Improving GNN Calibration with Discriminative Ability: Insights and Strategies

Yujie Fang[1], Xin Li*[1], Qianyu Chen[1], Mingzhong Wang[2]

[1] Beijing Institute of Technology
[2] University of the Sunshine Coast
{fangyujie,xinli,qychen}@bit.edu.cn, mwang@usc.edu.au

Abstract

The widespread adoption of Graph Neural Networks (GNNs) has led to an increasing focus on their reliability. To address the issue of underconfidence in GNNs, various calibration methods have been developed to gain notable reductions in calibration error. However, we observe that existing approaches generally fail to enhance consistently, and in some cases even deteriorate, GNNs' ability to discriminate between correct and incorrect predictions. In this study, we advocate the significance of discriminative ability and the inclusion of relevant evaluation metrics. Our rationale is twofold: 1) Overlooking discriminative ability can inadvertently compromise the overall quality of the model; 2) Leveraging discriminative ability can significantly inform and improve calibration outcomes. Therefore, we thoroughly explore the reasons why existing calibration methods have ineffectiveness and even degradation regarding the discriminative ability of GNNs. Building upon these insights, we conduct GNN calibration experiments across multiple datasets using a straightforward example model, denoted as DC(GNN). Its excellent performance confirms the potential of integrating discriminative ability as a key consideration in the calibration of GNNs, thereby establishing a pathway toward more effective and reliable network calibration.

Introduction

Graph Neural Networks (GNNs) have become a popular choice across various domains due to the ubiquitous presence of graph data and their exceptional performance. As GNNs gain increasing applications in safety-critical fields, such as autonomous driving (Weng et al. 2020), healthcare (Wang et al. 2023), and finance (Wang et al. 2022), their reliability has become a major concern. Contrary to the overconfidence issue in predictions made by conventional deep learning models, (Wang et al. 2021; Hsu et al. 2022) reveal that GNNs tend to exhibit a noteworthy trend of underconfidence. To address the issue of underconfidence in GNNs, primarily two strategies have been proposed to calibrate GNNs. One strategy integrates calibration-related loss functions into the training phase of GNNs, while another strategy adopts post-hoc (or post-processing) calibration techniques, applying them after GNNs have been trained.

*Corresponding author.

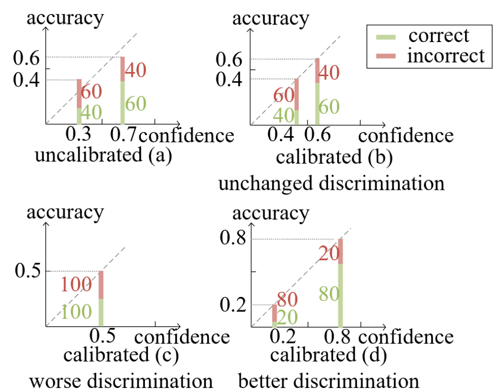


Figure 1: (a) presents the original model. The diagonal line indicates perfect calibration. Green for correct, red for incorrect; numbers indicate sample counts. (b), (c) and (d) showcase three different perfect calibration results with the same accuracy. Although they all achieve 0 calibration error, they exhibit different discriminative abilities between correct and incorrect predictions.

While traditional calibration evaluation metrics, such as ECE (Expected Calibration Error), may suggest good calibration results by these methods, they assume that well-calibrated confidence should accurately reflect the likelihood of correctness (DeGroot and Fienberg 1983). However, this assumption might not always be sufficient for evaluating a calibration model. As shown in Fig. 1, all three calibration methods can reduce the ECE to 0 while keeping the same classification accuracy, even though they have different discriminative abilities between correct and incorrect predictions: steady in (b), decreased in (c), and improved in (d).

Therefore, we argue that an effective calibration should also prioritize enhancing the discriminative ability in predictions, thus ensuring better alignment between confidence and accuracy. In essence, we propose that a good calibration model should strive to achieve two goals: reduce calibration error while simultaneously improving discriminative ability.

Given that AUROC (Area Under the Receiver Operating Characteristic) is a common measure of discriminative ability, we conducted an empirical study to evaluate existing GNN calibration methods for their calibration and discrimi-

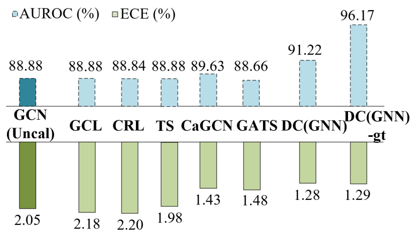


Figure 2: The calibration error (ECE) and AUROC of calibrated confidence obtained by various GNN calibration methods on the CS dataset.

native abilities. As depicted in Fig. 2, except DC(GNN) and DC(GNN)-gt, they have none or negligible improvement in discriminative ability between correct and incorrect predictions (notably better AUROC compared with the original GCN outputs) regardless of their achievement in calibrated error. The results are quite counterintuitive, especially for methods like CaGCN (Wang et al. 2021) and GATS (Hsu et al. 2022), which intentionally employ the negative log-likelihood (NLL) loss to elevate the confidence of correct predictions and lower the confidence of erroneous predictions, thus being supposed to achieve higher AUROC.

