, the calibration error is computed as $\text{acc}(B_m) - \text{conf}(B_m^{(l)})$. These discrepancies are illustrated as heatmap matrices in the last row of Figure 1, with $\hat{p}$ representing uncalibrated confidence and $\widetilde{p}$ representing calibrated confidence.

In Figure 1, the heatmap elements represent the differences between accuracy and average confidence. Deeper shades of **red** indicate that the calibrated confidence is lower than the accuracy (**under-confident**), while deeper shades of **green** indicate that confidence exceeds the accuracy (**over-confident**). Our findings show that calibration errors can vary significantly among nodes with the same level of neighborhood similarity but different confidence. Notably, both under- and over-confidence are observed in uncalibrated predictions within $\boldsymbol{s}(i) \in (0.1, 1.0]$ similarity intervals of the heatmap matrices.

Moreover, our analysis reveals that existing methods, which apply a unified policy to nodes with similar levels of neighborhood similarity, fail to achieve consistent calibration across diverse neighborhood similarity levels. While these methods may appear well-calibrated according to the heatmap bars in the second row, they demonstrate suboptimal results when their predictions are extended across confidence intervals. Specifically, CaGCN exhibits severe under-confidence in $\widetilde{p}_i \in (0.9, 1.0]$ confidence interval, with a maximum discrepancy of approximately 16.34% within the $\boldsymbol{s}(i) \in (0.1, 0.2]$ similarity range. GATS, on the other hand, demonstrates suboptimal calibration in regions of low prediction similarity, particularly in the $\widetilde{p}_i \in (0.2, 0.4]$ and $\widetilde{p}_i \in (0.6, 0.8]$ ranges, where the average discrepancies are 7.45% and 7.17% in the $\boldsymbol{s}(i) \in (0, 0.4]$ intervals, respectively. Hence, our observations suggest that a unified assumption to calibrating predictions based on neighborhood similarity cannot effectively achieve fine-grained calibration. We also provide an algorithmic perspective on the limitations of previous work, along with additional investigation results on more benchmark datasets, in the Appendix.

## Proposed Method

Given the limitation of earlier studies, we introduce SIMI-MAILBOX, a post-hoc calibration method designed to rectify miscalibration in GNNs across varying levels of neighborhood similarity. Building on our novel observation, SIMI-MAILBOX categorizes nodes based on both neighborhood similarity and confidence levels, ensuring that nodes within the same cluster exhibit similar calibration errors. Subsequently, our method employs group-specific temperature scaling to adjust the predictions of nodes in the designated cluster. These group-wise temperatures are tailored to correct the specific miscalibration associated with each group, instead of relying on a uniform tendency. The temperatures are optimized by directly minimizing the discrepancy between average confidence and accuracy within each cluster.

### Intuition: Topology Grouping Matters

| GNNs | | Cora | Citeseer | Pubmed | Computers | Photo |
|---|---|---|---|---|---|---|
| GCN | Node-wise | 6.139 | 1.957 | 1.370 | 40.370 | 7.200 |
| | Conf. | 0.065 | 0.060 | 0.068 | **0.060** | 0.052 |
| | Neig. Sim. | **0.057** | **0.046** | **0.061** | 0.062 | **0.047** |
| GAT | Node-wise | 8.656 | 2.570 | 1.614 | 44.980 | 14.550 |
| | Conf. | 0.068 | 0.062 | 0.068 | 0.048 | 0.053 |
| | Neig. Sim. | **0.056** | **0.044** | **0.055** | **0.045** | **0.047** |

Table 1: Variance of calibration errors ($\times\mathbf{100}$) involving neighborhood similarity sub-intervals (Neig. Sim.), confidence intervals (Conf), and total nodes (Node-wise).

For effective group-wise calibration, it is essential to categorize nodes in a manner that ensures they share a similar degree of miscalibration. This allows each group's temperature to be precisely tailored to address specific miscalibration levels rather than applying a broad, generalized adjustment. To this end, we present a novel observation suggesting that nodes with similar neighborhood prediction similarity $s(i)$ and confidence $\hat{p}_i$ share similar magnitudes of calibration errors. To substantiate this, we evaluate the variance of calibration errors under three different scenarios: (1) node-wise variance involving all nodes (specified as **Node-wise**), (2) variance within each confidence interval (specified as **Conf.**), and (3) variance within each neighborhood similarity sub-interval within each confidence interval (specified as **Neig. Sim.**).

