
Adaptive Cumulative Mass Calibration with Conformal Prediction

Daniil Kazantsev¹

Eric Moulines^{1,2}

Maxim Panov¹

Nikita Kotelevskii¹

Mohsen Guizani¹

¹Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates

²École pour l’informatique et les techniques avancées (EPITA), Paris, France

Abstract

Reliable probability estimates by classifiers are essential in high-risk applications. In practice, however, predicted probabilities are often miscalibrated, and many existing post-hoc calibration methods typically lack guarantees that a specific notion of calibration is achieved after the correction procedure is applied. We introduce a *set-based* perspective on calibration through the notion of *cumulative mass calibration* and the corresponding error measures. We propose a new calibration procedure based on conformal prediction that forms cumulative probabilities with guaranteed marginal coverage. We introduce an *adaptive* temperature scaling algorithm, with the temperature tuned for each input to satisfy the conformal coverage constraint. As we show, this procedure can be efficiently implemented. Across image classification tasks, particularly in settings with many classes, our method improves newly introduced calibration error measures (CMCE and α -CMCE) and standard metrics (such as ECE, cw-ECE, MCE) over the existing baselines.

The code is publicly available on GitHub  ¹.

1 INTRODUCTION

Classification models, particularly deep neural networks, are widely used in high-risk applications, including medical diagnosis and autonomous driving [Bojarski et al., 2016, Esteva et al., 2017]. In such settings, accurate probability estimates are crucial for decision-making. For example, when a model assigns a small probability to an outcome, a system might abstain from predicting or alert the user about the risk of error [Geifman and El-Yaniv, 2017, Hüllermeier and Waegeman, 2021, Franc et al., 2023]. However, modern

neural networks are notoriously miscalibrated, and their predicted probabilities are often unreliable [Guo et al., 2017, Kull et al., 2019, Zadrozny and Elkan, 2002].

The issue above has motivated extensive research on understanding and improving calibration. A key difficulty is that, given a finite dataset of one-hot labels, achieving and measuring *full multiclass (distributional) calibration* (see Definition 1) is practically challenging.

Therefore, in practice, more relaxed notions of calibration are used instead. For instance, Guo et al. [2017] showed that modern models tend to be overconfident and studied *confidence* calibration, using expected calibration error (ECE) as the metric. Kull et al. [2019] introduced *class-wise* calibration and the corresponding class-wise ECE (cw-ECE). Many other notions of calibration and associated metrics have been proposed [Nixon et al., 2019, Widmann et al., 2019, Zhao et al., 2021].

However, existing post-hoc calibration methods suffer from two significant problems. First, they typically do not directly target the associated calibration metric and thus provide no guarantee that the metric will improve [Vaicenavicius et al., 2019, Gupta et al., 2020]. For example, temperature scaling [Guo et al., 2017] does not guaranty of the improvement of ECE, and Dirichlet calibration [Kull et al., 2019] does not provably reduce cw-ECE. Second, methods based on one-vs-rest reductions (see, e.g., [Zadrozny and Elkan, 2002]) require fitting a separate calibrator per class, which becomes data-intensive as the number of classes grows [Le Coz et al., 2024]. Furthermore, existing notions of calibration have not been formulated in terms of cumulative sums of predicted probabilities. This is a consequential gap, as such cumulative quantities play an important and crucial role in decision-making [Zhao et al., 2021, Kiyani et al., 2025].

In this paper, we address these challenges and introduce new notions of calibration, corresponding error measures, and several practical post-hoc calibration procedures. Our key method is based on *conformal prediction* (CP; Shafer and Vovk, 2008, Angelopoulos et al., 2023). Specifically, we in-

¹https://github.com/stat-ml/conformal_probability_calibration

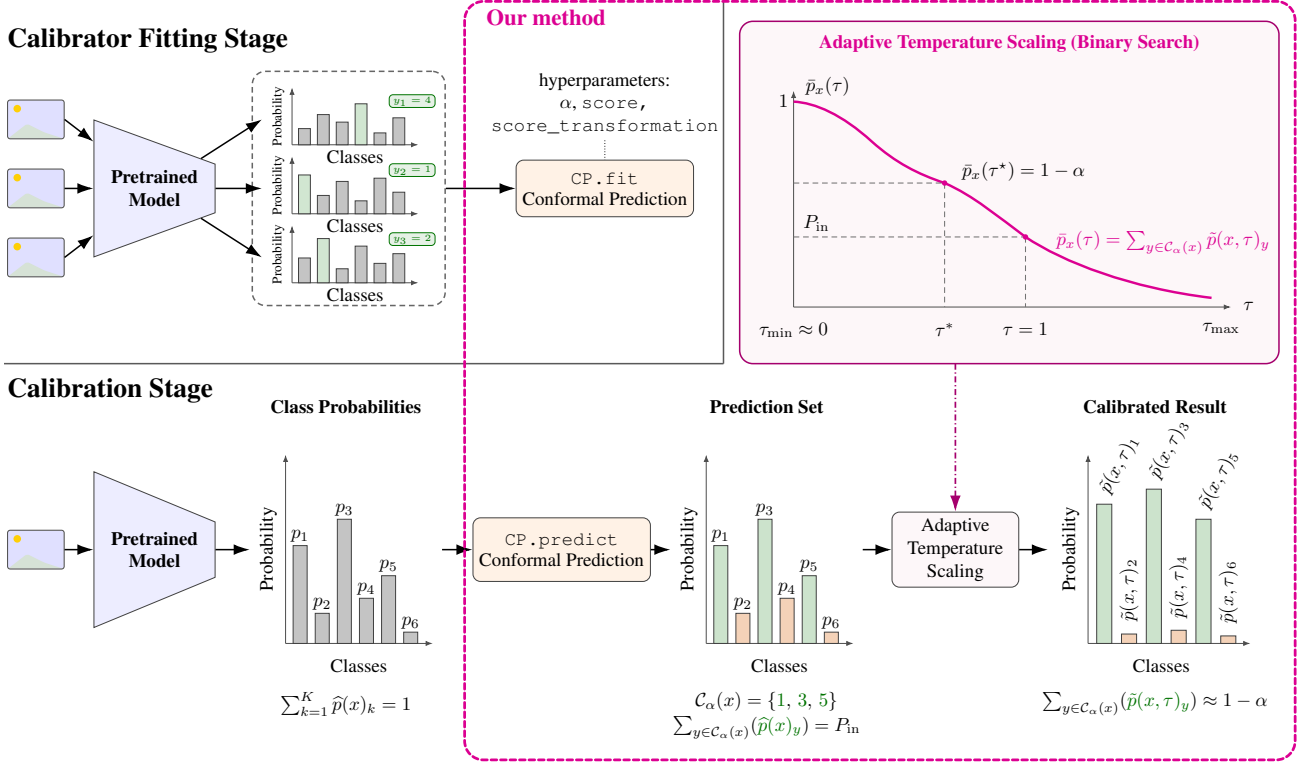


Figure 1: Flowchart of the proposed approach. In the first stage (calibrator fitting), a conformal predictor is fitted. In the second stage (calibration), the fitted conformal predictor is first applied to obtain the conformal set $C_\alpha(x)$. We then use binary search to determine τ^* , an approximate solution to an equation whose right-hand side is the constant $1 - \alpha$ and whose left-hand side is the total predicted probability mass of the elements in $C_\alpha(x)$, which is monotonically decreasing in τ .

Introduce a method that uses CP sets to calibrate probabilities via input *adaptive* temperature scaling. We show how to implement this procedure efficiently with one-dimensional binary search. Notably, the conformal-based calibration improves α -cumulative mass calibration (see Definition 4), and provides distribution-free marginal coverage guarantees; see equation (4). We also introduce an α -agnostic notion of calibration and its corresponding error measure, the *Cumulative Mass Calibration Error (CMCE)*, which considers calibration quality across all confidence levels.

Our main **contributions** are as follows.

1. We introduce new notions of calibration, defined for *cumulative masses*, that are necessary conditions for full multiclass (distributional) calibration; see Section 3.2.
2. We develop an efficient post-hoc calibration method that, based on conformal prediction sets, transforms predicted probabilities via *adaptive* temperature-scaling, improving α -mass calibration (see Definition 4) and corresponding error measures; see equation (4) and Section 4.
3. We propose a confidence level agnostic calibration measure, *Cumulative Mass Calibration Error (CMCE)*, and a method that directly optimizes it; see Section 3.3.

4. We demonstrate through large-scale experiments that our algorithms improve CMCE *and* other classical calibration errors, despite not being designed specifically for them; see Section 5.

2 RELATED WORK

Deep neural networks often provide *miscalibrated* probabilities, which pose a challenge to their reliable use in practice. For example, predicted confidences may not match empirical accuracies [Guo et al., 2017], class frequencies may be inaccurately estimated [Kull et al., 2019], and output can be unreliable for decision-making [Zhao et al., 2021], among other issues. These problems motivate different practical notions of calibration, algorithms, and evaluation metrics.

Two general strategies have been proposed for improving calibration: (i) training-time methods (e.g., label smoothing [Müller et al., 2019]), which require a specialized training procedure, and (ii) post-hoc calibration of an already pretrained model. While training-time methods can be effective, they modify the training process and often require retraining of the model from scratch. Consequently, post-hoc calibration methods are often preferred because they work directly with pretrained models.

The classical post-hoc calibration methods include Temperature Scaling [Guo et al., 2017], Platt Scaling [Platt et al., 1999], Isotonic Regression [Zadrozny and Elkan, 2002], and some parametric extensions such as Dirichlet Calibration [Kull et al., 2019] and Adaptive Temperature Scaling [Joy et al., 2023]. These methods are simple, widely used, and improve specific notions of calibration.

Some methods can be directly applied to a multiclass setting. For example, the classical temperature scaling fits a single global τ for all classes and covariates simultaneously. In contrast, some other approaches may work for multiclass classification indirectly by first fitting one-vs-rest binary classifiers and then aggregating them into a multiclass classifier. Therefore, when the number of classes K is large, these reductions become data-demanding [Le Coz et al., 2024] and can be impractical.

The classical temperature scaling, however, applies a single temperature to all inputs x , resulting in a non-adaptive procedure. To mitigate this issue, adaptive methods were proposed that learn input- or class-dependent temperatures [Tomani et al., 2022, Joy et al., 2023, Balanya et al., 2024]. These methods increase expressiveness but introduce additional models and hyperparameters, risk overfitting small calibration sets, and typically lack formal calibration guarantees.

Conformal prediction constructs *prediction sets* with guaranteed *marginal* coverage under the data exchangeability assumption [Shafer and Vovk, 2008, Angelopoulos et al., 2023, Plassier et al., 2024a,b]. While exact conditional guarantees are impossible without additional assumptions [Vovk, 2012, Foygel Barber et al., 2021], recent work shows that x -dependent but invertible score transforms can improve conditional behavior while retaining marginal validity [Colombo, 2024, Plassier et al., 2025]. However, standard CP returns sets rather than calibrated probability vectors. Recent studies examine the converse interaction, namely how calibration affects set efficiency [Huang et al., 2024, Xi et al., 2025], but not how conformal sets can be leveraged to produce calibrated probabilities (the direction we study in this paper).

3 CLASSIFIER CALIBRATION

We begin by introducing the necessary notations. We denote by $(X, Y) \sim P$ a random pair with a label $Y \in \mathcal{Y} = \{1, \dots, K\}$ and covariates $X \in \mathcal{X}$. A probabilistic classifier outputs a vector of predicted class probabilities $\hat{p}(x) = \{\hat{p}_1(x), \dots, \hat{p}_K(x)\} \in \Delta_K$, where $\Delta_K := \{q \in [0, 1]^K : \sum_{k=1}^K q_k = 1\}$ is the probability simplex. The ground-truth conditional distribution is denoted by $p^*(x)$. When there is no ambiguity, we write $\hat{p} := \hat{p}(x)$ and $p^* := p^*(x)$. The predicted label is $\arg \max_k \hat{p}_k(x)$, and the value $\max_k \hat{p}_k(x)$ is commonly referred to as the *confidence* in the calibration literature [Zhao et al., 2021].

3.1 BACKGROUND ON CALIBRATION

3.1.1 Notions of Calibration

Following Dwork et al. [2021], Zhao et al. [2021], we view calibration from the point of *indistinguishability*. A classifier $\hat{p}$ is considered calibrated with respect to some notion of calibration if a model $\hat{p}$ is statistically indistinguishable from the ground truth process p^* with respect to that notion of calibration.

Ideally, we would like $\hat{p}$ to coincide with p^* for each input x , in which case the model would be perfectly calibrated under any notion of calibration. In practice, however, this pointwise equality almost never holds. This has led to many relaxed notions of calibration in the literature. One such relaxation is *multiclass (distribution) calibration*.

Definition 1. A probabilistic classifier $\hat{p} : \mathcal{X} \rightarrow \Delta_K$ is said to be *multiclass (distribution) calibrated* if for every $k \in \mathcal{Y}$ the following equality holds:

$$P(Y = k \mid \hat{p}(X) = q) = q_k. \quad (1)$$

A multiclass-calibrated classifier can deviate significantly from p^* in a pointwise sense. A simple illustration arises for classifiers with low sharpness [Gruber and Buettner, 2022]. Moreover, multiclass calibration is difficult to verify empirically. For problems with many classes and limited data, checking whether (1) holds is not feasible, and it has motivated further relaxations with tractable error measures.

Arguably the most widely used calibration measures are *confidence calibration* [Guo et al., 2017] and *class-wise calibration* [Kull et al., 2019].

Definition 2. A probabilistic classifier $\hat{p} : \mathcal{X} \rightarrow \Delta_K$ is said to be *confidence calibrated* if for all $\beta \in [0, 1]$ the following equality holds:

$$P(Y = \arg \max_{k \in \mathcal{Y}} \hat{p}_k(X) \mid \max_{k \in \mathcal{Y}} \hat{p}_k(X) = \beta) = \beta. \quad (2)$$

Definition 3. A probabilistic classifier $\hat{p} : \mathcal{X} \rightarrow \Delta_K$ is said to be *class-wise calibrated* if for all $\beta \in [0, 1]$ and all $k \in \mathcal{Y}$ the following equality holds:

$$P(Y = k \mid \hat{p}_k(X) = \beta) = \beta. \quad (3)$$

Both (2) and (3) are *necessary* but not *sufficient* for multiclass calibration (1), as it is straightforward to verify.

Note that all these notions of calibration, by definition, are marginal in the sense that they are not required to hold for a particular x . Rather, they are defined on average for the class of indistinguishable inputs that satisfy the conditions in these definitions; see (2) and (3). E.g., for confidence calibration, we expect that (2) holds for all inputs x , for which $\max_{k \in \mathcal{Y}} \hat{p}_k(x) = \beta$ holds true.

3.1.2 Measures of Calibration Error

Unlike multiclass calibration, confidence and class-wise calibration have widely used practical error measures.

The **Expected Calibration Error** (ECE; Guo et al., 2017) evaluates how well a classifier is confidence calibrated:

$$\text{ECE} = \sum_{j=1}^b \frac{|B_j|}{N} |\text{Acc}(B_j) - \text{Conf}(B_j)|,$$

where N is the size of validation dataset, B_j is the j -th bin induced by predicted confidence, $\text{Acc}(B_j)$ is the accuracy in bin B_j , and $\text{Conf}(B_j)$ is the averaged predicted confidence in bin B_j .