Based on these observations, this paper aims to comprehensively investigate the reasons behind the failure of existing calibration methods. We identify that the causes include over-calibration, insufficient discriminative information, and overlooked distributional shifts. By addressing these concerns and emphasizing the discriminative ability of the calibration model, we develop a simple yet effective approach, DC(GNN) – a Discriminative Calibration model for GNNs. Experimental results across multiple datasets confirm that our DC(GNN) both reduces calibration error and enhances the discriminative ability.

Related Work

Deep Neural Network (DNN) Model Calibration

Various DNN calibration methods can be categorized into three classes (Gawlikowski et al. 2021): **Regularization methods applied during the training** modify the objective function or augment the training data to achieve calibration. For instance, label smoothing (Szegedy et al. 2016) effectively mitigates overconfidence. Similarly, introducing a negative entropy loss (Pereyra et al. 2017) can also achieve the same effect by penalizing high confidence values. **Post-hoc methods applied post-training** train the calibration model on a held-out validation set. Most post-hoc methods stem from temperature scaling (TS) (Guo et al. 2017), which rescales the output logits of the neural network with a temperature T. **Neural network uncertainty estimation methods** quantify both model and data uncertainty, typically using deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017; Mehrtash et al. 2020) and Bayesian methods (Izmailov et al. 2020; Foong et al. 2019).

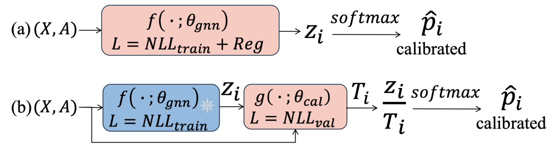


Figure 3: GNN calibration: (a) Regularization (b) Post-hoc.

GNN Calibration

While DNNs lean towards overconfidence, GNNs tend to exhibit underconfident. In GNN calibration, there are generally two types of loss functions that regularize the training phase: the graph calibration loss (GCL) (Wang, Yang, and Cheng 2022) and the confidence-reward loss (CRL) (Liu et al. 2022). The post-hoc calibration methods in GNNs, such as CaGCN (Wang et al. 2021), RBS (Liu et al. 2022), and GATS (Hsu et al. 2022), train additional networks on the validation set. They use the GNN output, such as logits, as the input for the calibration model and generate node-specific temperature coefficients. These methods differ in leveraging different factors to serve as the inputs. For instance, GATS utilizes a set of factors, including GNN homophily, to guide calibration model training.

A recent study (Zhu et al. 2022) observed that many regularization methods for DNN, which aim at alleviating overconfidence, inadvertently result in declined performance in failure prediction. This finding enforces the rationale behind our advocacy for the inclusion of promoting discriminative ability in calibration methods. In this paper, we start by understanding the underlying reasons for this observation, particularly in the context of GNN calibration. Thereafter, we focus on post-hoc calibration methods, which theoretically should offer superior discrimination ability compared to regularization methods. However, we discover that existing post-hoc methods have not achieved the desired level of discrimination ability, necessitating a more in-depth analysis to uncover the factors responsible for this limitation.

Problem Formulation

For consistent comparison, following the most prevailing methods (Hsu et al. 2022; Yang et al. 2022; Liu et al. 2022; Wang et al. 2021), the model proposed in this paper calibrates GNNs on semi-supervised node classification tasks.

Node Classification. Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $\mathcal{V} = \{v_1, \ldots, v_N\}$ is the set of nodes and $\mathcal{E}$ is the set of edges. $N = |\mathcal{V}|$ denotes the number of nodes in the graph. $A \in \{0,1\}^{N\times N}$ is the adjacent matrix that represents the connectivity between nodes. The feature and label of node v_i are represented as $x_i \in \mathbb{R}^d$ and $y_i \in \{1, \ldots, K\}$, respectively, where K is the number of classes. The output of the GNN model f parameterized by θ_{gnn} is expressed as $Z = f(X, A; \theta_{gnn})$, where $X = [x_i, \ldots, x_N]^T \in \mathbb{R}^{N\times d}$ is the input feature matrix, and $Z = [z_i, \ldots, z_N]^T \in \mathbb{R}^{N\times K}$ is the output logits. $\hat{p}_i = softmax(z_i)$ is the class probabilities for node v_i. $\hat{y}_i = argmax_y \hat{p}_{i,y}$ and $\hat{c}_i = max_y \hat{p}_{i,y}$ are the prediction and confidence for node v_i, respectively.

GNN Calibration. Fig. 3 illustrates the procedures for two primary categories of GNN calibration: (a) Regularization calibration methods and (b) Post-hoc calibration methods.

Metrics. We propose to use both ECE and AUROC to facilitate a comprehensive evaluation of calibration.