To explore the third scenario, we assess the variability in calibration errors across neighborhood similarity intervals within each confidence interval. Let $B_m^{(l)}$ represent the set of nodes in $l$-th neighborhood similarity interval and $m$-th confidence interval. The calibration error for each node $i$, defined as the absolute difference between its confidence and the accuracy associated with its confidence interval, is denoted as $D(i)$. We first calculate the variance of calibration error within each $B_m^{(l)}$, denoted as $V(B_m^{(l)})$:

$$\begin{aligned} V(B_m^{(l)}) &= \frac{1}{|B_m^{(l)}| - 1} \sum_{i \in B_m^{(l)}} (D(i) - \bar{D}_m^{(l)})^2, \\ D(i) &= |\text{Acc}(B_m) - \hat{p}_i|. \end{aligned} \tag{4}$$

where $\bar{D}_m^{(l)}$ represents the mean calibration error for nodes in $B_m^{(l)}$. We then average these variances over the collection $B^{\text{sim}} = \{B_1^{(1)}, B_1^{(2)}, ..., B_2^{(1)}, B_2^{(2)}, ...\}$, which incorporates all $B_m^{(l)}$ spanning the entire confidence intervals:

$$V^{\text{sim}} = \frac{1}{|B^{\text{sim}}|} \sum_{B_m^{(l)} \in B^{\text{sim}}} V(B_m^{(l)}). \tag{5}$$

Similarly, to assess the second scenario, we calculate the variability in calibration errors across confidence intervals $B^{\text{conf}} = \{B_1, B_2, ...\}$ by computing the variance within each $B_m$:

$$V(B_m) = \frac{1}{|B_m| - 1} \sum_{i \in B_m} (D(i) - \bar{D}_m)^2, \tag{6}$$

where $\bar{D}_m$ refers to the average calibration error for nodes within $B_m$. Following the approach used in GATS, we conceptualize node-wise calibration error as the calibration error of the confidence interval to which each node belongs. Consequently, the variance in calibration error related to individual nodes (the first scenario) is defined as the variance of all node-wise calibration errors.

As outlined in Table 1, the variance within **Neig. Sim.** shows the lowest, particularly when compared to the variance across all nodes (**Node-wise**). This demonstrates that nodes with comparable neighborhood predictions and confidence levels exhibit similar calibration error.

## SIMI-MAILBOX: A Topology-Grouping Strategy for Refining GNNs

Building on the observation discussed in previous section, SIMI-MAILBOX categorizes nodes by considering both neighborhood similarity and confidence levels. We estimate the neighborhood similarity for each node $i$ by computing the average representational similarity with its neighbors, denoted as MAILBOX $\mathcal{M}^{simi}(i)$:

$$\mathcal{M}^{simi}(i) = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \sigma(z_i^\mathsf{T} z_j), \tag{7}$$

where $z_i$ represents the output logits for node $i$ from trained GNN, and $\sigma$ is a sigmoid function. Nodes with similar MAILBOX values and confidence levels are then grouped into $N$ distinct clusters. More precisely, SIMI-MAILBOX constructs a feature vector $F_i^{simi} = [\bar{p}_i,\ \bar{\mathcal{M}}^{simi}(i)]^\mathsf{T}$ for each node $i$, with the first dimension representing normalized confidence $\bar{p}_i$ and the second dimension representing a normalized MAILBOX value via min-max scaling. Subsequently, KMeans clustering is applied to $F^{simi}$ to construct $N$ similarity-based clusters $C = \{C_1, ..., C_N\}$, ensuring the categorization adheres to both neighborhood similarity and confidence.

Once the categorization is completed, the original predictions for nodes within each cluster $C_n$ are scaled by a *group-specific* temperature $\boldsymbol{T}_n$, a learnable parameter designed to rectify the miscalibration within the $n$-th cluster:

$$\widetilde{p}_i = \max_k \sigma_{\text{sm}} \left( \frac{z_i}{\boldsymbol{T}_n} \right)_k \in \mathbb{R}, \quad i \in C_n. \tag{8}$$