The **class-wise Expected Calibration Error** (cw-ECE; Kull et al., 2019) evaluates class-wise calibration analogously.

$$\text{ECE}_{\text{cw}} = \sum_{k=1}^K \sum_{j=1}^b \frac{|B_j^{(k)}|}{NK} \cdot |\text{Acc}(B_j^{(k)}) - \text{Conf}(B_j^{(k)})|,$$

where each bin $B_j^{(k)}$ is formed separately for each class.

In literature, several post-hoc calibration algorithms have been proposed to reduce these errors. These include (non-adaptive) temperature scaling [Guo et al., 2017] for ECE and Dirichlet calibration [Kull et al., 2019] for cw-ECE. However, despite showing practical improvements in some cases, these algorithms do not, in general, reduce the corresponding error measure.

In what follows, we introduce a new set-based notion of calibration together with an algorithm that is guaranteed to reduce the corresponding calibration error and provides distribution-free marginal coverage guarantees.

3.2 CUMULATIVE MASS CALIBRATION

Practical motivation. We start this section with the motivation for the set-based calibration. Each notion of calibration helps different downstream decision-making rules [Zhao et al., 2021]. For example, confidence calibration aligns predicted confidence with empirical accuracy, leading to reliable selective prediction, abstention, or error detection. Class-wise calibration reduces gaps between predicted class probabilities and actual class frequencies, thereby lowering class-specific over- or underconfidence.

In many applications, however, systems operate with *prediction sets* rather than single labels [Sadinle et al., 2019, Romano et al., 2020, Angelopoulos et al., 2021]. For example, such systems might collect top-ranked labels until a target probability mass is reached. Based on the number of classes needed to get the target mass, the system decides whether to proceed. Then, it can proceed only if the resulting set of classes is sufficiently small. For decisions based

on these sets to be reliable, the cumulative mass of predicted probabilities must match the empirical coverage of the true labels. This motivates us to introduce new calibration notions, defined for cumulative masses.

Definitions. Let q be any categorical distribution over K classes and let π be a permutation of indices such that $q_{\pi(1)}, \dots, q_{\pi(K)}$ is nonincreasing. The *highest-probability region* of mass $1 - \alpha$ is:

$$\text{HPR}_{1-\alpha}[q] := \{\pi(1), \dots, \pi(m)\},$$

where $m := \min\{r : \sum_{k=1}^r q_{\pi(k)} \geq 1 - \alpha\}$.

Now, let us introduce two new notions of calibration.

Definition 4. A probabilistic classifier $\hat{p} : \mathcal{X} \rightarrow \Delta_K$ is said to be α -cumulative mass calibrated if for a given $\alpha \in [0, 1]$ the following inequality holds:

$$P(Y \in \text{HPR}_{1-\alpha}[\hat{p}(X)]) \geq 1 - \alpha.$$

Importantly, this condition holds for the ground-truth classifier p^* (see Appendix A). Equality need not hold, even for p^* , because the last included class may cause the cumulative mass to overshoot $1 - \alpha$.

Definition 5. A probabilistic classifier is said to be *cumulative mass calibrated* if it is α -cumulative mass calibrated for every $\alpha \in [0, 1]$.

Again, this condition holds for the ground-truth distribution p^* (see Appendix A).

Unlike notions that calibrate only a single class probability or a confidence score, which operate with scalars, cumulative mass calibration requires calibration of a *set of predicted probabilities*.

3.3 CUMULATIVE MASS CALIBRATION ERRORS (CMCE)

In this section, we introduce measures of calibration error that, analogously to ECE and cw-ECE, assess how well a classifier is calibrated with respect to the calibration notions defined above.

α -CMCE. We first define the *empirical coverage* of a model $\hat{p}$ on a validation set $\mathcal{D}_{\text{val}}$ of size N as the fraction of examples whose true label falls in the highest-probability region $\text{HPR}_{1-\alpha}(\hat{p})$:

$$\text{Coverage}(\hat{p}, \mathcal{D}_{\text{val}}) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(y_i \in \text{HPR}_{1-\alpha}[\hat{p}(x_i)]).$$

The α -CMCE is the deviation of coverage from the target coverage level $1 - \alpha$:

$$\alpha\text{-CMCE}(\hat{p}, \mathcal{D}_{\text{val}}) = |\text{Coverage}(\hat{p}, \mathcal{D}_{\text{val}}) - (1 - \alpha)|. \quad (4)$$

CMCE. Let us introduce CMCE, another measure of calibration error that does not require specifying α for computation. This measure is computed using binning, similarly to ECE and cw-ECE. For a validation set of size N , we form, for each example x , the K nested highest-probability sets induced by sorting the predicted probabilities. All such sets are collected into $\mathcal{S}$ (giving $|\mathcal{S}| = KN$), and the mass axis is partitioned into b bins with edges $0 = t_0 < t_1 < \dots < t_b = 1$. Each set $S \in \mathcal{S}$ is assigned to bin j according to its cumulative mass. The number of bins b is a hyperparameter, chosen as in ECE and cw-ECE.

Let $\mathcal{S}_j$ denote the collection of sets in bin j . Within each bin, we compute two quantities. The *empirical coverage* $\text{cov}(\mathcal{S}_j)$ is the fraction of sets in $\mathcal{S}_j$ that contain the true label y . The *average cumulative mass* $\text{mass}(\mathcal{S}_j)$ is the mean mass of sets in $\mathcal{S}_j$. The cumulative mass calibration error is

$$\text{CMCE} = \frac{1}{|\mathcal{S}|} \sum_{j=1}^b |\mathcal{S}_j| \cdot |\text{cov}(\mathcal{S}_j) - \text{mass}(\mathcal{S}_j)|, \quad (5)$$

where $|\mathcal{S}_j|$ is the number of sets in bin j . Since both coverage and mass lie in $[0, 1]$, we have $0 \leq \text{CMCE} \leq 1$, with smaller values indicating better calibration.

In the next section, we introduce a calibration algorithm that minimizes α -CMCE and provides α -cumulative mass calibrated classifier (see Definition 4). While we cannot provide analogous guarantees for CMCE, our experiments show that the algorithm also yields empirical improvements in CMCE.

4 MASS CALIBRATION ALGORITHMS

In this section, we introduce several algorithms that optimized the proposed variants of cumulative mass calibration.

4.1 DIRECT CMCE MINIMIZATION

We start with the most straightforward approach of direct CMCE optimization with *non-adaptive* temperature scaling. In its classical form [Guo et al., 2017], temperature scaling selects a single *global* temperature by maximizing predictive likelihood on a calibration set. For direct CMCE minimization, we instead select the temperature by minimizing CMCE (empirically) on a hold-out set. Since CMCE is not differentiable, we fix a predefined grid of temperature values and select the optimal one (see details in Appendix C).

The benefit of this algorithm, which we refer to as *Naive CMCE*, is that it optimizes CMCE directly, simultaneously for all levels α . However, it relies on a search over a non-differentiable objective and provides no guarantee that the resulting classifier satisfies the corresponding notion of calibration (see Definition 5). To obtain the guarantees, we construct an algorithm based on Conformal Prediction (CP)

to produce an α -cumulative-mass-calibrated classifier. CP provides a marginal coverage guarantee for the produced sets. Using the conformal set $\mathcal{C}_\alpha(X)$, we design a procedure that enforces the mass constraint at the chosen level α . We next review CP and then present the algorithm.

4.2 BACKGROUND ON CONFORMAL PREDICTION

Conformal prediction [Shafer and Vovk, 2008, Angelopoulos et al., 2023] is a framework that, under the exchangeability assumption, constructs a *set of labels* $\mathcal{C}_\alpha(X)$ with a *marginal* coverage guarantee:

$$P(Y \in \mathcal{C}_\alpha(X)) \geq 1 - \alpha, \quad (6)$$

for any specified error level $\alpha \in (0, 1)$.

To construct a CP label set (in a split conformal setup), one needs a pretrained model and a nonconformity score $s: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ that measures the error between a prediction and the correct target. The procedure requires a threshold q_α , selected from the scores $\{s(x_i, y_i)\}_{i=1}^n$ on a *calibration dataset*. The threshold is defined as the $\lceil (1 - \alpha)(n + 1) \rceil / n$ quantile of the empirical score distribution. Once the quantile is selected, the prediction set for an input x is formed as $\mathcal{C}_\alpha(x) := \{y: s(x, y) \leq q_\alpha\}$.

However, the coverage guarantee in (6) holds only *marginally* and says nothing about a set constructed for a specific input x . Ideally, one would like to have a *conditional* coverage guarantee $P(Y \in \mathcal{C}_\alpha(X) \mid X = x) \geq 1 - \alpha$ for any x and any α . However, this is impossible without additional distributional assumptions [Vovk, 2012, Foygel Barber et al., 2021]. Nevertheless, the following thoughtful procedure demonstrates the link between conditional conformal sets and perfect (point-wise) calibration.

Suppose that for every x and any confidence level α , we have valid conditional conformal sets $\{\mathcal{C}_\alpha(x)\}_{\alpha \in (0, 1)}$. Now let us fix an input x . Consider two boundary confidence levels $\alpha_1 \leq \alpha \leq \alpha_2$ such that the corresponding sets differ by exactly one label y . Specifically, this means $\mathcal{C}_{\alpha_1}(x) = \mathcal{C}_{\alpha_2}(x) \cup \{y\}$. Assume additionally that both α_1 and α_2 are minimal for the label sets they correspond to, and that the coverage is tight at these two levels. We can then recover the ground-truth probability of y as

$$\begin{aligned} p(Y = y \mid X = x) &= \\ P(Y \in \mathcal{C}_{\alpha_1}(X) \mid X = x) - P(Y \in \mathcal{C}_{\alpha_2}(X) \mid X = x) &= \\ &= \alpha_2 - \alpha_1. \end{aligned}$$

If we repeat this procedure for every label (by varying corresponding boundary levels α), we can recover the full conditional distribution $p(y \mid x)$ and reconstruct the ground-truth process for every x .

This preceding example demonstrates that valid (and exact at the boundaries) conditional sets can determine the full

conditional distribution in (1). Hence, conformal prediction can be used to improve calibration.

However, these valid conditional sets are unavailable, and in practice, we can only construct marginally valid sets $\mathcal{C}_\alpha(X)$ using conformal prediction. In the next section, we present a calibration algorithm based on this idea.

4.3 CUMULATIVE MASS CALIBRATION VIA CP

We propose a calibration algorithm, based on a marginal conformal prediction, with *adaptive* temperature scaling for each input x . The algorithm is *fundamentally different* from classical temperature scaling in [Guo et al., 2017], which fits a single global temperature shared across all inputs. Our method tunes an *input-dependent* temperature $\tau(x)$ which can be effectively found via one-dimensional binary search applied to a monotone function.

Below, we assume that both the pretrained model $\hat{p}$ and the procedure for constructing conformal sets are given. At inference time, using the uncalibrated model and its corresponding fitted CP procedure for specified level α , we produce a conformal set for input x . Then, we scale the model’s raw logits by the temperature $\tau(x) > 0$, which is tuned to satisfy a cumulative mass constraint for each input x . Once this temperature $\tau(x)$ is found, the classifier is calibrated in a sense that the cumulative mass over $\mathcal{C}_\alpha(X)$ matches (or, when not possible to match, exceeds) $1 - \alpha$. The detailed procedure is given in Algorithm 1.

In what follows, we omit explicit dependency on x for temperature: $\tau := \tau(x)$. The optimization problem for each input x is as follows:

$$\begin{aligned} \min_{\tau > 0} \quad & \sum_{y \in \mathcal{C}_\alpha(x)} \tilde{p}(x, \tau)_y, \\ \text{s.t.} \quad & \sum_{y \in \mathcal{C}_\alpha(x)} \tilde{p}(x, \tau)_y \geq 1 - \alpha, \end{aligned}$$

where $\tilde{p}(x, \tau) = \text{Softmax}[\log \hat{p}(x)/\tau]$.

The problem is one-dimensional for a scalar τ . Since the cumulative mass over $\mathcal{C}_\alpha(x)$ changes monotonically with τ , it can be solved via one-dimensional binary search.

4.4 CONSTRUCTION OF PREDICTIVE SETS

The proposed procedure linking valid conditional coverage to point-wise calibration (see Section 4.2) cannot be realized without additional assumptions [Vovk, 2012]. Therefore, in practice, for our *adaptive* temperature scaling algorithm, we use predictive sets from (split) CP, which provide only marginal (not conditional) coverage guarantees. Such marginally valid sets can be conservative, especially when per-class error rates are heterogeneous. As a result, a single global threshold may be too strict in some regions of $\mathcal{X}$.

Algorithm 1: CONFORMAL TEMPERATURE SCALING

Input: Calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$;

target error level $\alpha \in (0, 1)$;

probabilistic predictor $\hat{p}: \mathcal{X} \rightarrow \Delta^{K-1}$;

conformal prediction hyperparameters s, f_θ ;

temperature bounds $0 < \tau_{\min} < \tau_{\max}$.

Output: Calibrated probability vector $q(x) \in \Delta^{K-1}$.

Calibrator fitting stage:

Initialize conformal predictor CP with hyper-parameters: α, s, f_θ . And **fit** it on $\mathcal{D}_{\text{cal}}$, processed with $\hat{p}$.

Calibration stage:

1 **Conformal predictor call.** For a test input x , compute

$$\mathcal{C}_\alpha(x) \leftarrow \text{CP.predict}(\hat{p}(x)).$$

2 **if** $\mathcal{C}_\alpha(x) = \emptyset$ **or** $|\mathcal{C}_\alpha(x)| = K$ **then**

return $\hat{p}(x)$.

3 **Temperature-scaled probabilities.** For any $\tau > 0$, define

$$\bar{p}_x(\tau) = \sum_{y \in \mathcal{C}_\alpha(x)} \text{softmax}\left(\frac{\log \hat{p}(x)}{\tau}\right)_y = \sum_{y \in \mathcal{C}_\alpha(x)} \tilde{p}(x, \tau)_y.$$

4 **Temperature selection.** Find

$$\tau^* \in \arg \min_{\tau \in [\tau_{\min}, \tau_{\max}]} \bar{p}_x(\tau) \quad , \quad \text{s.t.} \quad \bar{p}_x(\tau) \geq 1 - \alpha,$$

using one-dimensional binary search over τ .

5 **return** $q = \tilde{p}(x, \tau)$.

Recent works propose adaptive *threshold* selection (do not confuse with adaptive temperature scaling) procedures that transform nonconformity scores in an x -dependent, invertible way while preserving marginal coverage [Colombo, 2024, Plassier et al., 2025]. These techniques also keep marginal coverage guarantees while often empirically improving conditional coverage. More precisely, let us define a transformation of the nonconformity score as:

$$\tilde{s}(x, y) = f_\theta(s(x, y), x), \tag{7}$$

where $f_\theta: \mathbb{R} \times \mathcal{X} \rightarrow \mathbb{R}$ is an invertible function parameterized by θ . We select a global threshold q_α for the transformed scores as in Section 4.2, and then define an x -adaptive threshold $\tilde{q}_\alpha(x) = f_\theta^{-1}(q_\alpha, x)$.

Accordingly, we consider two strategies for threshold selection in conformal prediction:

1. *Constant:* the classical global threshold q_α as in Section 4.2, with no additional training.
2. *Adaptive:* learn an invertible f_θ , select a global threshold based on the transformed scores, and adapt per x via $\tilde{q}_\alpha(x) = f_\theta^{-1}(q_\alpha, x)$. In our experiments, we use the method of Colombo [2024].