ECE measures the absolute difference between accuracy and confidence. We partition the range $[0, 1]$ into M equally spaced confidence intervals, with $B_m \in \{B_1, \ldots, B_M\}$ denoting the set of examples/nodes whose confidence falls in the m^{th} interval. ECE is defined as:

$$\sum_{m=1}^{M} \frac{|B_m|}{N} |acc(B_m) - conf(B_m)|. \tag{1}$$

where $acc(B_m)$ and $conf(B_m)$ are the accuracy and confidence of the model on B_m, respectively. AUROC serves as a metric to evaluate the discriminative ability of a model. It quantifies the model's performance by plotting the true positive rate against the false positive rate at various thresholds. The AUROC's value, ranging between 0 and 1, represents the area under this curve. A value closer to 1 indicates higher discriminative ability.

Deficiency Analysis for Calibration Failures

In this section, we identify three prominent reasons why existing calibration methods may fail to properly discriminate between correct and erroneous predictions.

Over-calibration

In scenarios where the model is overconfident on the overall dataset, calibration methods often attempt to reduce the model's confidence through parameter optimization. However, as correct predictions tend to have higher confidence scores than incorrect predictions, this approach can disproportionately reduce the confidence of correct predictions while leaving incorrect predictions relatively unchanged. As a result, the calibration may in fact worsen the correct predictions, even if it improves overall calibration performance.

Specifically, regularization losses, such as entropy augmentation, tend to show a bias toward reducing the confidence of data points with high confidence. This phenomenon can be exemplified by considering the gradient formula of entropy $\mathcal{H}$ within a binary classification context, given as $\frac{\partial(-\mathcal{H}(\hat{p}_i))}{\partial\theta_{gnn}} = \log\frac{\hat{c}_i}{1-\hat{c}_i} \times \frac{\partial\hat{c}_i}{\partial\theta_{gnn}}$, where the term $\log\frac{\hat{c}_i}{1-\hat{c}_i}$ is monotonically increasing with $\hat{c}_i$, indicating that higher confidence values ($\hat{c}_i$) get greater weight during gradient computation. As a result, the confidence of correct predictions would decrease more significantly than that of incorrect predictions, leading to greater overlap between the confidence distributions of both. Ultimately, the ability to differentiate between correct and incorrect predictions of the calibrated model is compromised.

In scenarios where the model is underconfident across the dataset, calibration methods primarily focus on increasing the confidence. However, similar to the previous issue, this approach can disproportionately increase the confidence of incorrect predictions against correct predictions. For instance, the entropy-penalty loss used in GCL is designed to assign greater weight to low-confidence data during gradient computation, which can result in a greater increase in confidence for incorrect predictions than for correct predictions. This also results in a reduction in the discriminative ability.

Remark 1 *Regularization calibration methods, due to their inherent biases, inadvertently reduce the model's discriminative ability between correct and incorrect predictions.*

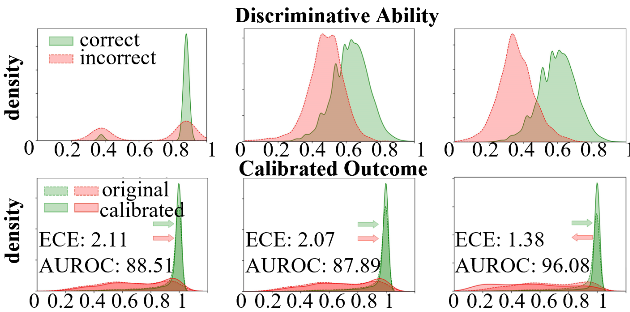
Insufficient Discriminative Information

Among all calibration methods, post-hoc calibration approaches are theoretically the most likely to achieve an increase in the discriminative ability between correct and incorrect predictions. This potential advantage stems from their training mechanism, which utilizes the NLL loss on the validation set. The NLL loss is inherently designed to elevate the confidence of correct predictions while reducing the confidence of incorrect predictions. Regrettably, as shown in Fig. 2, existing methods of this kind have also failed to achieve this goal. The ineffectiveness of these methods indicates that they have not fully followed the supervision provided by the NLL loss. In this section, we elucidate the underlying reasons behind this concerning phenomenon.

Firstly, the non-data-specific post-hoc calibration methods, such as TS, Vector Scaling (VS) (Guo et al. 2017) and Ensemble Temperature Scaling (ETS) (Zhang, Kailkhura, and Han 2020), are designed to calibrate either the entire dataset or specific classes as a whole, without accommodating the calibration at finer data granularity. The calibration process employed by these methods often neglects essential information needed to discriminate between correct and incorrect predictions. Hence, although they may succeed in reducing calibration error by aligning with the overall calibration trends of the dataset or specific classes, they cannot improve the model's discriminative ability. Taking TS as an example, it assigns the same temperature coefficient to all data points, thereby failing to achieve a more distinguished calibrated confidence among individual data points.

Regarding data-specific post-hoc methods, such as GATS, they incorporate individual-level factors as input to their calibration models. As indicated in related literature, these factors are strongly correlated with a reduction in calibration error. Their limited effectiveness in enhancing discriminative ability can be attributed to the insufficient discriminative information contained within their input factors. As depicted in the top row of Fig. 4, we can observe that the distributions of factors utilized by existing calibration methods show substantial overlap between correct and incorrect predictions, indicating a compromised discriminative ability. The second row of the figure demonstrates that the corresponding calibration results indeed struggle to effectively separate the confidence distributions for correct and incorrect predictions. However, it is noteworthy that incorporating information with sufficient discriminative power, such as ground-truth homophily, helps to achieve more appropriate calibration outcomes. This injection of discriminative information can remarkably enhance discriminative ability along with a significant reduction in calibration error.