The group-wise temperature $\boldsymbol{T} \in \mathbb{R}^N$ is then optimized with a new loss $\mathcal{L}_{simi}$ with standard cross-entropy loss $\mathcal{L}_{\text{CE}}$:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{simi}, \\ \mathcal{L}_{simi} &= \sum_{n=1}^{N} ||a_{val}^{(n)} - \frac{1}{|C_n|} \sum_{i \in C_n} \widetilde{p}_i||^2, \end{aligned} \tag{9}$$

where $\lambda$ is a scaling factor for $\mathcal{L}_{simi}$. During calibration, $\mathcal{L}_{\text{CE}}$ encourages the reduction of entropy for correctly predicted classes while increasing it for the incorrectly predicted ones. In parallel, $\mathcal{L}_{simi}$ minimizes the discrepancy between the average scaled confidence of all nodes and the accuracy of validation nodes $aval^{(n)}$ within each cluster. This approach directly adjusts the group-specific temperatures, with each $\boldsymbol{T}_n$ focused on minimizing the corresponding level of miscalibration.

**On Accuracy Preservation.** The post-hoc group-wise temperatures in SIMI-MAILBOX ensures that the relative ordering of predictions remains unchanged. Let $f : \mathbb{R}^K \to \mathbb{R}^K$ as a calibration function and $z_i = [z_{i1}, z_{i2}, ..., z_{iK}]^\mathsf{T}$ represents the logit vector for node $i$. We denote the group-specific temperature for the group to which node $i$ belongs as $\boldsymbol{T}_{g_i}$. Since the group-wise temperature $\boldsymbol{T}_{g_i}$ is uniformly applied to all elements of $z_i$, the order between elements in the calibrated logit $f_g(z_i)$ remains unchanged when subjected to the softmax operation $\sigma_{\text{sm}}$.

$$\begin{aligned} &f_g(z_i) = [z_{i1}/\boldsymbol{T}_{g_i}, z_{i2}/\boldsymbol{T}_{g_i}, ..., z_{iK}/\boldsymbol{T}_{g_i}], \\ &\breve{p}_i = \sigma_{\text{sm}}\left(f_g(z_i)\right). \end{aligned} \tag{10}$$

Thus, our method preserves the original classification accuracy, as the softmax function is order-preserving and scaling by $\boldsymbol{T}_{g_i}$ does not alter the relative ranking of logits.

**Comparison with Prior Studies.** While our work shares the post-hoc temperature scaling framework with previous GNN calibration methods, SIMI-MAILBOX introduces group-specific temperatures *independent of* node proximity or connectivity, thereby capturing high-level miscalibration patterns. Our method enables a more efficient optimization via $\mathcal{L}_{simi}$ due to its simplified number of parameters, compared to CaGCN and GATS requiring distinct temperatures for individual nodes. Moreover, the key distinction of our method lies in our discovery that nodes with similar neighborhood prediction similarity and confidence exhibit comparable calibration errors. This insight has not been explored in prior studies, as they do not fully consider the interplay between neighborhood similarity and confidence.

# Experiments

We validate the effectiveness of the proposed method under extensive experiments, leveraging two representative GNN architectures: GCN (Kipf and Welling 2016) and GAT (Veličković et al. 2017). The performance of our SIMI-MAILBOX is evaluated across eight small- and medium-scale benchmark graphs adopted in (Hsu et al. 2022): Cora, Citeseer, Pubmed (Sen et al. 2008), CoraFull (Bojchevski and Günnemann 2017), Coauthor CS, Computers, and Photo (Shchur et al. 2018). To further demonstrate the versatility, we extended our experiments to large-scale graphs, Arxiv (Hu et al. 2020) and Reddit (Zeng et al. 2019). More experiments including comparison with recent baselines, evaluations on heterophilous graphs and other GNN backbones, and hyperparameter robustness are provided in the Appendix.