4.5 PROPERTIES OF CALIBRATED CLASSIFIERS

Our calibrated classifiers preserve the *marginal* coverage guarantee from conformal prediction at the chosen level α , since by construction $P(Y \in \mathcal{C}_\alpha(X)) \geq 1 - \alpha$. Given the conformal set $\mathcal{C}_\alpha(x)$, our *adaptive* temperature scaling keeps the class ordering unchanged (due to monotonicity of temperature scaling) and tunes a single scalar temperature for that input.

Hence, after the calibration procedure, the probability mass within a predicted conformal set meets $1 - \alpha$ when feasible, and otherwise exceeds it. In particular, $\text{HPR}_{1-\alpha}[(\hat{p}(x, \tau))] = \mathcal{C}_\alpha(x)$ and $P(Y \in \text{HPR}_{1-\alpha}[(\hat{p}(x, \tau))]) \geq 1 - \alpha$. Moreover, the accuracy also remains the same. The mentioned marginal coverage guarantees hold only for the selected α . Therefore, the adaptive temperature scaling algorithm ensures α -cumulative mass-calibration of a classifier (see Definition 4).

In contrast to direct CMCE fitting, adaptive temperature scaling does not imply calibration for other α values (although in practice there could be improvements; see Section 5). As the CMCE error measure summarizes calibration across all cumulative masses, we treat it as a *diagnostic*. While we do not guarantee CMCE improvement, we often observe empirical improvements when we enforce the conformal mass constraint.

5 EXPERIMENTS

In this section, we evaluate proposed calibration methods in classification problems. We first consider a synthetic experiment in which the ground-truth probabilities p^* are known. We then consider real-world image classification datasets with a varying number of classes. As we will see, a large number of classes pose a particularly challenging calibration problem for many calibration methods.

Empirically, we observe that once a classifier is α -cumulative-mass calibrated, it often also improves CMCE, ECE [Guo et al., 2017], cw-ECE [Kull et al., 2019], and other popular calibration metrics.

Baselines. We compare our methods with popular calibration methods: Isotonic Regression [Zadrozny and Elkan, 2002], Platt Scaling [Platt et al., 1999], Global Temperature Scaling [Guo et al., 2017], Adaptive Temperature Scaling (Ada-TS; Joy et al., 2023), Dirichlet Scaling [Kull et al., 2019], and Venn-Abers calibration [Vovk and Petej, 2014]. We describe the baselines in Appendix H.

Our method. We propose a calibrator, $\text{TS}_{\text{(ours)}}$, that refers to *conformal temperature scaling*, shown in the flow chart in Figure 1 and presented in Algorithm 1.

Since the algorithm is based on conformal prediction sets, one needs to choose a nonconformity score. We consider

two options: APS (Adaptive Predicted Sets; Romano et al., 2020), defined as $s_{\text{APS}}(x, y) = \sum_{j=1}^k \hat{p}_{\pi(j)}(x)$ where $y = \pi(k)$, and MSP (Maximum Softmax Probability), defined as $s_{\text{MSP}}(x, y) = 1 - \hat{p}_y(x)$.

Finally, we consider two options for transforming the threshold. One can use a standard (split) CP procedure with a single *global threshold* (see Section 4.2) or an optional transformation that allows instance-dependent thresholds, as given by equation (7).

In our notation, we denote these choices by `Const` (Constant), the case when no threshold transformation is applied, and by `Ada` (Adaptive), the case when a normalizing-flow transformation is used [Colombo, 2024].

5.1 SYNTHETIC EXPERIMENT

We evaluate calibration on a synthetic dataset where the ground-truth probabilities $p^*(y | x)$ are available, which allows us to compute values of Kullback-Leibler divergence (KL) exactly, using p^* . Note that KL is a proper calibration error (it is 0 iff $p^* = \hat{p}$ for all x) [Gruber and Buettner, 2022], and the smaller it is, the better calibration is achieved. A complete description of the data-generating process is presented in Appendix G. Figure 2 summarizes the main results.

The proposed algorithms (Naive CMCE and TS) demonstrate strong performance across different metrics and numbers of classes K . Moreover, our *adaptive* TS (instantiated with constant score transformation, MSP scores, and $\alpha = 0.01$) consistently achieves low KL divergence and substantially reduces CMCE, despite not being explicitly optimized for it. Our adaptive TS is the only calibration procedure that does not degrade α -CMCE with the increase of K .

The synthetic experiment clearly shows that our *adaptive* TS calibration method shows reasonable results for a moderate number of classes, while significantly improving performance when the number of classes is large. We observe the same behaviour in real-world image datasets.

5.2 IMAGE CLASSIFICATION DATASETS

In this section, we consider popular visual datasets and study the calibration of image classifiers. We consider datasets with different numbers of classes, such as CIFAR-10, CIFAR-100 [Krizhevsky, 2009], ImageNet [Deng et al., 2009] with 1,000 classes, and iNaturalist [Van Horn et al., 2021], a dataset with a large (10,000) number of classes. More detailed description of the datasets is given in Appendix B. For all datasets, we use pretrained, publicly available models trained in a standard manner (without training-time methods to improve calibration). We provide a detailed

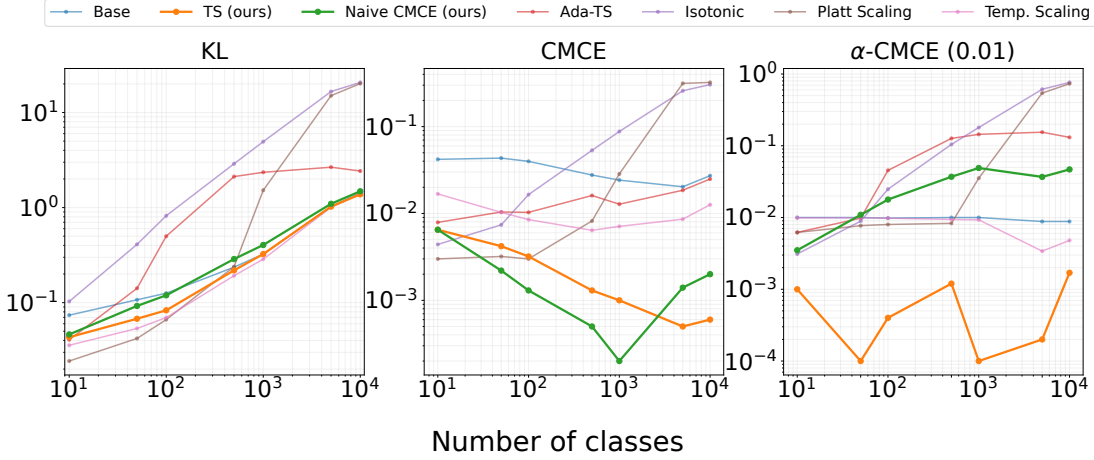


Figure 2: Performance metrics on synthetic data, as the number of classes K increases. TS tuned for $\alpha = 0.01$, MSP nonconformity score, and Constant transformation. KL computed with ground-truth $p^*(y | x)$. Our method (TS) consistently achieves better metrics as K increases.

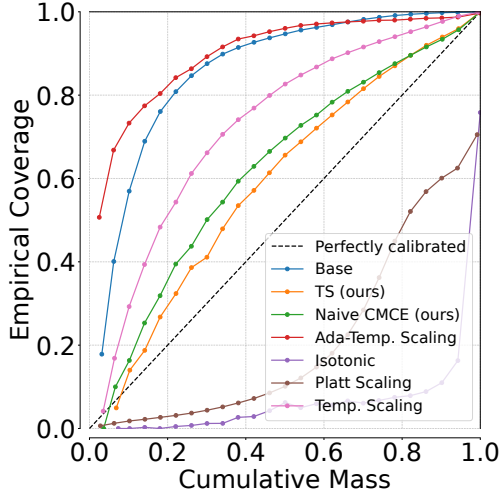


Figure 3: Cumulative mass calibration curves for different calibrators on iNaturalist21.

description of the models we use in Appendix D. Due to computational constraints, we do not evaluate some baselines on particular datasets. We specify which baselines are omitted for each dataset and provide details in E.

Experimental setup. Using available pretrained models for each dataset, we compute predictions (in the form of logits) for the corresponding validation data. Then, we split these predictions into two subsets: *calibration* (20%, to fit calibrators) and *test* (80%). We use 10 different splits to compute statistics for the metrics we consider. Once the calibrators are fitted to the corresponding calibration splits, we evaluate them over the test split and report the metrics.

Experimental results. Due to the space constraints, we move some of the results to Appendix I. Here, we keep the

most interesting ones. Since on some datasets (CIFAR-10, CIFAR-100, and ImageNet) all calibration methods keep the same accuracy as the base model, we do not report it in the tables. However, for iNaturalist, some baselines degrade accuracy, and we report this in a dedicated column in the table below.

CIFAR-10. It is well known [Kapoor et al., 2022] that standard image-classification datasets (e.g., CIFAR-10/100) contain almost no aleatoric uncertainty. Therefore, a model that predicts probability vectors close to one-hot vectors tends to be well-calibrated. Tables 2 and 3 in Appendix I reflect this: even uncalibrated models already perform strongly across most metrics. We observe that all calibration baselines and our proposed algorithms outperform the uncalibrated baselines. Among calibrators, Isotonic Regression and Temperature Scaling typically perform best. However, for α -CMCE, our algorithm with a constant transformation and $\alpha = 0.01$ gives, as expected, the best result. Nevertheless, the setting with few classes and minimal aleatoric uncertainty is not where our methods yield the most significant gain.

CIFAR-100. Table 1 (upper) reports results for CIFAR-100. In this and subsequent tables, entries that outperform the uncalibrated baseline are colored in green, while those that underperform are colored in orange. From Table 1 (upper), we see that our approaches consistently improve all metrics relative to the uncalibrated baseline, achieving the best and second-best results for the CMCE and α -CMCE. The test accuracy on CIFAR-100 is the same across all calibration methods and is 0.7264 ± 0.0023 for ResNet56 (the model we used).

iNaturalist21. Additionally, we consider the dataset with the largest number of classes (10,000). Table 1 (lower) reports different test metrics. From this table, we see that methods that aggregate binary calibrators into a multiclass

Table 1: **Upper table:** CIFAR-100 with ResNet-56; **Lower table:** iNaturalist21. Our calibrators use the MSP nonconformity score, a constant transformation, and either $\alpha = 0.05$ (upper) or $\alpha = 0.1$ (lower). Metrics* are scaled by 100. **Orange/Green** indicate worse/better performance than the uncalibrated model. **Bold** and underline mark the best and second-best values.

CIFAR-100						
Calibrator	ECE*	MCE*	NLL	Brier Score*	CMCE*	α -CMCE (0.05)*
Base	14.17 $\pm$ 0.10	32.36 $\pm$ 1.89	1.29 $\pm$ 0.01	41.78 $\pm$ 0.16	1.29 $\pm$ 0.01	8.09 $\pm$ 0.12
Ada-TS	15.04 $\pm$ 0.24	29.45 $\pm$ 2.93	2.37 $\pm$ 0.05	44.80 $\pm$ 0.26	1.95 $\pm$ 0.14	8.75 $\pm$ 0.25
Dirichlet	40.77 $\pm$ 1.01	73.60 $\pm$ 3.71	10.44 $\pm$ 0.55	82.06 $\pm$ 1.94	3.41 $\pm$ 0.17	35.76 $\pm$ 1.05
Isotonic	5.93 $\pm$ 0.29	32.52 $\pm$ 33.85	1.73 $\pm$ 0.07	39.28 $\pm$ 0.19	1.73 $\pm$ 0.16	2.27 $\pm$ 0.33
Platt Scaling	14.56 $\pm$ 0.23	21.73 $\pm$ 0.80	1.14 $\pm$ 0.01	41.72 $\pm$ 0.11	1.46 $\pm$ 0.06	3.42 $\pm$ 0.09
Temp. scaling	2.85 $\pm$ 0.15	11.41 $\pm$ 2.81	1.04 $\pm$ 0.01	38.24 $\pm$ 0.13	0.45 $\pm$ 0.03	1.42 $\pm$ 0.08
V.-Abers (OvA)	8.14 $\pm$ 0.44	16.67 $\pm$ 1.48	1.12 $\pm$ 0.01	39.36 $\pm$ 0.22	3.84 $\pm$ 0.15	4.44 $\pm$ 0.08
Naive CMCE (ours)	3.92 $\pm$ 0.42	11.17 $\pm$ 0.90	1.04 $\pm$ 0.01	38.33 $\pm$ 0.12	0.12 $\pm$ 0.08	0.57 $\pm$ 0.25
TS (ours)	4.26 $\pm$ 0.33	11.99 $\pm$ 0.94	1.04 $\pm$ 0.01	38.15 $\pm$ 0.16	0.35 $\pm$ 0.02	0.33 $\pm$ 0.21
iNaturalist21						
Calibrator	ECE*	MCE*	NLL	Brier Score*	CMCE*	α -CMCE (0.1)*
Base	49.73 $\pm$ 0.29	64.52 $\pm$ 0.82	2.04 $\pm$ 0.01	61.49 $\pm$ 0.26	5.55 $\pm$ 0.02	9.61 $\pm$ 0.04
Ada-TS	11.48 $\pm$ 0.53	26.97 $\pm$ 1.60	2.24 $\pm$ 0.10	35.98 $\pm$ 0.47	0.85 $\pm$ 0.12	3.03 $\pm$ 0.72
Isotonic	26.43 $\pm$ 1.27	38.47 $\pm$ 1.40	13.71 $\pm$ 0.12	80.19 $\pm$ 1.43	17.54 $\pm$ 0.60	40.01 $\pm$ 0.48
Platt Scaling	13.97 $\pm$ 0.45	29.66 $\pm$ 0.48	12.51 $\pm$ 0.09	68.02 $\pm$ 0.31	24.59 $\pm$ 0.27	32.98 $\pm$ 0.32
Temp. scaling	13.16 $\pm$ 8.35	25.60 $\pm$ 11.04	1.29 $\pm$ 0.10	35.70 $\pm$ 3.76	0.28 $\pm$ 0.35	4.64 $\pm$ 1.86
Naive CMCE (ours)	12.70 $\pm$ 3.73	25.73 $\pm$ 5.24	1.27 $\pm$ 0.03	35.09 $\pm$ 1.13	0.13 $\pm$ 0.01	4.75 $\pm$ 0.95
TS (ours)	7.45 $\pm$ 0.13	20.52 $\pm$ 8.27	1.38 $\pm$ 0.04	32.73 $\pm$ 0.45	0.08 $\pm$ 0.01	0.18 $\pm$ 0.53

calibrator become less effective and often worsen metrics. For instance, with Isotonic Regression, we observe a significant drop in accuracy and other metrics when the calibration set does not cover all labels. One way to bypass it is to require substantially more calibration data. However, it can be impractical, as even five calibration examples per class would mean 50,000 ($K = 10,000$) additional labeled images for post-hoc calibration. In contrast, our approaches are robust even when many classes are absent from the calibration set. The improvement is also evident in the cumulative mass calibration curves (see Figure 3), where our *adaptive* TS with MSP nonconformity score and global quantile for $\alpha = 0.1$ is nearly indistinguishable from the ideal curve, while also performing strongly on other metrics.

6 LIMITATIONS

There are several limitations of our approach.

1. There is a need to select a hyperparameter α to apply our proposed algorithms. While one may use the same α as for the plain CP procedure (there, it is dictated by application needs), in general, one may choose α for calibration based on other considerations. One future direction could be to develop a procedure that

automatically selects a specific α , or does not need it at all, while maintaining (marginal) CP guarantees.