Remark 2 *Insufficient discriminative information is the*



(a) binned-ratio in RBS (b) GNN homophily in GATS (c) ground-truth homophily

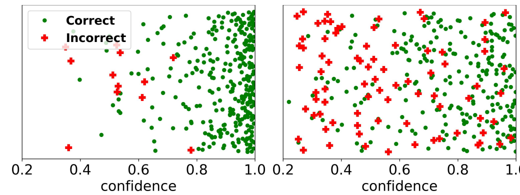
Figure 4: Discriminative abilities of different factors in existing methods, comparing to ground-truth homophily[1]. Second row: calibrated outcomes, ECE, AUROC using first-row factors for temperature coefficient via MLP on CS dataset.

primary reason why post-hoc calibration methods deviate from the objectives set by their training NLL loss. Incorporating adequate discriminative information enables more effective and appropriate calibration.

Overlooked Distribution Shifts

Calibration models generally rely on the outputs of the original model as their inputs. For instance, some calibration models directly employ the logits from the original model, while others utilize factors derived from these logits, which we denote as z for simplicity. In this section, we emphasize a critical aspect that has been largely overlooked - the significant distribution shift of z between the training and test datasets. As illustrated in Fig. 5, the training data typically exhibits a concentration of confidence values around 1, with the majority of training instances being correctly predicted. In contrast, these two features are not evident in the test dataset due to the shift in data distributions. However, some existing calibration models fail to take this issue into consideration and rely on the training data to select hyperparameters. Due to the predominance of correctly predicted instances in the training data, these models inadvertently become biased in hyperparameter selection. Specifically, it tends to focus on increasing confidence, often at the expense of neglecting necessary confidence reduction. Consequently, the calibration results may merely elevate the confidence across the entire dataset, rather than ensuring that the increase in confidence for correctly predicted instances exceeds that for incorrectly predicted ones.

[1]Following GATS, we compute node homophily as $H_i = \log\left(\frac{n_a+1}{n_d+1}\right)$, which quantifies the proportion of neighbors of node v_i that share the same label. Specifically, n_a and n_d represent the number of neighbors of v_i that have the same and different labels, respectively. This paper considers two variants of node homophily: GNN homophily, which uses the predicted labels of the GNN to determine n_a and n_d, and ground-truth homophily, which uses the real labels to compute n_a and n_d. For enhanced comparability between factors, we normalize these factors to the range of [0, 1].



(a) training set (b) test set

Figure 5: Distribution disparities in GNN outputs between training and test datasets, emphasizing: 1) factors like confidence distribution, and 2) correct sample ratio. This observation is derived from GCN's ouput on the Citeseer dataset.

Remark 3 *The distributional shift of the original model's outputs between the training and test sets can result in biased calibration outcomes for certain calibration models.*

Proposed Methodology

Based on the previous analysis, we devise a post-hoc calibration method for GNNs, namely DC(GNN), which aims to reduce calibration error while enhancing the discriminative ability. We aim to achieve calibration by leveraging signals with high discriminative ability. To approximate the discriminatory effect of the ground-truth information, signals need to be sufficiently informative. In addition, these signals can also complement each other across different datasets, thus enhancing the robustness of the calibration process.

Existing but Overlooked Discriminative Signals

We first identify three signals derived from the GNNs' outputs. Although they have been proposed in prior work, they have not been appropriately incorporated into existing calibration models.

- The **GNN output confidence** $C \in [0,1]^N$ can be viewed as the softened GNN prediction and is the target for calibration. As an input, it can determine the overall trend of calibration, and $C_i = \hat{c}_i$.
- The **GNN same-class-neighbor ratio** (Liu et al. 2022), denoted as $R \in [0,1]^N$, aims to utilize output probabilities to better approximate the ground-truth homophily. The nodewise formula is as follows:

$$\forall v_i \in \mathcal{V},\ R_i = \frac{1}{|\mathcal{N}_i|} \sum_{v_j \in \mathcal{N}_i} \hat{p}_{j,\hat{y}_i}, \tag{2}$$

  where $\mathcal{N}_i$ denotes the set of neighbors of node v_i, and $\hat{p}_{j,\hat{y}_i}$ corresponds to the probability that GNN predicts node v_j belonging to the same class $\hat{y}_i$ of node v_i .
- The **uncertainty mass** (Jøsang 2016) $U \in [0,1]^N$ represents the vacuity of evidence for a prediction. While the first two signals rely on probability distributions to portray first-order uncertainty, uncertainty mass represents a form of second-order uncertainty. Its value can represent the uncertainty of first-order probabilities, providing an alternative perspective for discrimination information.

$$\forall v_i \in \mathcal{V}, U_i = \frac{K}{K + \sum_{k=1}^{K} exp(z_{i,k})}, \tag{3}$$