| Methods | | UnCal. | TS | VS | ETS | CaGCN | GATS | Ours |
|---|---|---|---|---|---|---|---|---|
| Cora | GCN | 12.43 ± 4.24 | 3.87 ± 1.22 | 4.30 ± 1.28 | 3.78 ± 1.25 | 5.22 ± 1.45 | 3.55 ± 1.28 | **1.97 ± 0.44** |
| | GAT | 14.88 ± 4.30 | 3.42 ± 1.00 | 3.45 ± 1.13 | 3.32 ± 0.92 | 3.81 ± 1.00 | 3.05 ± 0.78 | **2.08 ± 0.45** |
| Citeseer | GCN | 12.54 ± 8.58 | 5.27 ± 1.70 | 5.15 ± 1.46 | 5.10 ± 1.76 | 6.60 ± 1.76 | 4.49 ± 1.53 | **2.66 ± 0.53** |
| | GAT | 16.65 ± 7.98 | 5.08 ± 1.48 | 4.62 ± 1.58 | 5.01 ± 1.46 | 4.86 ± 1.68 | 4.01 ± 1.42 | **2.86 ± 0.56** |
| Pubmed | GCN | 7.30 ± 1.56 | 1.27 ± 0.30 | 1.46 ± 0.29 | 1.26 ± 0.31 | 1.05 ± 0.33 | 0.95 ± 0.32 | **0.75 ± 0.15** |
| | GAT | 10.38 ± 1.89 | 1.15 ± 0.46 | 1.05 ± 0.36 | 1.13 ± 0.47 | 0.99 ± 0.34 | 0.98 ± 0.36 | **0.69 ± 0.16** |
| Computers | GCN | 2.96 ± 0.76 | 2.62 ± 0.55 | 2.70 ± 0.61 | 2.59 ± 0.72 | 1.70 ± 0.53 | 2.15 ± 0.52 | **1.02 ± 0.26** |
| | GAT | 1.58 ± 0.56 | 1.44 ± 0.35 | 1.44 ± 0.40 | 1.42 ± 0.43 | 1.82 ± 0.63 | 1.36 ± 0.34 | **0.95 ± 0.37** |
| Photo | GCN | 2.11 ± 0.97 | 1.68 ± 0.68 | 1.75 ± 0.67 | 1.63 ± 0.84 | 1.98 ± 0.53 | 1.46 ± 0.51 | **1.01 ± 0.36** |
| | GAT | 2.18 ± 1.54 | 1.56 ± 0.63 | 1.65 ± 0.70 | 1.57 ± 0.78 | 2.04 ± 0.74 | 1.49 ± 0.65 | **0.97 ± 0.53** |
| CS | GCN | 1.72 ± 1.28 | 1.01 ± 0.24 | 0.94 ± 0.28 | 0.97 ± 0.22 | 2.32 ± 1.12 | 0.90 ± 0.29 | **0.58 ± 0.19** |
| | GAT | 1.48 ± 0.79 | 1.07 ± 0.34 | 1.01 ± 0.40 | 1.03 ± 0.31 | 2.27 ± 1.13 | 0.85 ± 0.23 | **0.72 ± 0.43** |
| Physics | GCN | 0.56 ± 0.33 | 0.51 ± 0.19 | 0.46 ± 0.15 | 0.51 ± 0.19 | 0.88 ± 0.47 | 0.45 ± 0.15 | **0.28 ± 0.11** |
| | GAT | 0.55 ± 0.24 | 0.56 ± 0.20 | 0.56 ± 0.21 | 0.55 ± 0.20 | 1.06 ± 0.40 | **0.43 ± 0.16** | 0.48 ± 0.22 |
| CoraFull | GCN | 6.49 ± 1.28 | 5.55 ± 0.45 | 5.79 ± 0.43 | 5.49 ± 0.46 | 5.92 ± 2.84 | 3.74 ± 0.63 | **3.46 ± 1.31** |
| | GAT | 5.25 ± 1.32 | 4.41 ± 0.50 | 4.42 ± 0.49 | 4.36 ± 0.50 | 6.80 ± 3.81 | 3.46 ± 0.46 | **2.64 ± 1.02** |

Table 2: ECE results (reported in percentage) for our proposed calibration method and baselines. A lower ECE indicates better calibration performance. The best and second best performances are represented by bold and underline texts.

| Methods | | UnCal. | CaGCN | GATS | Ours |
|---|---|---|---|---|---|
| Arxiv | GCN | 4.92 ± 0.36 | 1.97 ± 0.16 | 0.75 ± 0.06 | **0.71 ± 0.13** |
| | SAGE | 3.00 ± 0.89 | 1.84 ± 0.19 | 2.05 ± 0.28 | **0.98 ± 0.23** |
| Reddit | GCN | 8.55 ± 1.28 | 1.86 ± 0.19 | 2.56 ± 0.59 | **0.35 ± 0.05** |
| | SAGE | 11.30 ± 1.99 | 2.14 ± 0.35 | 4.66 ± 0.57 | **0.73 ± 0.15** |

Table 3: ECE results (in percentage) for our method and baselines on large-scale datasets.