2. Our adaptive TS calibration procedure uses classical *marginally* valid conformal sets. However, as shown in the thoughtful procedure in Section 3, only valid *conditional* coverage sets may lead to the full calibration.

7 CONCLUSION

In this paper, we introduced two notions of *cumulative mass calibration* that require calibration of cumulative probabilities, which is a necessary condition for full multiclass calibration. We defined α -CMCE and CMCE, practical empirical measures of cumulative mass calibration.

Building on split CP, we propose an *adaptive* TS calibration procedure that ensures α -cumulative mass calibration. The proposed method keeps a finite-sample, distribution-free *marginal* coverage guarantee at level $1 - \alpha$ from CP. The calibrator is especially effective in settings with a large number of classes, consistently reducing α -CMCE, CMCE, and other calibration metrics, despite not being explicitly optimized for it.

References

- Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.
- Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.
- Sergio A Balanya, Juan Maroñas, and Daniel Ramos. Adaptive temperature scaling for robust calibration of deep neural networks. *Neural Computing and Applications*, 36(14):8073–8095, 2024.
- Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.
- Nicolò Colombo. Normalizing flows for conformal regression. In *Uncertainty in Artificial Intelligence*, pages 881–893. PMLR, 2024.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.
- Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.
- Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.
- Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.
- Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Sebastian Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. *Advances in Neural Information Processing Systems*, 35: 8618–8632, 2022.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
- Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723, 2020.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20331–20347, 2024.
- Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3): 457–506, 2021.
- Tom Joy, Francesco Pinto, Ser-Nam Lim, Philip HS Torr, and Puneet K Dokania. Sample-dependent adaptive temperature scaling for improved calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14919–14926, 2023.
- Sanyam Kapoor, Wesley J Maddox, Pavel Izmailov, and Andrew G Wilson. On uncertainty, tempering, and data augmentation in bayesian classification. In *Advances in Neural Information Processing Systems*, volume 35, pages 18211–18225, 2022.
- Shayan Kiyani, George J Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. In *International Conference on Machine Learning*, pages 30943–30965. PMLR, 2025.
- Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, ON, Canada, 2009. CIFAR-10/100 dataset report.

- Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Adrien Le Coz, Stéphane Herbin, and Faouzi Adjed. Confidence calibration of classifiers with many classes. In *Advances in Neural Information Processing Systems*, volume 37, pages 77686–77725, 2024.
- Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in neural information processing systems*, volume 33, pages 7498–7512, 2020.
- Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.
- Vincent Plassier, Alexander Fishkov, Mohsen Guizani, Maxim Panov, and Eric Moulines. Probabilistic conformal prediction with approximate conditional validity. In *International Conference on Learning Representations*, 2024a.
- Vincent Plassier, Nikita Kotelevskii, Aleksandr Rubashevskii, Fedor Noskov, Maksim Velikanov, Alexander Fishkov, Samuel Horvath, Martin Takac, Eric Moulines, and Maxim Panov. Efficient conformal prediction under data heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 4879–4887. PMLR, 2024b.
- Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In *Forty-second International Conference on Machine Learning*, 2025.
- John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.
- Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591. Curran Associates, Inc., 2020.
- Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.
- Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.
- Christian Tomani, Daniel Cremers, and Florian Buettner. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In *European Conference on Computer Vision*, pages 555–569. Springer, 2022.
- Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd international conference on artificial intelligence and statistics*, pages 3459–3467. PMLR, 2019.
- Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisín Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12884–12893, 2021.
- Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pages 475–490. PMLR, 2012.
- Vladimir Vovk and Ivan Petej. Venn-Abers predictors. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 829–838, 2014.
- Vladimir Vovk, Ivan Petej, and Valentina Fedorova. Large-scale probabilistic predictors with and without guarantees of validity. *Advances in Neural Information Processing Systems*, 28, 2015.
- David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. *Advances in neural information processing systems*, 32, 2019.
- Huajun Xi, Jianguo Huang, Kangdao Liu, Lei Feng, and Hongxin Wei. Does confidence calibration improve conformal prediction? *Transactions on Machine Learning Research*, 2025.
- Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.
- Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions:

A novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems*, volume 34, pages 22313–22324, 2021.

Adaptive Cumulative Mass Calibration with Conformal Prediction (Supplementary Material)

Daniil Kazantsev¹ Eric Moulines^{1,2} Maxim Panov¹ Nikita Kotelevskii¹ Mohsen Guizani¹

¹Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates

²École pour l'informatique et les techniques avancées (EPITA), Paris, France

A AUXILIARY THEOREMS

Lemma A.1. *For any $\alpha \in [0, 1]$, the ground-truth classifier p^* is α -cumulative mass calibrated, meaning*

$$P(Y \in \text{HPR}_{1-\alpha}[p^*(X)]) \geq 1 - \alpha.$$

Proof. Fix any $x \in \mathcal{X}$ and let $q \in \Delta^K$ denote the conditional distribution $q_k = P(Y = k \mid X = x)$, so $q = p^*(x)$. Let $S = \text{HPR}_{1-\alpha}[q]$. By definition of $\text{HPR}_{1-\alpha}[q]$, the set S contains the m most likely labels under q , where m is the smallest integer such that:

$$\sum_{k=1}^m q_{\pi(k)} \geq 1 - \alpha.$$

Since $q_k = P(Y = k \mid X = x)$, we have:

$$P(Y \in S \mid X = x) = \sum_{k=1}^m P(Y = \pi(k) \mid X = x) = \sum_{k=1}^m q_{\pi(k)} \geq 1 - \alpha.$$

Now apply the law of total probability

$$P(Y \in \text{HPR}_{1-\alpha}[p^*(X)]) = \mathbb{E} \left[\mathbb{1}(Y \in \text{HPR}_{1-\alpha}[p^*(X)] \mid X) \right] \geq 1 - \alpha.$$

This proves the claim. □

Remark A.2. *Equality need not hold, even for p^* . By construction, $\text{HPR}_{1-\alpha}[q]$ is the smallest prefix of sorted probabilities whose cumulative sum reaches at least $1 - \alpha$. If no prefix sum equals exactly $1 - \alpha$, then the last included class makes the cumulative mass strictly larger than $1 - \alpha$.*

Lemma A.3. *The ground-truth classifier p^* is cumulative mass calibrated.*

Proof. By Definition 5, it suffices to show that p^* is α -cumulative mass calibrated for an arbitrary $\alpha \in [0, 1]$. Fix α . For any $x \in \mathcal{X}$, let $q = p^*(x)$ and let $S_\alpha(x) = \text{HPR}_{1-\alpha}[q]$. By construction of $\text{HPR}_{1-\alpha}[q]$, we have

$$\sum_{k \in S_\alpha(x)} q_k \geq 1 - \alpha.$$

Since $q_k = P(Y = k \mid X = x)$, it follows that

$$P(Y \in S_\alpha(x) \mid X = x) = \sum_{k \in S_\alpha(x)} P(Y = k \mid X = x) = \sum_{k \in S_\alpha(x)} q_k \geq 1 - \alpha.$$

Taking expectation over X yields

$$P(Y \in \text{HPR}_{1-\alpha}[p^*(X)]) = \mathbb{E} \left[\mathbb{1}(Y \in \text{HPR}_{1-\alpha}[p^*(X)] \mid X) \right] \geq 1 - \alpha.$$

Since α was arbitrary, p^* is α -cumulative mass calibrated for every $\alpha \in [0, 1]$, and therefore it is cumulative mass calibrated. $\square$

B DATASETS DESCRIPTION

We conduct experiments on four widely used image classification benchmarks:

- **CIFAR-10** [Krizhevsky, 2009]: Consists of natural images of size 32×32 pixels, evenly distributed across 10 classes.
- **CIFAR-100** [Krizhevsky, 2009]: Similar in structure to CIFAR-10, but with 100 classes.
- **ImageNet** [Deng et al., 2009]: ImageNet is a large-scale hierarchical dataset of millions of labeled real-world images across thousands of object categories, widely used as a benchmark for computer vision research. We use a subset of ImageNet, comprising 34,745 images and 1,000 classes.
- **iNaturalist 2021** [Van Horn et al., 2021]: The dataset consists of real-world photographs collected from the citizen-science platform iNaturalist. Images capture plants, animals, fungi, and other natural species in their natural environments. This dataset comprises 10,000 species.

C NAIVE CMCE ALGORITHM

In this section, we describe one of the algorithms proposed in this paper, the *Naive CMCE* procedure.

Cumulative mass calibration (Definition 5) is a necessary condition for full multiclass calibration. It is also stronger than an α -cumulative mass calibration, since the latter targets only a single level α . Our *adaptive* temperature scaling algorithm cannot be tuned for all α simultaneously. It relies on conformal prediction, which constructs predictive sets $\mathcal{C}_\alpha(x)$ for a fixed, user-chosen α .

This motivates a simple baseline that directly calibrates a base classifier by reducing empirical CMCE on a hold-out dataset. We refer to this baseline as *Naive CMCE*.

We use temperature scaling, but we tune it differently from both adaptive and classical temperature scaling [Guo et al., 2017]. Classical temperature scaling fits a single global temperature by maximizing predictive likelihood. In Naive CMCE, we fit a single global temperature by minimizing empirical CMCE.

Let $\hat{p}(x) \in \Delta^K$ denote the predictive distribution of the pretrained model. For $\tau > 0$, we define the temperature-scaled probabilities as

$$\hat{p}_\tau(x) = \text{Softmax}[\log \hat{p}(x)/\tau].$$

This transformation preserves the class ordering for each input x and adjusts only the confidence.

Given a hold-out calibration set $\mathcal{D}$, we select the temperature by

$$\hat{\tau} \in \arg \min_{\tau \in \mathcal{T}} \text{CMCE}(\hat{p}_\tau; \mathcal{D}),$$

where $\mathcal{T} \subset (0, \infty)$ is a predefined grid. We use a grid search since CMCE is not differentiable with respect to τ .

We construct $\mathcal{T}$ to be uniform in logarithmic space. We set $\tau_{\min} = 10^{-3}$ and $\tau_{\max} = 10^3$.

This baseline provides no distribution-free guarantees, including no coverage guarantees. Empirically, however, minimizing CMCE often improves other calibration metrics as well. But as our results show, minimizing CMCE usually improves other calibration metrics as well, supporting the importance of cumulative mass calibration.

D MODELS DESCRIPTION

Models. We employ pretrained models available through the HuggingFace library and the Birder project.

- **CIFAR-10.** We run experiments using ResNet-20 and ResNet-56 [He et al., 2016] architectures pretrained on the dataset.
- **CIFAR-100.** We use ResNet-56, as well as a Vision Transformer (ViT) [Dosovitskiy et al., 2021] with approximately 86M parameters, pretrained on ImageNet by Google.
- **ImageNet.** We evaluate a Vision Transformer (ViT) with approximately 86M parameters pretrained on ImageNet.
- **iNaturalist21.** We use a Vision Transformer (ViT) with approximately 91.6M parameters.

E COMPUTATIONAL CONSTRAINTS

Baselines omitted due to computational constraints. Due to computational constraints, we do not evaluate all baselines on every dataset. In particular, Dirichlet calibration scales poorly with the number of classes. In the official implementation we used, fitting materializes a per-sample outer product $XX^\top$ during reparameterization, resulting in a tensor of shape (n_{cal}, C, C) , where n_{cal} is the number of calibration samples and C is the number of classes. For ImageNet ($C=1000$) with $n_{\text{cal}}=6000$, this requires storing 6×10^9 floating-point values, and for iNaturalist the requirement exceeds 10^{12} values, which is not feasible in our experimental setup.

Venn–Abers with incomplete class coverage. We also omit Venn–Abers calibration when the calibration split does not include all classes. Venn–Abers is a binary calibrator and is commonly extended to the multiclass case via a one-vs-all reduction. This reduction requires, for each class, at least one positive example in the calibration data. On iNaturalist, the calibration split does not cover all classes; hence, some one-vs-all subproblems have no positive instances and the corresponding calibrations are undefined. As a result, the method cannot be applied reliably in this setting.

F EXPERIMENTAL SETUP DETAILS

Pipeline. Our evaluation pipeline is organized as follows:

1. We compute predictions (logits) with pretrained models for the datasets discussed above.
2. We split the data into two subsets: *calibration* and *test*, according to the declared proportion. We use 10 splits for computing metric distributions, and we allocate 20% of the data for the calibration set.
3. We train the calibrators on the calibration data. Hence, the methods receive logits and true classes.
4. We compute the calibrators’ prediction for the test part of the data and evaluate metrics.

Visualization.

- **Result tables.** The tables report the mean and variance of each evaluation metric across different experiments. Values highlighted in **green** indicate improvements over the uncalibrated model, while values in **orange** denote degradations. The best-performing results for each metric (except coverage) are emphasized with **bold and underlined** formatting.
- **Our methods arguments comparison table.** The table compares some metrics (usually CMCE) for different setups. Underlined are these metrics, which show the best performance across columns (usually, it corresponds with the best choice of α). With **bold** font are highlighted metrics, which outperform across rows (usually, it corresponds with the best choice of our method and score type for a fixed α)
- **Cumulative mass calibration curve.** The graph curve illustrates the relationship between the cumulative predicted probability mass of top-ranked labels and the observed coverage of the truth class along one bin. A perfectly calibrated model corresponds to the diagonal line, while deviations from this line indicate over- or under-confidence. We use aggregation over 25 bins.

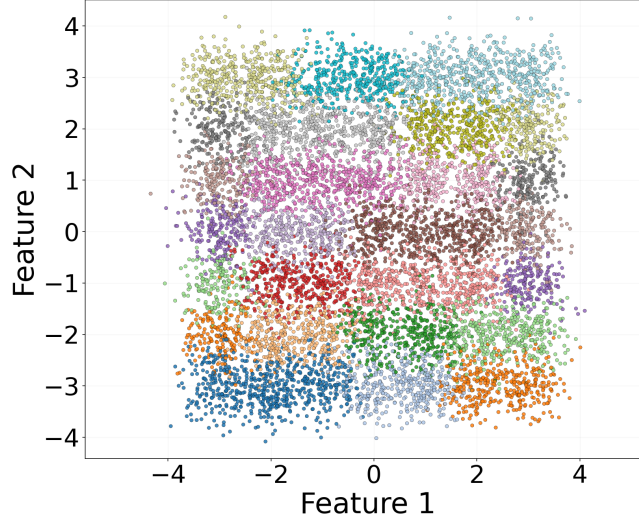


Figure 4: Samples from the synthetic dataset with $K = 49$; points are colored by class.

Metrics. We evaluate classifiers using the following metrics:

- **Brier Score** is the mean squared error between predicted probabilities and one-hot true labels:

$$\text{BS} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K (\mathbb{1}\{y_i = k\} - \hat{p}_k(x_i))^2.$$

- **NLL** (Negative Log-Likelihood) is the average log loss of the probability assigned to the true label:

$$\text{NLL} = -\frac{1}{n} \sum_{i=1}^n \log \hat{p}_{y_i}(x_i).$$

- **ECE** (Expected Calibration Error) [Guo et al., 2017] is the average absolute calibration gap across b confidence bins:

$$\text{ECE} = \sum_{i=1}^b \frac{|B_i|}{n} |\text{Acc}(B_i) - \text{Conf}(B_i)|,$$

where B_i is the i -th bin induced by predicted confidence, $\text{Acc}(B_i)$ is the empirical accuracy in B_i , and $\text{Conf}(B_i)$ is the mean predicted confidence in B_i .