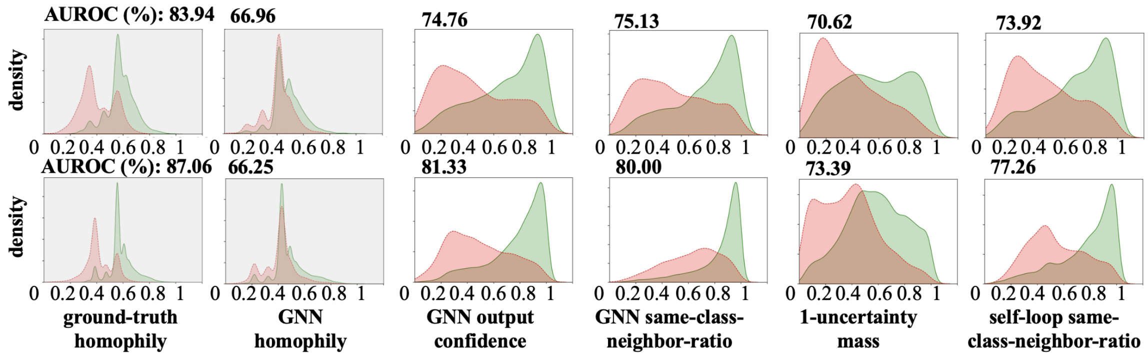


Figure 6: Discriminative abilities of signals in distinguishing correct and incorrect predictions, depicted by distribution separation, are also reported through AUROC values. Lesser overlap indicates higher discriminative ability. To align with the general trend where higher signal values correlate with correct predictions, we visualize 1-uncertainty mass rather than uncertainty mass. The first and second rows of the figure are from the Citeseer and Pubmed datasets, respectively.

where z_i corresponds to the logits of node v_i in the output of the GNN.

Fig. 6 depicts the discriminative ability of different signals by presenting the distribution of correctly and incorrectly predicted nodes on the Citeseer and Pubmed datasets. The degree of overlap between the two distributions indicates the effectiveness of a signal in differentiating between correct and incorrect predictions. It is evident that all three existing signals we have selected can effectively reduce the overlap area compared with GNN homophily commonly adopted in GNN calibration models, e.g., (Hsu et al. 2022).

Proposed (Novel) Signal: Self-Loop Same-Class-Neighbor Ratio

Fig. 7 presents a concrete example from the Citeseer dataset. Whichever signal from the previous section or their combination is used, node v_{685} is always, albeit incorrectly, inferred to be a correctly predicted node. Due to the neighborhood aggregation mechanism of GNNs, they are susceptible to being misled by the dominant class in a node's neighborhood, leading to erroneous predictions. Moreover, this issue is generally undetectable through factors solely based on GNN output information. To overcome the potential misguidance arising from neighborhood information, we propose the **self-loop same-class-neighbor ratio**, denoted as $R^{sl} \in [0,1]^N$, which is computed using a self-loop mechanism. The output class probabilities $S \in [0,1]^{N \times K}$ in the self-loop mechanism are obtained as follows:

$$S = softmax(f(X, I; \theta_{gnn})), \tag{4}$$

where I is the identity matrix. The self-loop same-class-neighbor ratio R^{sl} is computed for each node as follows:

$$\forall v_i \in \mathcal{V},\ R_i^{sl} = \frac{1}{|\mathcal{N}_i|} \sum_{v_j \in \mathcal{N}_i} \hat{s}_{j,\hat{y}_i}, \tag{5}$$

where $\hat{s}_{j,\hat{y}_i}$ denotes the probability assigned by the self-loop mechanism to node v_j as belonging to class $\hat{y}_i$.

Fig. 7 illustrates how the self-loop same-class-neighbor-ratio provides a more accurate assessment of whether misclassification has occurred. The effectiveness of this signal

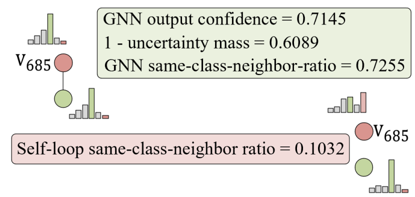


Figure 7: Visualization of signal values for node v_{685} in Citeseer dataset. Node colors correspond to real labels. The classification probabilities from GCN (left) and the self-loop mechanism (right) are represented with bars, with the GNN predicting v_{685} as the green class (highest prob. for the green class). Node v_{685} is a misclassified node undetectable by three signals derived from the GNN (green box), but is identifiable by the self-loop same-class-neighbor ratio (red box).

is further supported by the results shown in Fig. 6, which presents the AUROC scores achieved by the self-loop same-class-neighbor ratio. The self-loop mechanism is designed to provide calibration information by emphasizing node features and neglecting possible misguidance from neighborhoods. This approach takes advantage of the trained parameters of the GNN to update the node representations, thus eliminating the need for additional parameters.