| Methods | | CaGCN | GATS | Ours |
|---|---|---|---|---|
| Arxiv | GCN | 20.84 ± 2.69 | 48.89 ± 11.39 | **7.10 ± 0.94 (-41.79 sec)** |
| | SAGE | 23.02 ± 4.44 | 61.67 ± 16.89 | **4.85 ± 0.65 (-56.82 sec)** |
| Reddit | GCN | 55.98 ± 13.76 | 72.90 ± 19.98 | **11.04 ± 0.30 (-61.86 sec)** |
| | SAGE | 78.13 ± 27.35 | 192.01 ± 177.57 | **9.91 ± 0.95 (-182.1 sec)** |

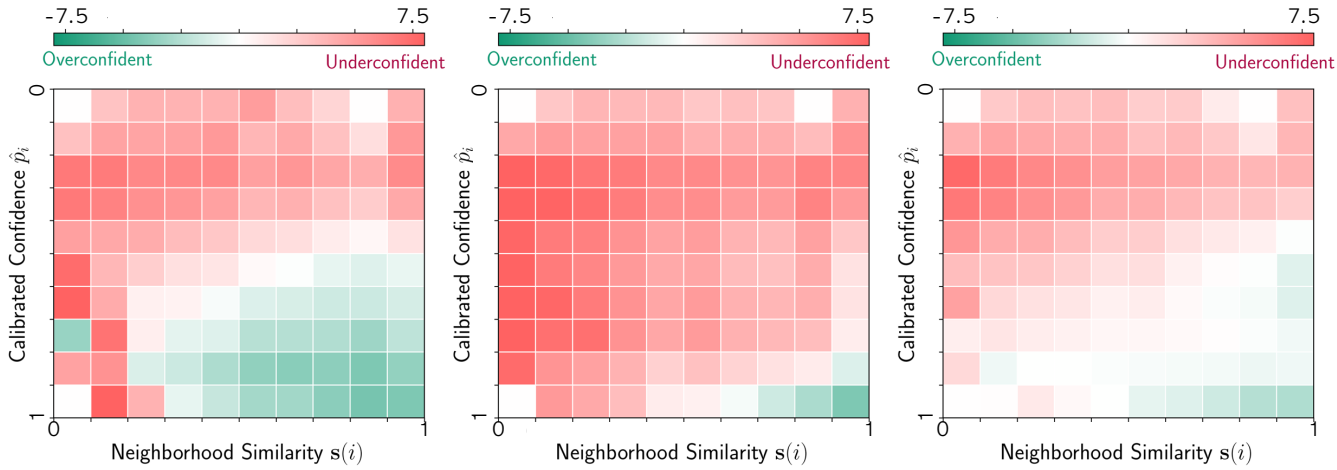
Table 4: Calibration duration (in seconds) for our method and baselines on large-scale datasets.

**Baselines.** In alignment with precedent studies, we compare our method against classical calibration methods: temperature scaling (TS), vector scaling (VS) (Guo et al. 2017), and ensemble temperature sclaing (ETS) (Zhang, Kailkhura, and Han 2020) and GNN-specialized calibration baselines: CaGCN (Wang et al. 2021) and GATS (Hsu et al. 2022). We provide an additional experiments to compare SIMI-MAILBOX and GPN (Stadler et al. 2021) and GNNSafe (Wu et al. 2023) for out-of-detection task in the Appendix.

**Experimental Setup.** We undertake our experiments following the experimental protocols of GATS (Hsu et al. 2022) in the scope of semi-supervised node classification. Details of the experiment configurations are provided in the Appendix. To assess the calibration performance, we use ECE as a principal metric (Naeini, Cooper, and Hauskrecht 2015), following the common practice (Hsu et al. 2022). The optimal calibration models are chosen based on the lowest validation ECE on training set. Additional calibration metrics, including class-wise ECE (Kull et al. 2019; Nixon et al. 2019), Kernel Density Estimation-based ECE (Zhang, Kailkhura, and Han 2020), Brier Score (Brier et al. 1950), and Negative Log-likelihood, are provided in the Appendix.

**Results on Small- and Medium-scale Graphs.** Table 2 shows that SIMI-MAILBOX outperforms baselines in 15 of 16 settings. Notably, our method pioneers in achieving an error rate below $3\%$ on Cora and Citeseer datasets, with a significant lead on Cora using GCN, breaking into the $1\%$ error range. SIMI-MAILBOX also demonstrates marked improvements on Pubmed and CS datasets, first achieving ECE reductions to within the $[0.5, 0.8]$ range. Even on Computers and Photo datasets, where the original predictions are already well-calibrated, SIMI-MAILBOX further reduces calibration errors to below $1\%$ with GAT. Additionally, consistent improvement is observed on the CoraFull dataset, with our method achieving the first 2% error range using GAT.