- **cw-ECE** (Class-wise Expected Calibration Error) [Kull et al., 2019] averages per-class calibration gaps over b bins:

$$\text{ECE}_{\text{cw}} = \sum_{j=1}^K \sum_{i=1}^b \frac{|B_i^{(j)}|}{nK} \cdot |\text{Acc}(B_i^{(j)}) - \text{Conf}(B_i^{(j)})|,$$

where each bin $B_i^{(j)}$ is obtained separately for each class.

G SYNTHETIC EXPERIMENT DETAILS

Data. In this synthetic experiment, we generate a 2D dataset with K classes, where K can be up to 10,000. See example of a synthetic dataset in Figure 4

Each class k is represented by an isotropic Gaussian, with shared variance $\sigma = 0.35$ and centers, uniformly placed in a square grid. We draw labels uniformly, and sample inputs from an isotropic Gaussian around their class center:

$$p(x \mid y = k) = \mathcal{N}(\mu_k, \sigma^2 I).$$

Given any x , the ground-truth class probabilities form a normalized Gaussian mixture with equal priors:

$$p(y = k \mid x) \propto \exp\left(-\frac{1}{2\sigma^2}\|x - \mu_k\|_2^2\right).$$

This construction is simple, gives high accuracy when σ is small, and still contains aleatoric uncertainty due to overlap between nearby classes.

Model, loss, and optimization. We train a small, fully connected classifier network that maps $x \in \mathbb{R}^2$ to K logits. To help with many-class separation on grids, we add random Fourier features (RFF; Liu et al., 2020) before the MLP:

$$\phi(x) = [x, \sin(2\pi B^\top x), \cos(2\pi B^\top x)],$$

where $B \in \mathbb{R}^{2 \times F}$ has i.i.d. Gaussian entries.

We train the classifier with cross-entropy loss and AdamW optimizer.

H BASELINES DESCRIPTION

In our experiments, we compare several post-hoc calibration methods. When a method is binary by design, we apply it to each class in a one-vs-rest (OvR) way and then renormalize:

$$q_k(x) = g_k(z_k(x)) \Rightarrow \tilde{p}_k(x) = \frac{q_k(x)}{\sum_{j=1}^K q_j(x)}, \quad (8)$$

where $z_k = z_k(x)$ is the score (or logit) for class k and g_k is the binary calibrator.

Platt scaling [Platt et al., 1999] (binary, we use OvR for multiclass). Fits a sigmoid to map scores to probabilities:

$$\hat{p}(y_k = 1 \mid z_k) = \sigma(Az_k + B) = \frac{1}{1 + e^{-(Az_k + B)}}.$$

Then, we renormalize these binary probabilities as in (8) to obtain categorical probabilities.

Temperature scaling [Guo et al., 2017] (multiclass). Divides all logits by a single positive scalar $T > 0$ before softmax:

$$\hat{p}_i = \text{Softmax}(z/T)_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}.$$

This parameter is tuned via the maximization (gradient ascent) of the likelihood over the calibration split of the data.

Isotonic regression [Zadrozny and Elkan, 2002] (binary, we use OvR for multiclass). Learns a monotone, piecewise-constant map g from score to probability with no parametric form:

$$\hat{p}(y_k = 1 \mid z_k) = g(z_k),$$

with g non-decreasing (which is achieved by the Pool Adjacent Violators Algorithm (PAVA)). Again, for categorical probabilities, we renormalize these binary probabilities as in (8).

Dirichlet calibration [Kull et al., 2019] (multiclass). Learns a parametric mapping on the simplex using features of the uncalibrated scores, implemented as a multinomial logistic regression:

$$\hat{p} = \text{Softmax}(W\phi(z) + b),$$

where $\phi(z)$ are fixed transformations of scores (e.g., logits or log-softmax and simple interactions).

Venn-Abers [Vovk and Petej, 2014, Vovk et al., 2015] (binary; OvR for multiclass). Venn-Abers turns a scoring binary classifier into a (multi-)probabilistic predictor with calibration guarantees: for a test input x , it considers the two hypothetical labels $y \in \{0, 1\}$, assigns (x, y) to a Venn category induced by the score, and uses isotonic regression to produce a probability pair (p_0, p_1) that act as a lower and upper estimate of true class probability.

We report a single probability via the standard Venn-Abers aggregation; for multiclass classification, we use the one-vs-rest approach. We follow the IVAP implementation from [Vovk et al., 2015] (Algorithms 5-6) and utilize the publicly available code¹. In our experiments, we will emphasize that we use OvA (one-vs-all), which corresponds to the OvR strategy.

¹<https://github.com/ip200/venn-acters/tree/main>

Sample-dependent adaptive temperature [Joy et al., 2023]. The idea is to predict temperature values on a per-data-point basis to address the limitation of a single temperature parameter, which could reduce the individual confidence of the predictions, irrespective of whether the classification of a given input is correct or incorrect. For this, they fit a separate temperature prediction module (a small neural network) that operates independently on each input sample.

The calibrated probability is then computed as follows:

$$\hat{p}_i(x) = \text{Softmax}(z/T(x))_i = \frac{e^{z_i/T(x)}}{\sum_j e^{z_j/T(x)}},$$

where z also depends on x , but we omitted it for clarity.

Optimization and splits. All calibrators are fit on a held-out calibration split only (test split is never used for fitting).

I ADDITIONAL EXPERIMENTAL RESULTS

In this section, we present additional experimental results on the considered datasets. In the tables we present below, we will use the following shortcuts to display the strategy for threshold selection (see Section 4.4 in the main part): `C` stands for *Constant* (identical transformation of the conformal threshold), and `Ada` stands for *Adaptive*, as a particular way to incorporate dependency on x via a normalizing flow [Colombo, 2024].

I.1 CIFAR-10

We consider a classification problem over CIFAR-10. We present complete results in Table 4 (for ResNet20) and in Table 5 (for ResNet56). Additionally, we present the best instances of our algorithms in Table 2 and Table 3 in a similar manner. We see that for such a small dataset, our algorithm does not provide a significant gain. Nonetheless, our methods consistently outperform the uncalibrated model across all metrics. In particular, Naive CMCE achieves the lowest CMCE while also improving NLL and Brier score compared to the base model. Temperature Scaling and Isotonic Regression remain strong baselines.

The accuracy for ResNet20 and ResNet56 are same for all calibration methods, and are (0.9256 ± 0.0011) and (0.9439 ± 0.001) correspondingly.

Table 2: Evaluation results on CIFAR-10 and ResNet-20 (MSP score, Constant transformation, $\alpha = 0.01$ for ours). **Orange/Green** indicate worse/better performance than the uncalibrated model. **Bold** and underline mark the best and second-best values.

Calibrator	ECE	MCE	cw-ECE	NLL	Brier Score	CMCE	α -CMCE (0.01)
Base	.04 ± .00	.32 ± .21	.01 ± .00	.282 ± .005	.12 ± .00	.007 ± .000	.007 ± .001
Ada-TS	.03 ± .00	.35 ± .23	.01 ± .00	.328 ± .047	.12 ± .00	.004 ± .001	.004 ± .004
Dirichlet	.01 ± .00	.45 ± .31	.01 ± .00	.242 ± .007	.12 ± .00	.002 ± .001	.005 ± .001
Isotonic	.02 ± .00	.34 ± .25	.01 ± .00	.299 ± .030	.11 ± .00	.003 ± .001	.004 ± .001
Platt Scaling	.04 ± .00	.29 ± .05	.01 ± .00	.256 ± .002	.12 ± .00	.006 ± .001	.008 ± .000
Temp. scaling	.02 ± .01	.35 ± .24	.01 ± .00	.242 ± .013	.11 ± .00	.003 ± .002	.003 ± .004
V.-Abers (OvA)	.02 ± .00	.20 ± .04	.01 ± .00	.246 ± .003	.11 ± .00	.010 ± .000	.009 ± .001
Naive CMCE (ours)	.01 ± .00	.17 ± .06	.01 ± .00	.235 ± .004	.11 ± .00	.002 ± .000	.006 ± .001
TS (ours)	.02 ± .00	.27 ± .13	.01 ± .00	.241 ± .004	.11 ± .00	.003 ± .000	.001 ± .001

I.2 CIFAR-100

Here, we consider classification over CIFAR-100. For this dataset, we consider ResNet-56 and Visual Transformer (ViT). Similarly to CIFAR-10, here we present full results in Table 7 (ViT) and Table 6 (ResNet56); in Tables 8 and 1 (main text) we present the best results.

Table 3: Evaluation results on CIFAR-10 and ResNet-56 (MSP score, Constant transformation, $\alpha = 0.01$ for ours). Metrics* are scaled by 100. **Orange/Green** indicate worse/better performance than the uncalibrated model. **Bold** and underline mark the best and second-best values.

Calibrator	ECE	MCE	cw-ECE	NLL	Brier Score	CMCE	α -CMCE (0.01)
Base	.04 $\pm$.00	.41 $\pm$.03	.01 $\pm$.00	.254 $\pm$.006	.09 $\pm$.00	.006 $\pm$.000	.010 $\pm$.001
Ada-TS	.02 $\pm$.00	.26 $\pm$.03	.01 $\pm$.00	.241 $\pm$.018	.09 $\pm$.00	<u>.003 $\pm$.001</u>	.000 $\pm$.002
Dirichlet	.02 $\pm$.00	.40 $\pm$.20	<u>.01 $\pm$.00</u>	.202 $\pm$.007	.09 $\pm$.00	.003 $\pm$.001	.004 $\pm$.001
Isotonic	<u>.01 $\pm$.00</u>	.22 $\pm$.02	.01 $\pm$.00	.265 $\pm$.015	<u>.09 $\pm$.00</u>	.003 $\pm$.000	.003 $\pm$.000
Platt Scaling	.02 $\pm$.00	.27 $\pm$.11	.01 $\pm$.00	<u>.194 $\pm$.004</u>	.09 $\pm$.00	.004 $\pm$.001	.006 $\pm$.001
Temp. scaling	.01 $\pm$.00	.25 $\pm$.03	.01 $\pm$.00	.193 $\pm$.004	.09 $\pm$.00	.003 $\pm$.000	.006 $\pm$.001
V.-Abers (OvA)	.02 $\pm$.00	<u>.22 $\pm$.04</u>	.01 $\pm$.00	.200 $\pm$.002	.09 $\pm$.00	.009 $\pm$.001	.008 $\pm$.000
Naive CMCE (ours)	.02 $\pm$.00	.28 $\pm$.04	.01 $\pm$.00	.196 $\pm$.005	.09 $\pm$.00	.002 $\pm$.000	.004 $\pm$.001
TS (ours)	.02 $\pm$.00	.34 $\pm$.05	.01 $\pm$.00	.200 $\pm$.005	.09 $\pm$.00	.003 $\pm$.000	<u>.001 $\pm$.001</u>

Table 4: Evaluation results on CIFAR-10 and ResNet-20. **Orange/Green** indicate worse/better performance than the uncalibrated model. **Bold** and underline mark the best and second-best values.

Calibrator	α	Score Transf.	ECE	MCE	cw-ECE	NLL	Brier Score	CMCE	α -CMCE (0.2)	α -CMCE (0.1)	α -CMCE (0.05)	α -CMCE (0.01)
Base			.04 $\pm$.00	.32 $\pm$.21	.01 $\pm$.00	.282 $\pm$.005	.12 $\pm$.00	.007 $\pm$.000	.149 $\pm$.001	.059 $\pm$.001	.019 $\pm$.001	.007 $\pm$.001
Ada-TS			.03 $\pm$.00	.35 $\pm$.23	.01 $\pm$.00	.328 $\pm$.047	.12 $\pm$.00	<u>.004 $\pm$.001</u>	.156 $\pm$.004	.068 $\pm$.004	.026 $\pm$.005	.004 $\pm$.004
Dirichlet			.01 $\pm$.00	.45 $\pm$.31	.01 $\pm$.00	.242 $\pm$.007	.12 $\pm$.00	<u>.002 $\pm$.001</u>	.162 $\pm$.003	.076 $\pm$.001	.035 $\pm$.001	.005 $\pm$.001
Isotonic			.02 $\pm$.00	.34 $\pm$.25	<u>.01 $\pm$.00</u>	.299 $\pm$.030	.11 $\pm$.00	.003 $\pm$.001	.165 $\pm$.001	.077 $\pm$.002	.034 $\pm$.002	.004 $\pm$.001
Platt Scaling			.04 $\pm$.00	.29 $\pm$.05	.01 $\pm$.00	.256 $\pm$.002	.12 $\pm$.00	.006 $\pm$.001	.169 $\pm$.001	.084 $\pm$.001	.040 $\pm$.001	.008 $\pm$.000
Temp. scaling			.02 $\pm$.01	.35 $\pm$.24	.01 $\pm$.00	.242 $\pm$.013	.11 $\pm$.00	.003 $\pm$.002	.159 $\pm$.005	.072 $\pm$.006	.032 $\pm$.005	.003 $\pm$.004
V.-Abers (OvA)			.02 $\pm$.00	<u>.20 $\pm$.04</u>	.01 $\pm$.00	.246 $\pm$.003	.11 $\pm$.00	<u>.010 $\pm$.000</u>	.173 $\pm$.001	.085 $\pm$.001	.043 $\pm$.000	.009 $\pm$.001
Naive CMCE			.01 $\pm$.00	.17 $\pm$.06	.01 $\pm$.00	.235 $\pm$.004	.11 $\pm$.00	.002 $\pm$.000	.164 $\pm$.001	.077 $\pm$.002	.036 $\pm$.001	.006 $\pm$.001
TS (ours)	0.01	APS Const	.16 $\pm$.00	.24 $\pm$.02	.03 $\pm$.00	.375 $\pm$.003	.14 $\pm$.00	.063 $\pm$.001	.191 $\pm$.001	.098 $\pm$.000	.049 $\pm$.000	.010 $\pm$.000
TS (ours)	0.01	APS Ada	.04 $\pm$.00	.52 $\pm$.24	.01 $\pm$.00	.306 $\pm$.012	.12 $\pm$.00	.007 $\pm$.000	.146 $\pm$.002	.055 $\pm$.002	.015 $\pm$.003	.012 $\pm$.002
TS (ours)	0.01	MSP Const	.02 $\pm$.00	.27 $\pm$.13	.01 $\pm$.00	<u>.241 $\pm$.004</u>	<u>.11 $\pm$.00</u>	.003 $\pm$.000	.160 $\pm$.001	.074 $\pm$.001	.034 $\pm$.001	<u>.001 $\pm$.001</u>
TS (ours)	0.01	MSP Ada	.03 $\pm$.00	.42 $\pm$.27	.01 $\pm$.00	.272 $\pm$.013	.12 $\pm$.00	.004 $\pm$.001	.150 $\pm$.001	.061 $\pm$.003	.021 $\pm$.004	.008 $\pm$.004
TS (ours)	0.05	APS Const	.40 $\pm$.01	.62 $\pm$.00	.08 $\pm$.00	.801 $\pm$.010	.34 $\pm$.00	.186 $\pm$.003	.199 $\pm$.000	.100 $\pm$.000	.050 $\pm$.000	.010 $\pm$.000
TS (ours)	0.05	APS Ada	.05 $\pm$.00	.58 $\pm$.25	.01 $\pm$.00	.325 $\pm$.013	.13 $\pm$.00	.008 $\pm$.000	.139 $\pm$.002	<u>.046 $\pm$.002</u>	<u>.004 $\pm$.002</u>	.010 $\pm$.001
TS (ours)	0.05	MSP Const	<u>.01 $\pm$.00</u>	.55 $\pm$.19	.01 $\pm$.00	.301 $\pm$.010	.12 $\pm$.00	.007 $\pm$.001	.148 $\pm$.003	.050 $\pm$.004	.000 $\pm$.004	.002 $\pm$.001
TS (ours)	0.05	MSP Ada	.03 $\pm$.01	.45 $\pm$.26	.01 $\pm$.00	.301 $\pm$.015	.12 $\pm$.00	.009 $\pm$.001	.146 $\pm$.003	.053 $\pm$.004	.010 $\pm$.005	.002 $\pm$.001
TS (ours)	0.1	APS Const	.48 $\pm$.01	.77 $\pm$.00	.03 $\pm$.00	1.252 $\pm$.028	.48 $\pm$.01	.254 $\pm$.005	.199 $\pm$.000	.100 $\pm$.000	.050 $\pm$.000	.010 $\pm$.000
TS (ours)	0.1	APS Ada	.04 $\pm$.00	.76 $\pm$.19	.01 $\pm$.00	.328 $\pm$.013	.13 $\pm$.00	.007 $\pm$.000	<u>.139 $\pm$.002</u>	.044 $\pm$.002	.021 $\pm$.002	.008 $\pm$.001
TS (ours)	0.1	MSP Const	.05 $\pm$.00	.32 $\pm$.22	.01 $\pm$.00	.317 $\pm$.003	.13 $\pm$.00	.023 $\pm$.000	.145 $\pm$.001	.048 $\pm$.002	.037 $\pm$.001	.005 $\pm$.001
TS (ours)	0.1	MSP Ada	.07 $\pm$.00	.43 $\pm$.25	.02 $\pm$.00	.326 $\pm$.005	.13 $\pm$.00	.023 $\pm$.000	.146 $\pm$.002	.053 $\pm$.002	.031 $\pm$.002	.003 $\pm$.001
TS (ours)	0.2	APS Const	.25 $\pm$.00	.83 $\pm$.01	.03 $\pm$.00	.671 $\pm$.007	.28 $\pm$.00	.108 $\pm$.001	.159 $\pm$.001	.070 $\pm$.001	.029 $\pm$.001	.000 $\pm$.001
TS (ours)	0.2	APS Ada	.08 $\pm$.10	.64 $\pm$.26	.01 $\pm$.01	.381 $\pm$.172	.15 $\pm$.07	.026 $\pm$.046	.141 $\pm$.008	.067 $\pm$.003	.024 $\pm$.003	.005 $\pm$.003
TS (ours)	0.2	MSP Const	.17 $\pm$.00	.32 $\pm$.21	.03 $\pm$.00	.406 $\pm$.005	.15 $\pm$.00	.055 $\pm$.001	.149 $\pm$.001	.071 $\pm$.001	.032 $\pm$.001	.004 $\pm$.001
TS (ours)	0.2	MSP Ada	.16 $\pm$.00	.34 $\pm$.20	.03 $\pm$.00	.401 $\pm$.002	.15 $\pm$.00	.054 $\pm$.001	<u>.146 $\pm$.003</u>	.074 $\pm$.001	.033 $\pm$.001	.003 $\pm$.001