Scaling Temperature

In this paper, we concatenate the discriminative features previously described as input to train the calibration model g. The output of the calibration model is a temperature vector $T \in \mathbb{R}^N$ that corresponds to all examples/nodes, denoted as:

$$T = g(C||R||U||R^{sl}; \theta_{cal}), \tag{6}$$

where g is a neural network parameterized by θ_{cal}, and in our experiments, we used a simple MLP (Multi-Layer Perceptron). By using the nodewise logit z_i obtained from the GNN output and the temperature T_i generated from the calibration model (as described in Fig. 3), we can compute calibrated node-level predictions. It is worth noting that a key contribution of our work is the strategic proposition for designing more appropriate calibration methods—leveraging signals with high discriminative ability. As a result, our methodology is not constrained to the use of a simple MLP that takes these four specific signals as input. Generally, it is adaptable and can be extended to encompass a broader range of signals with high discriminative ability, as well as to accommodate more complex calibration network structures.

Experiments

This section evaluates GNN calibration methods on a variety of datasets (see Appendix). The experiments utilized two representative GNN backbones: GCN (Kipf and Welling 2016) and GAT (Veličković et al. 2017).

Experimental Settings

Baselines. We compared our model, DC(GNN), with eight baseline models that represent various categories of calibration techniques as follows:

- Regularization calibration methods: Graph Calibration Loss (**GCL**) and Confidence-Reward Loss (**CRL**).
- Non-node-specific Post-hoc calibration methods: Temperature Scaling (**TS**), Vector Scaling (**VS**), and Ensemble Temperature Scaling (**ETS**).
- Node-specific Post-hoc calibration methods: **CaGCN**, **RBS**, and Graph Attention Temperature Scaling (**GATS**).

In addition, we include the results from the uncalibrated backbone (**Uncal**) for reference.

Implementation Details. For GCL, CRL, and RBS, we performed a grid search to find the optimal hyperparameters, such as the coefficient of the regularization term. For other baseline models, we followed the experimental settings in GATS. We conducted experiments on 8 benchmark datasets, with each experiment consisting of 5 splits (train/val/test: 10%-5%-85%). We initialized the model 5 times for each split and employed 3-fold internal cross-validation. In total, we performed 75 runs on each dataset per model. In addition to ECE and AUROC, we also report the accuracy of two regularization methods. The additional experimental details are provided in the Appendix.

Performance Comparison

Table 1 presents the calibration results with the GCN backbone. Results for the GAT backbone are in Appendix. As observed, DC(GNN) achieves the best overall calibration results, yielding both excellent ECE values and substantial improvements in AUROC. A thorough analysis follows below.

Discriminative ability-driven approach enables more appropriate calibration. Compared to Uncal, some calibration methods achieve a reduction in ECE but worse AUROC performance. For instance, CRL on datasets Pubmed and Physics, and CaGCN on datasets Cora and Citeseer, demonstrate this behavior. These methods reduce calibration error at the cost of sacrificing the inherent quality of GNNs, making them not a favorable trade-off. A notable observation is that DC(GNN) remarkably improves AUROC. On the CS and Physics datasets, DC(GNN) outperforms the second-best method by 1.59% and 0.69%, respectively. Even the most competitive baselines, GATS and CaGCN, cannot consistently perform well across all datasets. However, DC(GNN) achieves the best results in terms of both ECE and AUROC on nearly all datasets. This confirms that the discriminative ability-driven calibration strategy, as exemplified by DC(GNN), is both robust and remarkably effective. It not only achieves the alignment of accuracy and confidence but also enhances the overall quality of GNN.

Regularization offers limited efficacy in GNN calibration. Compared to post-hoc calibration methods, the two regularization-based GNN calibration approaches, GCL and CRL, neither significantly reduce ECE nor improve AUROC. In fact, they even underperform on datasets like Cora and Citeseer compared to uncalibrated results. Moreover, they fail to achieve higher classification accuracy. These phenomena could be attributed to the regularization terms introduced by GCL and CRL. These terms, aimed at increasing confidence in the training set, seem to have a marginal influence on GNNs due to high confidence levels already achieved under the original classification NLL loss. This also explains why the drawbacks of over-calibration associated with regularization methods have not been observed; both methods maintain stable AUROC across most datasets. In summary, these regularization-based calibration methods exhibit limited effectiveness in mitigating underconfidence.

Node-specific calibration is effective. Non-node-specific calibration methods, such as TS, VS, and ETS, which apply uniform scaling to logits, show limitations in reducing ECE and only get marginal improvement in AUROC. In comparison, CaGCN, GATS, and DC(GNN), which consider node neighborhood information, provide node-specific temperature assignment, resulting in lower overall calibration errors. Therefore, node-specific calibration proves to be an effective strategy. Importantly, our DC(GNN) shares the same simplicity as non-node-specific calibration methods.

In summary, DC(GNN)'s effectiveness comes from the use of discriminative signals, eliminating the need for intricate and specialized designs. The concise and straightforward nature of DC(GNN) makes it broadly applicable to various datasets and reliably performs in most scenarios.