**Results on Large-scale Graphs.** To further demonstrate the versatility of our method, we extended our experiments to large-scale graphs, following the evaluation protocol in (Hu et al. 2020). We employed GCN and GraphSAGE (SAGE)(Hamilton, Ying, and Leskovec 2017), which are representative architectures for large-scale benchmark datasets. As shown in Table3, SIMI-MAILBOX outperforms all baselines to a considerable extent, achieving an error rate below 1% in all examined settings. This superiority is particularly notable in the Reddit dataset with SAGE, where our method reduces miscalibration by 10.57% compared to the uncalibrated baseline. In addition to calibration performance, we also measured the total execution time for each run, as presented in Table 4. Our method significantly improves time efficiency across all experiments, with a notable reduction in execution time on the Reddit dataset, decreasing by 61.86 and 182.10 seconds with GCN and SAGE. This




(a) CoraFull (CaGCN) (b) CoraFull (GATS) (c) CoraFull (Ours)

Figure 2: Qualitative analysis of our calibration results on CoraFull dataset, compared with CaGCN and GATS. Each cell in the heatmap represents the subtraction of the average confidence of calibrated nodes from the accuracy, with color and intensity indicating the magnitude of this discrepancy. Throughout diverse neighborhood similarity levels, our method facilitates a better reduction in the gap between accuracy and confidence compared to baselines.

gain is attributed to the simplified group-wise temperature approach, which allows for rapid optimization with only few parameters ($N$ clusters), in contrast to baselines that rely on complex GNNs for deriving node-wise temperature.

**SIMI-MAILBOX on Self-training.** In addition to improving calibration, calibrated predictions can be applied in self-training, utilizing pseudo-labels generated from unlabeled samples. As evidenced by (Rizve et al. 2021), poorly-calibrated models have a risk to choose pseudo-labeled samples with high confidence but incorrect classifications. Hence, confidence adjusted through calibration methods can lead to the selection of more accurate and high-confidence samples, improving classification accuracy. We broaden our evaluation of SIMI-MAILBOX to self-training scenarios, initially explored in CaGCN. Adhering to the same evaluation protocol in (Wang et al. 2021), we validate the effectiveness of our method in generating qualified pseudo-labels over baselines. Detailed results of this experiment are provided in the Appendix.

**Effectiveness on Diverse Neighborhood Topology.** To further validate the effectiveness of our method across different levels of neighborhood similarity, we present a qualitative comparison in Figure 2, utilizing a consistent dataset (CoraFull) and architecture (GCN) in preceding section. Similar to earlier analyses, the x-axis partitions nodes into intervals based on neighborhood prediction similarity, while the y-axis categorizes them by confidence intervals. Each cell in the heatmap represents the subtraction of average confidence of calibrated nodes from the accuracy. Deeper shades of **red** indicate that calibrated confidence is lower than accuracy (**under-confident**), while deeper shades of **green** signify that confidence exceeds accuracy (**over-confident**). Ideally, a perfectly calibrated model would produce a uniformly white heatmap, indicating perfect alignment between confidence and accuracy. As illustrated, SIMI-MAILBOX significantly reduces the discrepancy between accuracy and average confidence across varying similarity levels compared to baseline methods. This improvement is particularly evident in the patterns identified in the previous analysis, where our method mitigates discrepancies in the $\boldsymbol{s}(i) \in (0.1, 0.2]$ range within the $\hat{p}_i \in (0.9, 1.0]$ interval for CaGCN, and addresses the prevalent under-confidence observed with GATS in the $\boldsymbol{s}(i) \in (0.0, 0.4]$ range within the $\hat{p}_i \in (0.6, 0.8]$ intervals.

## Conclusion

In this study, we presented a novel analysis that identifies the limitations of uniform design principles in existing GNN calibration methods, particularly based on neighborhood similarity. To address these limitations, we proposed SIMI-MAILBOX, a novel calibration method that employs group-specific temperatures to refine miscalibration in nodes categorized by both neighborhood similarity and confidence. Comprehensive experiments have demonstrated the effectiveness of SIMI-MAILBOX, supported by extensive empirical and technical analysis. As for future work, we are dedicated to developing a theoretical foundation for our method.