We observe two patterns in our results: (i) performance is model-dependent, and calibrators tend to help ResNet-56 more than ViT; and (ii) Temperature Scaling often provides the best ECE/NLL among classical baselines, while our Naive CMCE yields the best (or near-best) CMCE and the tightest control of coverage near the target.

The accuracy for ResNet56 architecture are same (0.7264 ± 0.0023) and for ViT (0.8947 ± 0.0011).

I.3 IMAGENET

In Table 10, we present the full results for ImageNet (ViT), and in Table 9, we display the best ones. From the results, we see that most of the methods do not the uncalibrated baseline across all metrics except cumulative mass ones. Our methods show better performance compared with the baselines calibrations.

Table 10 indicates that standard baselines frequently degrade calibration on this many-class task (e.g., Venn-Abers and Platt increase ECE/MCE substantially). The uncalibrated model achieves a strong NLL/Brier score.

Still, our methods substantially reduce CMCE: e.g., TS (MSP, Const, $\alpha = 0.01$) achieves the second-best CMCE among calibrated models and improves ECE/MCE relative to most competitors, while maintaining competitive coverage near both targets. Moreover, all of our methods improve α -CMCE for α 's used during the calibration procedure.

Table 5: Evaluation results on CIFAR-10 and ResNet-56. **Orange/Green** indicate worse/better performance than the uncalibrated model. **Bold** and underline mark the best and second-best values.

Calibrator	α	Score Transf.	ECE	MCE	cw-ECE	NLL	Brier Score	CMCE	α -CMCE (0.2)	α -CMCE (0.1)	α -CMCE (0.05)	α -CMCE (0.01)
Base			.04 $\pm$.00	.41 $\pm$.03	.01 $\pm$.00	.254 $\pm$.006	.09 $\pm$.00	.006 $\pm$.000	.157 $\pm$.001	.064 $\pm$.001	.020 $\pm$.001	.010 $\pm$.001
Ada-TS			.02 $\pm$.00	.26 $\pm$.03	.01 $\pm$.00	.241 $\pm$.018	.09 $\pm$.00	<u>.003 $\pm$.001</u>	.169 $\pm$.002	.077 $\pm$.002	.034 $\pm$.002	.000 $\pm$.002
Dirichlet			.02 $\pm$.00	.40 $\pm$.20	<u>.01 $\pm$.00</u>	.202 $\pm$.007	.09 $\pm$.00	<u>.003 $\pm$.001</u>	.166 $\pm$.002	.076 $\pm$.002	.033 $\pm$.002	.004 $\pm$.001
Isotonic			.01 $\pm$.00	.22 $\pm$.02	.01 $\pm$.00	.265 $\pm$.015	.09 $\pm$.00	.003 $\pm$.000	.172 $\pm$.002	.080 $\pm$.002	.036 $\pm$.001	.003 $\pm$.000
Platt Scaling			.02 $\pm$.00	.27 $\pm$.11	.01 $\pm$.00	<u>.194 $\pm$.004</u>	.09 $\pm$.00	.004 $\pm$.001	.169 $\pm$.002	.079 $\pm$.002	.035 $\pm$.001	.006 $\pm$.001
Temp. scaling			<u>.01 $\pm$.00</u>	.25 $\pm$.03	.01 $\pm$.00	.193 $\pm$.004	.09 $\pm$.00	.003 $\pm$.000	.168 $\pm$.002	.079 $\pm$.001	.035 $\pm$.001	.006 $\pm$.001
V.-Abers (OvA)			.02 $\pm$.00	<u>.22 $\pm$.04</u>	.01 $\pm$.00	.200 $\pm$.002	.09 $\pm$.00	.009 $\pm$.001	.179 $\pm$.001	.087 $\pm$.001	.042 $\pm$.001	.008 $\pm$.000
Naive CMCE			.02 $\pm$.00	.28 $\pm$.04	.01 $\pm$.00	.196 $\pm$.005	.09 $\pm$.00	.002 $\pm$.000	.166 $\pm$.002	.075 $\pm$.002	.032 $\pm$.001	.004 $\pm$.001
TS (ours)	0.01 APS	Const	.08 $\pm$.00	.26 $\pm$.00	.02 $\pm$.00	.253 $\pm$.006	.09 $\pm$.00	.036 $\pm$.002	.183 $\pm$.001	.093 $\pm$.000	.047 $\pm$.000	.010 $\pm$.000
TS (ours)	0.01 APS	Ada	.04 $\pm$.00	.45 $\pm$.08	.01 $\pm$.00	.279 $\pm$.014	.10 $\pm$.00	.007 $\pm$.000	.154 $\pm$.002	.060 $\pm$.002	.015 $\pm$.003	.015 $\pm$.003
TS (ours)	0.01 MSP	Const	.02 $\pm$.00	.34 $\pm$.05	.01 $\pm$.00	.200 $\pm$.005	.09 $\pm$.00	.003 $\pm$.000	.163 $\pm$.001	.072 $\pm$.001	.030 $\pm$.001	<u>.001 $\pm$.001</u>
TS (ours)	0.01 MSP	Ada	.03 $\pm$.00	.40 $\pm$.05	.01 $\pm$.00	.240 $\pm$.019	.09 $\pm$.00	.004 $\pm$.001	.157 $\pm$.003	.065 $\pm$.004	.021 $\pm$.004	.010 $\pm$.003
TS (ours)	0.05 APS	Const	.22 $\pm$.01	.56 $\pm$.01	.05 $\pm$.00	.450 $\pm$.015	.17 $\pm$.01	.107 $\pm$.004	.195 $\pm$.000	.098 $\pm$.000	.049 $\pm$.000	.010 $\pm$.000
TS (ours)	0.05 APS	Ada	.04 $\pm$.00	.52 $\pm$.18	.01 $\pm$.00	.289 $\pm$.008	.10 $\pm$.00	.007 $\pm$.000	.150 $\pm$.002	.054 $\pm$.002	<u>.007 $\pm$.003</u>	.011 $\pm$.001
TS (ours)	0.05 MSP	Const	.00 $\pm$.00	.60 $\pm$.16	.01 $\pm$.00	.282 $\pm$.019	.10 $\pm$.00	.010 $\pm$.001	.150 $\pm$.003	.050 $\pm$.003	.000 $\pm$.003	.001 $\pm$.001
TS (ours)	0.05 MSP	Ada	.03 $\pm$.01	.44 $\pm$.07	.01 $\pm$.00	.248 $\pm$.008	.10 $\pm$.00	.013 $\pm$.001	.155 $\pm$.002	.060 $\pm$.003	.014 $\pm$.004	.003 $\pm$.001
TS (ours)	0.1 APS	Const	.30 $\pm$.01	.70 $\pm$.04	.04 $\pm$.00	.677 $\pm$.016	.25 $\pm$.01	.155 $\pm$.004	.197 $\pm$.000	.099 $\pm$.000	.049 $\pm$.000	.009 $\pm$.000
TS (ours)	0.1 APS	Ada	.04 $\pm$.00	.74 $\pm$.23	.01 $\pm$.00	.271 $\pm$.010	.10 $\pm$.00	.006 $\pm$.001	.150 $\pm$.002	.054 $\pm$.002	.024 $\pm$.003	.009 $\pm$.003
TS (ours)	0.1 MSP	Const	.08 $\pm$.00	.41 $\pm$.03	.02 $\pm$.00	.270 $\pm$.005	.10 $\pm$.00	.032 $\pm$.000	.157 $\pm$.001	.064 $\pm$.001	.039 $\pm$.001	.007 $\pm$.000
TS (ours)	0.1 MSP	Ada	.08 $\pm$.00	.38 $\pm$.07	.02 $\pm$.00	.283 $\pm$.005	.10 $\pm$.00	.030 $\pm$.001	.155 $\pm$.002	.060 $\pm$.003	.035 $\pm$.002	.005 $\pm$.001
TS (ours)	0.2 APS	Const	.29 $\pm$.01	.81 $\pm$.00	.04 $\pm$.00	.645 $\pm$.014	.25 $\pm$.01	.133 $\pm$.003	.175 $\pm$.001	.082 $\pm$.001	.037 $\pm$.001	.004 $\pm$.001
TS (ours)	0.2 APS	Ada	.03 $\pm$.00	.77 $\pm$.05	.01 $\pm$.00	.263 $\pm$.010	.10 $\pm$.00	.008 $\pm$.003	.152 $\pm$.004	.075 $\pm$.005	.029 $\pm$.004	.005 $\pm$.002
TS (ours)	0.2 MSP	Const	.18 $\pm$.00	.39 $\pm$.00	.03 $\pm$.00	.388 $\pm$.006	.13 $\pm$.00	.064 $\pm$.001	.157 $\pm$.001	.071 $\pm$.001	.027 $\pm$.001	.003 $\pm$.001
TS (ours)	0.2 MSP	Ada	.18 $\pm$.00	.39 $\pm$.00	.03 $\pm$.00	.383 $\pm$.004	.13 $\pm$.00	.063 $\pm$.000	.157 $\pm$.001	.072 $\pm$.001	.029 $\pm$.001	.002 $\pm$.001

I.4 INATURALIST21

Finally, in Table 11 we present the best results for iNaturalist21, obtained with ViT, and in Table 1 we present the best results for our model.

Table 11 shows that OvR-style binary calibration methods (Isotonic, Platt) harm accuracy and worsen several key calibration metrics on this dataset with a large number of classes. By contrast, our methods preserve accuracy and improve NLL, Brier, and especially CMCE, while meeting coverage targets more reliably. Moreover, a particular instance is the best across all key calibration metrics (see Table 1). Temperature Scaling is competitive on NLL/Brier but is worse than our methods in terms of CMCE and target coverage.

I.5 TAKEAWAYS ACROSS DATASETS

Based on the observed results, we can make the following observations:

1. Our methods most closely match the desired $1 - \alpha$ coverage across CIFAR-100, ImageNet, and iNaturalist.
2. We consistently attain the best or near-best CMCE, highlighting the benefit of calibrating cumulative mass relative to conformal sets.
3. In large- K settings (ImageNet, iNaturalist), OvR binary methods often degrade calibration and even accuracy. In contrast, our methods remain stable and competitive on key calibration metrics while substantially improving set-mass alignment.

Table 6: Evaluation results on CIFAR-100 and ResNet-56. **Orange/Green** indicate worse/better performance than the uncalibrated model. **Bold** and underline mark the best and second-best values.