Ablation Study

To further validate the effectiveness of each signal, we conducted ablation studies by incrementally adding signals as inputs to the DC(GNN). Table 2 presents the results on four datasets; additional results are in Appendix. As evident from the overall trends depicted in the table, incorporating more signals leads to superior and more robust performance in terms of AUROC and ECE. These results affirm the rationale behind our model's utilization of the combined set of discriminative signals. There are indeed a few instances where the performance decreases after adding a specific signal. This discrepancy arises from different signals having distinct strengths, making it challenging for a single signal to be universally applicable across all datasets. Furthermore, observing adding the self-loop same-class neighbor ratio R^{sl} highlights the impact of our newly introduced signal R^{sl} on AUROC across datasets, such as Pubmed (81.62% to 82.47%)

Model	Dataset	Cora	Citeseer	Pubmed	Computers	Photo	CS	Physics	CoraFull
Uncal	ACC	83.81±1.05	71.16±0.63	86.60±0.19	88.79±0.66	92.69±0.50	92.15±0.24	95.41±0.26	64.22±0.58
	ECE	6.13±4.75	8.32±5.03	4.31±1.67	3.07±0.89	2.24±1.95	2.05±0.50	1.54±0.49	7.99±0.93
	AUROC	83.88±1.11	76.01±0.83	81.33±0.48	82.49±0.80	88.07±0.98	88.88±0.57	91.19±0.73	77.43±0.48
GCL	ACC	83.80±1.02	71.18±0.64	86.69±0.18	89.00±0.60	92.78±0.37	92.15±0.25	95.41±0.26	64.23±0.59
	ECE	6.29±4.63	8.76±6.59	1.37±0.38	2.30±0.36	1.65±0.53	2.18±0.60	1.55±0.48	7.87±0.64
	AUROC	83.87±1.12	76.06±0.92	81.29±0.50	82.71±0.72	88.44±0.67	88.88±0.62	91.20±0.72	77.46±0.45
CRL	ACC	83.88±1.05	71.09±0.64	85.73±0.43	87.22±0.69	92.79±0.31	92.12±0.30	95.24±0.24	64.29±0.53
	ECE	6.99±4.69	9.34±7.40	2.00±0.85	2.28±0.59	1.77±0.78	2.20±0.54	1.40±0.41	7.81±0.63
	AUROC	83.70±1.15	75.99±0.95	79.71±0.61	82.12±0.89	88.24±0.74	88.84±0.69	90.78±0.62	77.50±0.46
TS	ECE	3.93±1.45	5.29±1.22	1.29±0.29	2.53±0.52	1.54±0.39	1.98±0.45	1.15±0.45	7.67±0.50
	AUROC	84.18±1.06	76.16±0.82	81.47±0.45	82.61±0.75	88.21±0.79	88.88±0.55	91.15±0.71	77.47±0.44
VS	ECE	3.80±1.41	5.41±1.11	1.43±0.34	2.66±0.55	1.61±0.46	1.89±0.51	1.17±0.49	7.64±0.47
	AUROC	84.24±1.03	76.27±0.92	81.41±0.47	82.62±0.74	88.12±0.85	89.14±0.41	91.29±0.59	77.52±0.43
ETS	ECE	3.87±1.44	5.26±1.21	1.28±0.30	2.40±0.56	1.45±0.42	1.75±0.45	0.95±0.35	7.57±0.50
	AUROC	84.18±1.06	76.16±0.82	81.47±0.45	82.61±0.75	88.21±0.79	88.88±0.55	91.15±0.71	77.47±0.44
CaGCN	ECE	5.22±1.60	6.53±1.49	<u>1.05±0.41</u>	<u>1.85±0.57</u>	2.01±0.70	<u>1.43±0.39</u>	0.63±0.38	7.39±1.80
	AUROC	82.97±1.58	75.47±1.10	<u>82.30±0.74</u>	**83.85±1.08**	87.02±1.39	<u>89.63±0.50</u>	<u>91.95±0.70</u>	77.99±0.99
RBS	ECE	4.92±2.01	6.90±2.30	1.51±0.38	2.44±0.52	2.21±0.59	2.50±0.47	1.19±0.46	11.19±1.69
	AUROC	84.04±1.29	76.21±0.85	81.53±0.44	82.89±0.73	88.31±0.81	88.97±0.51	91.17±0.68	<u>78.03±0.43</u>
GATS	ECE	<u>3.55±1.46</u>	<u>4.61±1.35</u>	**1.03±0.33**	2.19±0.47	<u>1.34±0.34</u>	1.48±0.28	<u>0.58±0.12</u>	<u>5.47±0.80</u>
	AUROC	**84.32±1.14**	<u>76.36±0.81</u>	81.80±0.54	82.78±0.75	<u>88.39±0.70</u>	88.66±0.50	91.04±0.68	77.71±0.49
DC(GNN)	ECE	**3.35±1.50**	**3.94±1.41**	1.18±0.35	**1.82±0.59**	**1.30±0.49**	**1.02±0.33**	**0.54±0.17**	**2.49±0.56**
	AUROC	<u>84.25±1.28</u>	**76.51±1.36**	**82.76±0.50**	<u>83.20±0.76</u>	**89.01±0.81**	**91.22±0.56**	**92.64±0.56**	**78.37±0.63**

Table 1: GCN calibration results (mean ± std), presented in terms of ECE (%, lower is better) and AUROC (%, higher is better). The best and runner-up results are highlighted with bold and underline, respectively.

and CS (88.70% to 90.64%). Unlike signals derived from GNN output, the self-loop mechanism, focusing on the intrinsic information of nodes, provides a unique perspective for GNN calibration. We also present a variant of our model, DC(GNN)-gt, which utilizes ground truth homophily as the input signal. It achieves impressive calibration results across datasets, although it is outperformed by DC(GNN) on certain datasets. This observation implies that the aggregation of multiple discriminative signals can enhance the robustness of the calibration model.