## Acknowledgments

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)). We sincerely appreciate Polymerize for their generous support of our manuscript. Special thanks also go to Changhun Kim and Taewon Kim for constructive comments for this manuscript.

## References


Ahn, S. J.; and Kim, M. 2021. Variational graph normalized autoencoders. In *Proceedings of the 30th ACM international conference on information & knowledge management*, 2827–2831.

Angelopoulos, A.; Bates, S.; Malik, J.; and Jordan, M. I. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Bojchevski, A.; and Günnemann, S. 2017. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. *arXiv preprint arXiv:1707.03815*.

Brier, G. W.; et al. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1): 1–3.

Cauchois, M.; Gupta, S.; and Duchi, J. C. 2021. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22(81): 1–42.

Depeweg, S.; Hernandez-Lobato, J.-M.; Doshi-Velez, F.; and Udluft, S. 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, 1184–1193. PMLR.

Dusenberry, M.; Jerfel, G.; Wen, Y.; Ma, Y.; Snoek, J.; Heller, K.; Lakshminarayanan, B.; and Tran, D. 2020. Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning*, 2782–2792. PMLR.

Elinas, P.; Bonilla, E. V.; and Tiao, L. 2020. Variational inference for graph convolutional networks in the absence of graph data and adversarial settings. *Advances in neural information processing systems*, 33: 18648–18660.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hasanzadeh, A.; Hajiramezanali, E.; Boluki, S.; Zhou, M.; Duffield, N.; Narayanan, K.; and Qian, X. 2020. Bayesian graph neural networks with adaptive connection sampling. In *International conference on machine learning*, 4094–4104. PMLR.

Hébert-Johnson, U.; Kim, M.; Reingold, O.; and Rothblum, G. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, 1939–1948. PMLR.

Hou, Z.; Liu, X.; Dong, Y.; Wang, C.; Tang, J.; et al. 2022. Graphmae: Self-supervised masked graph autoencoders. *arXiv preprint arXiv:2205.10803*.

Hsu, H. H.-H.; Shen, Y.; and Cremers, D. 2022. A Graph Is More Than Its Nodes: Towards Structured Uncertainty-Aware Learning on Graphs. In *NeurIPS 2022 Workshop: New Frontiers in Graph Learning*.

Hsu, H. H.-H.; Shen, Y.; Tomani, C.; and Cremers, D. 2022. What Makes Graph Neural Networks Miscalibrated? *Advances in Neural Information Processing Systems*, 35: 13775–13786.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33: 22118–22133.

Huang, K.; Jin, Y.; Candes, E.; and Leskovec, J. 2024. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, 36.

Jiang, Z.; Araki, J.; Ding, H.; and Neubig, G. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9: 962–977.

Jung, S.; Seo, S.; Jeong, Y.; and Choi, J. 2023. Scaling of class-wise training losses for post-hoc calibration. In *International Conference on Machine Learning*, 15421–15434. PMLR.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kull, M.; Perello Nieto, M.; Kängsepp, M.; Silva Filho, T.; Song, H.; and Flach, P. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Lee, J. B.; Rossi, R.; and Kong, X. 2018. Graph classification using structural attention. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1666–1674.

Liu, T.; Liu, Y.; Hildebrandt, M.; Joblin, M.; Li, H.; and Tresp, V. 2022. On Calibration of Graph Neural Networks for Node Classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Ma, X.; and Blaschko, M. B. 2021. Meta-cal: Well-controlled post-hoc calibration by ranking. In *International Conference on Machine Learning*, 7235–7245. PMLR.

Maddox, W. J.; Izmailov, P.; Garipov, T.; Vetrov, D. P.; and Wilson, A. G. 2019. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32.

Minderer, M.; Djolonga, J.; Romijnders, R.; Hubis, F.; Zhai, X.; Houlsby, N.; Tran, D.; and Lucic, M. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34: 15682–15694.

Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P.; and Dokania, P. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33: 15288–15299.

Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Nixon, J.; Dusenberry, M. W.; Zhang, L.; Jerfel, G.; and Tran, D. 2019. Measuring Calibration in Deep Learning. In *CVPR workshops*, volume 2.