Calibrator	α	Score Transf.	ECE	MCE	cw-ECE	NLL	Brier Score	CMCE	α -CMCE (0.2)	α -CMCE (0.1)	α -CMCE (0.05)	α -CMCE (0.01)
Base			.14 $\pm$.00	.32 $\pm$.02	.00 $\pm$.00	1.288 $\pm$.011	.42 $\pm$.00	.013 $\pm$.000	.002 $\pm$.002	.061 $\pm$.002	.081 $\pm$.001	.066 $\pm$.001
Ada-TS			.15 $\pm$.00	.29 $\pm$.03	.00 $\pm$.00	2.368 $\pm$.047	.45 $\pm$.00	.019 $\pm$.001	.021 $\pm$.003	.055 $\pm$.003	.087 $\pm$.003	.099 $\pm$.002
Dirichlet			.41 $\pm$.01	.74 $\pm$.04	.01 $\pm$.00	10.440 $\pm$.545	.82 $\pm$.02	.034 $\pm$.002	.212 $\pm$.009	.310 $\pm$.010	.358 $\pm$.011	.393 $\pm$.012
Isotonic			.06 $\pm$.00	.33 $\pm$.34	<u>.00 $\pm$.00</u>	1.728 $\pm$.070	.39 $\pm$.00	.017 $\pm$.002	.062 $\pm$.004	.001 $\pm$.004	.023 $\pm$.003	.031 $\pm$.004
Platt Scaling			.15 $\pm$.00	.22 $\pm$.01	.00 $\pm$.00	1.139 $\pm$.006	.42 $\pm$.00	.015 $\pm$.001	.131 $\pm$.002	.068 $\pm$.001	.034 $\pm$.001	.007 $\pm$.000
Temp. scaling			.03 $\pm$.00	<u>.11 $\pm$.03</u>	.00 $\pm$.00	1.038 $\pm$.007	.38 $\pm$.00	.004 $\pm$.000	.093 $\pm$.001	.039 $\pm$.002	.014 $\pm$.001	.002 $\pm$.001
V.-Abers (OvA)			.08 $\pm$.00	.17 $\pm$.01	.00 $\pm$.00	1.119 $\pm$.008	.39 $\pm$.00	.038 $\pm$.002	.138 $\pm$.003	.077 $\pm$.002	.044 $\pm$.001	.010 $\pm$.000
Naive CMCE			.04 $\pm$.00	.11 $\pm$.01	.00 $\pm$.00	1.041 $\pm$.007	.38 $\pm$.00	.001 $\pm$.001	.081 $\pm$.004	.027 $\pm$.003	<u>.006 $\pm$.003</u>	.003 $\pm$.002
TS (ours)	0.01 APS	Const	.18 $\pm$.01	.41 $\pm$.02	.00 $\pm$.00	1.417 $\pm$.025	.45 $\pm$.01	.098 $\pm$.005	.169 $\pm$.002	.087 $\pm$.002	.045 $\pm$.001	.009 $\pm$.000
TS (ours)	0.01 APS	Ada	.15 $\pm$.02	.35 $\pm$.04	.00 $\pm$.00	1.337 $\pm$.062	.43 $\pm$.02	.031 $\pm$.041	.034 $\pm$.077	.033 $\pm$.067	.057 $\pm$.057	.052 $\pm$.035
TS (ours)	0.01 MSP	Const	<u>.04 $\pm$.00</u>	.22 $\pm$.03	.00 $\pm$.00	1.046 $\pm$.007	.38 $\pm$.00	.008 $\pm$.001	.092 $\pm$.003	.035 $\pm$.002	.014 $\pm$.002	.002 $\pm$.001
TS (ours)	0.01 MSP	Ada	.14 $\pm$.00	.33 $\pm$.02	.00 $\pm$.00	1.301 $\pm$.013	.42 $\pm$.00	.013 $\pm$.000	.001 $\pm$.002	.062 $\pm$.002	.083 $\pm$.002	.069 $\pm$.002
TS (ours)	0.05 APS	Const	.19 $\pm$.01	.39 $\pm$.01	.00 $\pm$.00	1.404 $\pm$.029	.45 $\pm$.01	.092 $\pm$.004	.164 $\pm$.001	.083 $\pm$.001	.041 $\pm$.001	.008 $\pm$.001
TS (ours)	0.05 APS	Ada	.15 $\pm$.00	.47 $\pm$.27	.00 $\pm$.00	1.333 $\pm$.013	.42 $\pm$.00	.013 $\pm$.000	.003 $\pm$.002	.067 $\pm$.003	.088 $\pm$.002	.069 $\pm$.001
TS (ours)	0.05 MSP	Const	<u>.04 $\pm$.00</u>	.12 $\pm$.01	.00 $\pm$.00	<u>1.039 $\pm$.008</u>	.38 $\pm$.00	<u>.004 $\pm$.000</u>	.074 $\pm$.003	.022 $\pm$.002	.003 $\pm$.002	<u>.001 $\pm$.001</u>
TS (ours)	0.05 MSP	Ada	.13 $\pm$.00	.33 $\pm$.03	.00 $\pm$.00	1.315 $\pm$.010	.42 $\pm$.00	.012 $\pm$.000	.002 $\pm$.001	.069 $\pm$.001	.093 $\pm$.001	.061 $\pm$.002
TS (ours)	0.1 APS	Const	.20 $\pm$.01	.38 $\pm$.01	.00 $\pm$.00	1.411 $\pm$.025	.46 $\pm$.01	.093 $\pm$.003	.168 $\pm$.001	.084 $\pm$.001	.043 $\pm$.001	.008 $\pm$.001
TS (ours)	0.1 APS	Ada	.15 $\pm$.00	.44 $\pm$.25	.00 $\pm$.00	1.349 $\pm$.017	.42 $\pm$.00	.013 $\pm$.000	.003 $\pm$.002	.070 $\pm$.002	.081 $\pm$.001	.068 $\pm$.001
TS (ours)	0.1 MSP	Const	.06 $\pm$.00	.14 $\pm$.03	<u>.00 $\pm$.00</u>	1.065 $\pm$.009	.39 $\pm$.00	.005 $\pm$.000	.069 $\pm$.005	.003 $\pm$.004	.009 $\pm$.003	.000 $\pm$.001
TS (ours)	0.1 MSP	Ada	.11 $\pm$.00	.35 $\pm$.02	.00 $\pm$.00	1.285 $\pm$.024	.42 $\pm$.00	.009 $\pm$.000	.002 $\pm$.004	.071 $\pm$.004	.068 $\pm$.003	.054 $\pm$.003
TS (ours)	0.2 APS	Const	.21 $\pm$.01	.37 $\pm$.01	.00 $\pm$.00	1.402 $\pm$.017	.46 $\pm$.00	.092 $\pm$.003	.169 $\pm$.001	.088 $\pm$.000	.043 $\pm$.000	.007 $\pm$.000
TS (ours)	0.2 APS	Ada	.13 $\pm$.00	.37 $\pm$.08	.00 $\pm$.00	1.320 $\pm$.017	.42 $\pm$.00	.010 $\pm$.001	<u>.001 $\pm$.003</u>	.047 $\pm$.003	.070 $\pm$.002	.062 $\pm$.001
TS (ours)	0.2 MSP	Const	.06 $\pm$.01	.54 $\pm$.10	<u>.00 $\pm$.00</u>	1.194 $\pm$.018	.41 $\pm$.00	.011 $\pm$.000	.001 $\pm$.006	<u>.003 $\pm$.005</u>	.015 $\pm$.004	.017 $\pm$.003
TS (ours)	0.2 MSP	Ada	.12 $\pm$.01	.32 $\pm$.02	.00 $\pm$.00	1.267 $\pm$.014	.42 $\pm$.00	.012 $\pm$.000	.003 $\pm$.002	.030 $\pm$.004	.046 $\pm$.004	.041 $\pm$.003

Table 7: Evaluation results on CIFAR-100 and ViT-B16. Metrics* are scaled by 100. **Orange/Green** indicate worse/better performance than the uncalibrated model. **Bold** and underline mark the best and second-best values.

Calibrator	α	Score Transf.	ECE	MCE	cw-ECE	NLL	Brier Score	CMCE	α -CMCE (0.2)	α -CMCE (0.1)	α -CMCE (0.05)	α -CMCE (0.01)
Base			.05 $\pm$.00	.34 $\pm$.09	.00 $\pm$.00	<u>.437 $\pm$.002</u>	.17 $\pm$.00	.003 $\pm$.000	.132 $\pm$.001	.049 $\pm$.001	.013 $\pm$.001	<u>.001 $\pm$.001</u>
Ada-TS			.07 $\pm$.00	.36 $\pm$.06	.00 $\pm$.00	.889 $\pm$.032	.18 $\pm$.00	.008 $\pm$.001	.128 $\pm$.001	.036 $\pm$.001	.007 $\pm$.002	.035 $\pm$.002
Dirichlet			.13 $\pm$.01	.70 $\pm$.15	.00 $\pm$.00	3.134 $\pm$.463	.27 $\pm$.01	.012 $\pm$.003	.065 $\pm$.007	<u>.033 $\pm$.008</u>	.082 $\pm$.009	.118 $\pm$.010
Isotonic			<u>.04 $\pm$.00</u>	.21 $\pm$.03	<u>.00 $\pm$.00</u>	1.067 $\pm$.116	.16 $\pm$.00	.014 $\pm$.002	.138 $\pm$.003	.049 $\pm$.003	.007 $\pm$.004	.024 $\pm$.004
Platt Scaling			.04 $\pm$.00	.49 $\pm$.40	.00 $\pm$.00	.404 $\pm$.006	.16 $\pm$.00	.007 $\pm$.001	.144 $\pm$.002	.064 $\pm$.001	.033 $\pm$.002	.008 $\pm$.001
Temp. scaling			.04 $\pm$.02	.30 $\pm$.06	.00 $\pm$.00	.446 $\pm$.029	<u>.16 $\pm$.00</u>	.027 $\pm$.016	.159 $\pm$.013	.080 $\pm$.013	.043 $\pm$.006	.010 $\pm$.000
V.-Abers (OvA)			.10 $\pm$.00	<u>.30 $\pm$.04</u>	.00 $\pm$.00	.484 $\pm$.004	.17 $\pm$.00	.044 $\pm$.001	.183 $\pm$.001	.095 $\pm$.001	.049 $\pm$.000	.010 $\pm$.000
Naive CMCE			.06 $\pm$.00	.32 $\pm$.01	.00 $\pm$.00	.456 $\pm$.006	.17 $\pm$.00	.001 $\pm$.001	.126 $\pm$.001	.041 $\pm$.001	.005 $\pm$.002	.009 $\pm$.003
TS (ours)	0.01 APS	Const	<u>.04 $\pm$.00</u>	.40 $\pm$.04	.00 $\pm$.00	.455 $\pm$.004	<u>.16 $\pm$.00</u>	.029 $\pm$.002	.161 $\pm$.001	.081 $\pm$.001	.043 $\pm$.001	.010 $\pm$.000
TS (ours)	0.01 APS	Ada	.06 $\pm$.00	.57 $\pm$.31	.00 $\pm$.00	.468 $\pm$.008	.17 $\pm$.00	<u>.002 $\pm$.001</u>	.124 $\pm$.002	.041 $\pm$.002	.005 $\pm$.002	.005 $\pm$.001
TS (ours)	0.01 MSP	Const	.06 $\pm$.00	.43 $\pm$.09	.00 $\pm$.00	.455 $\pm$.002	.17 $\pm$.00	.004 $\pm$.001	.124 $\pm$.001	.038 $\pm$.002	<u>.002 $\pm$.002</u>	.000 $\pm$.002
TS (ours)	0.01 MSP	Ada	.05 $\pm$.00	.39 $\pm$.14	<u>.00 $\pm$.00</u>	.448 $\pm$.004	.17 $\pm$.00	.004 $\pm$.000	.128 $\pm$.001	.044 $\pm$.001	.007 $\pm$.002	.004 $\pm$.003
TS (ours)	0.05 APS	Const	.10 $\pm$.00	.45 $\pm$.03	.00 $\pm$.00	.562 $\pm$.003	.18 $\pm$.00	.066 $\pm$.001	.184 $\pm$.001	.096 $\pm$.001	.050 $\pm$.000	.010 $\pm$.000
TS (ours)	0.05 APS	Ada	.06 $\pm$.01	.68 $\pm$.22	.00 $\pm$.00	.486 $\pm$.014	.18 $\pm$.00	.006 $\pm$.006	.126 $\pm$.008	.041 $\pm$.006	.003 $\pm$.005	.002 $\pm$.001
TS (ours)	0.05 MSP	Const	<u>.04 $\pm$.00</u>	.36 $\pm$.19	<u>.00 $\pm$.00</u>	.456 $\pm$.002	.17 $\pm$.00	.014 $\pm$.000	.122 $\pm$.000	.040 $\pm$.001	.001 $\pm$.002	.009 $\pm$.000
TS (ours)	0.05 MSP	Ada	<u>.03 $\pm$.00</u>	.38 $\pm$.06	.00 $\pm$.00	.472 $\pm$.005	.17 $\pm$.00	.013 $\pm$.001	.126 $\pm$.002	.036 $\pm$.003	.007 $\pm$.004	.005 $\pm$.001
TS (ours)	0.1 APS	Const	.16 $\pm$.00	.46 $\pm$.02	.00 $\pm$.00	.648 $\pm$.004	.20 $\pm$.00	.091 $\pm$.001	.189 $\pm$.000	.099 $\pm$.000	.050 $\pm$.000	.010 $\pm$.000
TS (ours)	0.1 APS	Ada	.12 $\pm$.09	.60 $\pm$.16	.00 $\pm$.00	.619 $\pm$.136	.21 $\pm$.04	.058 $\pm$.045	.149 $\pm$.014	.060 $\pm$.013	.028 $\pm$.008	.002 $\pm$.004
TS (ours)	0.1 MSP	Const	.01 $\pm$.00	.40 $\pm$.12	<u>.00 $\pm$.00</u>	.543 $\pm$.007	.18 $\pm$.00	.029 $\pm$.000	<u>.109 $\pm$.001</u>	.009 $\pm$.001	.032 $\pm$.001	.002 $\pm$.001
TS (ours)	0.1 MSP	Ada	.07 $\pm$.00	.34 $\pm$.08	.00 $\pm$.00	.484 $\pm$.004	.17 $\pm$.00	.030 $\pm$.000	.129 $\pm$.000	.043 $\pm$.001	.036 $\pm$.002	.009 $\pm$.000
TS (ours)	0.2 APS	Const	.22 $\pm$.00	.60 $\pm$.03	.01 $\pm$.00	.760 $\pm$.002	.25 $\pm$.00	.116 $\pm$.001	.185 $\pm$.001	.094 $\pm$.000	.048 $\pm$.000	.009 $\pm$.000
TS (ours)	0.2 APS	Ada	.13 $\pm$.06	.62 $\pm$.15	.00 $\pm$.00	.629 $\pm$.090	.21 $\pm$.03	.066 $\pm$.030	.155 $\pm$.008	.076 $\pm$.005	.032 $\pm$.005	.003 $\pm$.002
TS (ours)	0.2 MSP	Const	.16 $\pm$.00	.34 $\pm$.09	.00 $\pm$.00	.556 $\pm$.002	.19 $\pm$.00	.064 $\pm$.000	.132 $\pm$.001	.078 $\pm$.001	.042 $\pm$.001	.010 $\pm$.000
TS (ours)	0.2 MSP	Ada	.14 $\pm$.01	.37 $\pm$.08	.00 $\pm$.00	.554 $\pm$.007	.19 $\pm$.00	.060 $\pm$.002	.125 $\pm$.002	.076 $\pm$.004	.037 $\pm$.002	.008 $\pm$.001

Table 8: Evaluation results on CIFAR-100 and ViT-B16 model where our calibrators have $\alpha = 0.01$, MSP nonconformity score and constant transformation. Metrics* are scaled by 100. **Orange/Green** indicate worse/better performance than the uncalibrated model. **Bold** and underline mark the best and second-best values.

Calibrator	ECE	MCE	cw-ECE	NLL	Brier Score	CMCE	α -CMCE (0.01)
Base	.05 $\pm$.00	.34 $\pm$.09	<u>.00 $\pm$.00</u>	<u>.437 $\pm$.002</u>	.17 $\pm$.00	<u>.003 $\pm$.000</u>	<u>.001 $\pm$.001</u>
Ada-TS	.07 $\pm$.00	.36 $\pm$.06	<u>.00 $\pm$.00</u>	.889 $\pm$.032	.18 $\pm$.00	.008 $\pm$.001	.035 $\pm$.002
Dirichlet	.13 $\pm$.01	.70 $\pm$.15	.00 $\pm$.00	3.134 $\pm$.463	.27 $\pm$.01	.012 $\pm$.003	.118 $\pm$.010
Isotonic	<u>.04 $\pm$.00</u>	.21 $\pm$.03	.00 $\pm$.00	1.067 $\pm$.116	.16 $\pm$.00	.014 $\pm$.002	.024 $\pm$.004
Platt Scaling	.04 $\pm$.00	.49 $\pm$.40	.00 $\pm$.00	.404 $\pm$.006	.16 $\pm$.00	.007 $\pm$.001	.008 $\pm$.001
Temp. scaling	.04 $\pm$.02	.30 $\pm$.06	.00 $\pm$.00	.446 $\pm$.029	<u>.16 $\pm$.00</u>	.027 $\pm$.016	.010 $\pm$.000
V-Abers (OvA)	.10 $\pm$.00	<u>.30 $\pm$.04</u>	.00 $\pm$.00	.484 $\pm$.004	.17 $\pm$.00	.044 $\pm$.001	.010 $\pm$.000
Naive CMCE	.06 $\pm$.00	.32 $\pm$.01	<u>.00 $\pm$.00</u>	.456 $\pm$.006	.17 $\pm$.00	.001 $\pm$.001	.009 $\pm$.003
TS (ours)	.06 $\pm$.00	.43 $\pm$.09	<u>.00 $\pm$.00</u>	.455 $\pm$.002	.17 $\pm$.00	.004 $\pm$.001	.000 $\pm$.002

Table 9: Evaluation results on ImageNet with calibrators configured using $\alpha = 0.01$, the APS nonconformity score, and adaptive transformation. Metrics* denote values scaled by a factor of 100. **Red** indicates performance worse than the uncalibrated model, whereas **Green** indicates performance better than the uncalibrated model. **Bold** highlights the best value, and underline highlights the second-best value.