Model	Dataset	Pubmed	Computers	CS	CoraFull
C	ECE	1.33±0.38	2.40±0.61	1.20±0.33	3.38±0.92
	AUROC	81.47±0.45	82.58±0.74	88.65±0.51	77.38±0.50
$C\&R$	ECE	1.28±0.35	2.37±0.58	1.18±0.33	3.16±0.72
	AUROC	81.62±0.49	82.68±0.73	88.70±0.56	77.21±0.63
$C\&R$	ECE	1.40±0.34	2.36±0.59	1.06±0.35	3.11±0.77
$\&R^{sl}$	AUROC	82.47±0.53	82.72±0.80	90.64±0.63	77.43±0.75
DC	ECE	1.18±0.35	1.82±0.59	1.02±0.33	2.49±0.56
(GNN)	AUROC	82.76±0.50	83.20±0.76	91.22±0.56	78.37±0.63
DC(G	ECE	1.75±0.27	1.48±0.28	1.19±0.30	5.09±0.38
NN)-gt	AUROC	91.66±0.31	94.84±0.57	96.17±0.37	86.01±0.43

Table 2: Ablation study.

Conclusion

This paper provided a in-depth analysis of existing calibration models for GNNs. We conclude that the assessment of GNN outputs should consider two aspects: accuracy, reflecting the model's ability to distinguish between different classes, and AUROC, reflecting the model's ability to discriminate between correct and incorrect predictions. The calibration model should also align with such intentions. We observed that while calibration models conventionally emphasize accuracy, AUROC has generally been overlooked. This neglect poses a potential risk of compromising the original model's quality. Furthermore, our findings highlight that a model's discriminative ability is crucial for effective calibration, as demonstrated by our experimental results. This study makes three primary contributions: 1) We provide insights into the limitations of existing calibration methods, some of which, counterintuitive, degrade a model's discriminative ability. These analyses contribute to the advancement of calibration research. 2) We discover the necessity of incorporating AUROC as an essential metric in calibration models. 3) We introduce a straightforward discriminative ability-driven model, denoted as DC(GNN), which achieves more appropriate calibration outcomes with a significant increase in AUROC and a concurrent reduction in ECE.

Acknowledgments

This work was partially supported by the NSFC under Grants 92270125 and 62276024, as well as the National Key Research and Development Program of China No.2022YFC3302101.

References

DeGroot, M. H.; and Fienberg, S. E. 1983. The Comparison and Evaluation of Forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2): 12–22.

Foong, A. Y.; Li, Y.; Hernández-Lobato, J. M.; and Turner, R. E. 2019. 'In-Between'Uncertainty in Bayesian Neural Networks. *arXiv preprint arXiv:1906.11537*.

Gawlikowski, J.; Tassi, C. R. N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. 2021. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.

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.

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.

Izmailov, P.; Maddox, W. J.; Kirichenko, P.; Garipov, T.; Vetrov, D.; and Wilson, A. G. 2020. Subspace inference for Bayesian deep learning. In *Uncertainty in Artificial Intelligence*, 1169–1179. PMLR.

Jøsang, A. 2016. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer Publishing Company, Incorporated, 1st edition. ISBN 3319423355.

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

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

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.

Mehrtash, A.; Wells, W. M.; Tempany, C. M.; Abolmaesumi, P.; and Kapur, T. 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12): 3868–3878.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

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

Wang, J.; Zhang, S.; Xiao, Y.; and Song, R. 2022. A Review on Graph Neural Network Methods in Financial Applications. *Journal of Data Science*, 20(2): 111–134.

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, S.; Zeng, Z.; Yang, X.; and Zhang, X. 2023. Self-Supervised Graph Learning for Long-Tailed Cognitive Diagnosis. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 110–118. AAAI Press.

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.

Weng, X.; Wang, Y.; Man, Y.; and Kitani, K. M. 2020. Gnn3dmot: Graph neural network for 3d multi-object tracking with 2d-3d multi-feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6499–6508.

Yang, X.; Wang, J.; Zhao, X.; Li, S.; and Tao, Z. 2022. Calibrate Automated Graph Neural Network via Hyperparameter Uncertainty. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4640–4644.

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.

Zhu, F.; Cheng, Z.; Zhang, X.-Y.; and Liu, C.-L. 2022. Rethinking confidence calibration for failure prediction. In *European Conference on Computer Vision*, 518–536. Springer.