Pal, S.; Regol, F.; and Coates, M. 2019. Bayesian graph convolutional neural networks using non-parametric graph learning. *arXiv preprint arXiv:1910.12132*.

Park, J.; Song, J.; and Yang, E. 2021. Graphens: Neighbor-aware ego network synthesis for class-imbalanced node classification. In *International Conference on Learning Representations*.

Perez-Lebel, A.; Morvan, M. L.; and Varoquaux, G. 2022. Beyond calibration: estimating the grouping loss of modern neural networks. *arXiv preprint arXiv:2210.16315*.

Platt, J.; et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3): 61–74.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33: 3581–3591.

Rong, Y.; Huang, W.; Xu, T.; and Huang, J. 2019. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*.

Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Galligher, B.; and Eliassi-Rad, T. 2008. Collective classification in network data. *AI magazine*, 29(3): 93–93.

Shchur, O.; Mumme, M.; Bojchevski, A.; and Günnemann, S. 2018. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*.

Shi, S.; Chen, J.; Qiao, K.; Yang, S.; Wang, L.; and Yan, B. 2022. Select and Calibrate the Low-confidence: Dual-Channel Consistency based Graph Convolutional Networks. *arXiv preprint arXiv:2205.03753*.

Stadler, M.; Charpentier, B.; Geisler, S.; Zügner, D.; and Günnemann, S. 2021. Graph posterior network: Bayesian predictive uncertainty for node classification. *Advances in Neural Information Processing Systems*, 34: 18033–18048.

Sui, Y.; Wang, X.; Wu, J.; Lin, M.; He, X.; and Chua, T.-S. 2022. Causal attention for interpretable and generalizable graph classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1696–1705.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Wang, M.; Yang, H.; and Cheng, Q. 2022. GCL: Graph Calibration Loss for Trustworthy Graph Neural Network. In *Proceedings of the 30th ACM International Conference on Multimedia*, 988–996.

Wang, X.; Liu, H.; Shi, C.; and Yang, C. 2021. Be confident! towards trustworthy graph neural networks via confidence calibration. *Advances in Neural Information Processing Systems*, 34: 23768–23779.

Wen, Y.; Tran, D.; and Ba, J. 2020. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*.

Wu, Q.; Chen, Y.; Yang, C.; and Yan, J. 2023. Energy-based out-of-distribution detection for graph neural networks. *arXiv preprint arXiv:2302.02914*.

Xing, C.; Arik, S.; Zhang, Z.; and Pfister, T. 2019. Distance-based learning from errors for confidence calibration. *arXiv preprint arXiv:1912.01730*.

Xu, K.; Li, C.; Tian, Y.; Sonobe, T.; Kawarabayashi, K.-i.; and Jegelka, S. 2018. Representation learning on graphs with jumping knowledge networks. In *International conference on machine learning*, 5453–5462. PMLR.

Yang, J.-Q.; Zhan, D.-C.; and Gan, L. 2023. Beyond Probability Partitions: Calibrating Neural Networks with Semantic Aware Grouping. *arXiv preprint arXiv:2306.04985*.

Yun, S.; Kim, S.; Lee, J.; Kang, J.; and Kim, H. J. 2021. Neo-gnns: Neighborhood overlap-aware graph neural networks for link prediction. *Advances in Neural Information Processing Systems*, 34: 13683–13694.

Zargarbashi, S. H.; Antonelli, S.; and Bojchevski, A. 2023. Conformal prediction sets for graph neural networks. In *International Conference on Machine Learning*, 12292–12318. PMLR.

Zeng, H.; Zhou, H.; Srivastava, A.; Kannan, R.; and Prasanna, V. 2019. Graphsaint: Graph sampling based inductive learning method. *arXiv preprint arXiv:1907.04931*.

Zhang, J.; Kailkhura, B.; and Han, T. Y.-J. 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*, 11117–11128. PMLR.

Zhang, M.; and Chen, Y. 2018. Link prediction based on graph neural networks. *Advances in neural information processing systems*, 31.

Zhao, X.; Chen, F.; Hu, S.; and Cho, J.-H. 2020. Uncertainty aware semi-supervised learning on graph data. *Advances in Neural Information Processing Systems*, 33: 12827–12836.

Zhu, Z.; Zhang, Z.; Xhonneux, L.-P.; and Tang, J. 2021. Neural bellman-ford networks: A general graph neural network framework for link prediction. *Advances in Neural Information Processing Systems*, 34: 29476–29490.