Calibrator	ECE*	MCE*	NLL	Brier Score*	CMCE*	α -CMCE (0.01)*
Base	2.41 $\pm$ 0.05	8.92 $\pm$ 1.23	0.70 $\pm$ 0.00	26.06 $\pm$ 0.05	0.43 $\pm$ 0.01	0.40 $\pm$ 0.02
Ada-TS	10.81 $\pm$ 0.41	25.53 $\pm$ 1.31	1.69 $\pm$ 0.10	30.45 $\pm$ 0.30	0.58 $\pm$ 0.10	7.21 $\pm$ 0.48
Isotonic	4.43 $\pm$ 0.08	11.35 $\pm$ 1.00	2.12 $\pm$ 0.02	27.36 $\pm$ 0.15	2.56 $\pm$ 0.09	5.71 $\pm$ 0.09
Platt Scaling	12.89 $\pm$ 0.19	20.67 $\pm$ 1.23	0.81 $\pm$ 0.00	28.62 $\pm$ 0.07	1.26 $\pm$ 0.02	0.92 $\pm$ 0.02
Temp. scaling	<u>2.49 $\pm$ 0.04</u>	<u>9.08 $\pm$ 1.04</u>	0.70 $\pm$ 0.00	26.06 $\pm$ 0.05	0.42 $\pm$ 0.01	0.39 $\pm$ 0.03
V-Abers (OvA)	23.93 $\pm$ 0.25	33.66 $\pm$ 0.77	1.00 $\pm$ 0.00	33.32 $\pm$ 0.14	7.92 $\pm$ 0.04	1.00 $\pm$ 0.00
Naive CMCE	5.82 $\pm$ 0.09	16.18 $\pm$ 0.99	0.73 $\pm$ 0.00	26.72 $\pm$ 0.07	0.02 $\pm$ 0.00	0.93 $\pm$ 0.06
TS (ours)	2.87 $\pm$ 0.06	13.34 $\pm$ 9.96	<u>0.73 $\pm$ 0.00</u>	<u>26.35 $\pm$ 0.07</u>	<u>0.34 $\pm$ 0.01</u>	0.01 $\pm$ 0.04

Table 10: Evaluation results on Imagenet. textcolorredOrange/Green indicate worse/better performance than the uncalibrated model. **Bold** and underline mark the best and second-best values.

Calibrator	α	Score Transf.	ECE	MCE	cw-ECE	NLL	Brier Score	CMCE	α -CMCE (0.2)	α -CMCE (0.1)	α -CMCE (0.05)	α -CMCE (0.01)
Base			<u>.02 $\pm$.00</u>	.09 $\pm$.01	.00 $\pm$.00	.704 $\pm$.002	.26 $\pm$.00	.004 $\pm$.000	.131 $\pm$.001	.059 $\pm$.001	.026 $\pm$.001	.004 $\pm$.000
Ada-TS			.11 $\pm$.00	.26 $\pm$.01	.00 $\pm$.00	1.688 $\pm$.096	.30 $\pm$.00	.006 $\pm$.001	.069 $\pm$.004	.014 $\pm$.004	.052 $\pm$.004	.072 $\pm$.005
Isotonic			.04 $\pm$.00	.11 $\pm$.01	.00 $\pm$.00	2.124 $\pm$.025	.27 $\pm$.00	.026 $\pm$.001	.089 $\pm$.002	<u>.010 $\pm$.001</u>	.029 $\pm$.001	.057 $\pm$.001
Platt Scaling			.13 $\pm$.00	.21 $\pm$.01	.00 $\pm$.00	.806 $\pm$.004	.29 $\pm$.00	.013 $\pm$.000	.167 $\pm$.001	.087 $\pm$.000	.047 $\pm$.000	.009 $\pm$.000
Temp. scaling			<u>.02 $\pm$.00</u>	.09 $\pm$.01	.00 $\pm$.00	.704 $\pm$.002	.26 $\pm$.00	.004 $\pm$.000	.130 $\pm$.001	.058 $\pm$.001	.025 $\pm$.001	.004 $\pm$.000
V-Abers (OvA)			.24 $\pm$.00	.34 $\pm$.01	.00 $\pm$.00	1.001 $\pm$.005	.33 $\pm$.00	.079 $\pm$.000	.195 $\pm$.000	.100 $\pm$.000	.050 $\pm$.000	.010 $\pm$.000
Naive CMCE			.06 $\pm$.00	.16 $\pm$.01	.00 $\pm$.00	.734 $\pm$.003	.27 $\pm$.00	.000 $\pm$.000	.102 $\pm$.001	.031 $\pm$.001	<u>.002 $\pm$.000</u>	.009 $\pm$.001
TS (ours)	0.01	APS Const	.16 $\pm$.00	.41 $\pm$.01	<u>.00 $\pm$.00</u>	1.104 $\pm$.009	.32 $\pm$.00	.073 $\pm$.001	.186 $\pm$.001	.095 $\pm$.000	.048 $\pm$.000	.010 $\pm$.000
TS (ours)	0.01	APS Ada	.03 $\pm$.00	.13 $\pm$.00	.00 $\pm$.00	.727 $\pm$.003	.26 $\pm$.00	.003 $\pm$.000	.123 $\pm$.001	.052 $\pm$.001	.020 $\pm$.001	.000 $\pm$.000
TS (ours)	0.01	MSP Const	.04 $\pm$.00	.15 $\pm$.00	.00 $\pm$.00	.723 $\pm$.003	.27 $\pm$.00	<u>.001 $\pm$.000</u>	.099 $\pm$.001	.029 $\pm$.001	.004 $\pm$.001	.001 $\pm$.001
TS (ours)	0.01	MSP Ada	.03 $\pm$.00	<u>.09 $\pm$.01</u>	.00 $\pm$.00	<u>.720 $\pm$.005</u>	<u>.26 $\pm$.00</u>	.004 $\pm$.000	.129 $\pm$.000	.056 $\pm$.001	.022 $\pm$.001	.002 $\pm$.001
TS (ours)	0.05	APS Const	.22 $\pm$.00	.49 $\pm$.01	.00 $\pm$.00	1.244 $\pm$.007	.35 $\pm$.00	.093 $\pm$.001	.188 $\pm$.000	.097 $\pm$.000	.049 $\pm$.000	.010 $\pm$.000
TS (ours)	0.05	APS Ada	.04 $\pm$.00	.28 $\pm$.15	.00 $\pm$.00	.756 $\pm$.007	.27 $\pm$.00	.003 $\pm$.000	.114 $\pm$.002	.042 $\pm$.003	.010 $\pm$.003	.002 $\pm$.001
TS (ours)	0.05	MSP Const	.04 $\pm$.00	.17 $\pm$.03	.00 $\pm$.00	.754 $\pm$.004	.27 $\pm$.00	.006 $\pm$.000	.101 $\pm$.001	.031 $\pm$.002	.002 $\pm$.003	.000 $\pm$.001
TS (ours)	0.05	MSP Ada	.02 $\pm$.00	.09 $\pm$.01	.00 $\pm$.00	.727 $\pm$.005	.26 $\pm$.00	.008 $\pm$.000	.127 $\pm$.002	.052 $\pm$.002	.017 $\pm$.003	.004 $\pm$.001
TS (ours)	0.1	APS Const	.24 $\pm$.00	.52 $\pm$.00	.00 $\pm$.00	1.267 $\pm$.007	.36 $\pm$.00	.097 $\pm$.001	.187 $\pm$.000	.096 $\pm$.000	.049 $\pm$.000	.010 $\pm$.000
TS (ours)	0.1	APS Ada	.08 $\pm$.09	.35 $\pm$.12	<u>.00 $\pm$.00</u>	.869 $\pm$.216	.29 $\pm$.04	.022 $\pm$.041	.126 $\pm$.034	.048 $\pm$.026	.023 $\pm$.014	<u>.000 $\pm$.005</u>
TS (ours)	0.1	MSP Const	.04 $\pm$.00	.34 $\pm$.03	.00 $\pm$.00	.841 $\pm$.007	.28 $\pm$.00	.013 $\pm$.000	.097 $\pm$.001	.003 $\pm$.002	.003 $\pm$.001	.010 $\pm$.001
TS (ours)	0.1	MSP Ada	.04 $\pm$.00	.09 $\pm$.01	.00 $\pm$.00	.743 $\pm$.002	.27 $\pm$.00	.014 $\pm$.000	.126 $\pm$.001	.051 $\pm$.001	.030 $\pm$.000	.006 $\pm$.000
TS (ours)	0.2	APS Const	.24 $\pm$.00	.52 $\pm$.00	.00 $\pm$.00	1.263 $\pm$.005	.37 $\pm$.00	.098 $\pm$.001	.185 $\pm$.000	.097 $\pm$.000	.049 $\pm$.000	.009 $\pm$.000
TS (ours)	0.2	APS Ada	.03 $\pm$.00	.38 $\pm$.05	.00 $\pm$.00	.785 $\pm$.003	.27 $\pm$.00	.007 $\pm$.000	.108 $\pm$.002	.054 $\pm$.002	.019 $\pm$.001	.002 $\pm$.000
TS (ours)	0.2	MSP Const	.07 $\pm$.00	.12 $\pm$.02	.00 $\pm$.00	.878 $\pm$.003	.29 $\pm$.00	.033 $\pm$.000	<u>.075 $\pm$.001</u>	.061 $\pm$.001	.025 $\pm$.001	.002 $\pm$.001
TS (ours)	0.2	MSP Ada	.09 $\pm$.00	.14 $\pm$.00	<u>.00 $\pm$.00</u>	.805 $\pm$.003	.28 $\pm$.00	.030 $\pm$.000	.120 $\pm$.001	.067 $\pm$.001	.032 $\pm$.001	.005 $\pm$.001

Table 11: Calibration results on iNaturalist21 (10k classes). Binary OvR methods often reduce accuracy and worsen calibration, whereas our proposed approaches preserve accuracy and achieve superior CMCE, coverage, and other key calibration metrics.

Calibrator	α	Score Transf.	Accuracy	ECE	MCE	cw-ECE	NLL	Brier Score	CMCE	α -CMCE (0.2)	α -CMCE (0.1)	α -CMCE (0.05)	α -CMCE (0.01)
Base			.78 $\pm$.00	.50 $\pm$.00	.65 $\pm$.01	.0001 $\pm$.0000	2.043 $\pm$.015	.61 $\pm$.00	.056 $\pm$.000	.188 $\pm$.001	.096 $\pm$.000	.048 $\pm$.000	.010 $\pm$.000
Ada-TS			.78 $\pm$.00	.11 $\pm$.01	.27 $\pm$.02	.0000 $\pm$.0000	2.238 $\pm$.098	.36 $\pm$.00	.009 $\pm$.001	.052 $\pm$.007	.030 $\pm$.007	.067 $\pm$.007	.086 $\pm$.006
Isotonic			.45 $\pm$.00	.26 $\pm$.01	.38 $\pm$.01	.0001 $\pm$.0000	13.707 $\pm$.119	.80 $\pm$.01	.175 $\pm$.006	.309 $\pm$.004	.400 $\pm$.005	.445 $\pm$.005	.480 $\pm$.004
Platt Scaling			.47 $\pm$.00	.14 $\pm$.00	.30 $\pm$.00	.0001 $\pm$.0000	12.506 $\pm$.090	.68 $\pm$.00	.246 $\pm$.003	.232 $\pm$.003	.330 $\pm$.003	.379 $\pm$.003	.418 $\pm$.003
Temp. scaling			.78 $\pm$.00	.13 $\pm$.08	.26 $\pm$.11	.0001 $\pm$.0000	1.293 $\pm$.100	.36 $\pm$.04	.003 $\pm$.004	.121 $\pm$.021	.046 $\pm$.019	.013 $\pm$.016	.004 $\pm$.009
Naive CMCE			.78 $\pm$.00	.13 $\pm$.04	.26 $\pm$.05	.0001 $\pm$.0000	1.274 $\pm$.026	.35 $\pm$.01	.001 $\pm$.000	.121 $\pm$.011	.048 $\pm$.009	.014 $\pm$.009	.003 $\pm$.006
TS (ours)	0.01	APS Const	.78 $\pm$.00	.26 $\pm$.01	.39 $\pm$.02	.0001 $\pm$.0000	1.427 $\pm$.020	.41 $\pm$.01	.007 $\pm$.001	.148 $\pm$.003	.071 $\pm$.002	.035 $\pm$.002	.007 $\pm$.001
TS (ours)	0.01	APS Ada	.78 $\pm$.00	.44 $\pm$.02	.61 $\pm$.01	.0001 $\pm$.0000	1.920 $\pm$.050	.57 $\pm$.01	.047 $\pm$.002	.177 $\pm$.003	.087 $\pm$.002	.042 $\pm$.001	.006 $\pm$.001
TS (ours)	0.01	MSP Const	.78 $\pm$.00	.14 $\pm$.01	.24 $\pm$.02	.0001 $\pm$.0000	1.267 $\pm$.016	.35 $\pm$.00	.001 $\pm$.000	.119 $\pm$.003	.047 $\pm$.002	.016 $\pm$.002	.000 $\pm$.001
TS (ours)	0.01	MSP Ada	.78 $\pm$.00	.49 $\pm$.00	.65 $\pm$.01	.0001 $\pm$.0000	2.040 $\pm$.015	.61 $\pm$.00	.054 $\pm$.000	.186 $\pm$.001	.094 $\pm$.001	.046 $\pm$.001	.008 $\pm$.001
TS (ours)	0.05	APS Const	.78 $\pm$.00	.28 $\pm$.01	.42 $\pm$.01	.0001 $\pm$.0000	1.472 $\pm$.019	.42 $\pm$.00	.010 $\pm$.001	.151 $\pm$.001	.072 $\pm$.002	.035 $\pm$.001	.007 $\pm$.001
TS (ours)	0.05	APS Ada	.78 $\pm$.00	.39 $\pm$.01	.56 $\pm$.01	.0001 $\pm$.0000	1.823 $\pm$.015	.53 $\pm$.00	.040 $\pm$.000	.163 $\pm$.004	.077 $\pm$.002	.034 $\pm$.002	.003 $\pm$.001
TS (ours)	0.05	MSP Const	.78 $\pm$.00	.08 $\pm$.01	.13 $\pm$.01	.0001 $\pm$.0000	1.244 $\pm$.018	.32 $\pm$.00	.001 $\pm$.000	.094 $\pm$.005	.028 $\pm$.003	.001 $\pm$.003	.013 $\pm$.002
TS (ours)	0.05	MSP Ada	.78 $\pm$.00	.45 $\pm$.01	.65 $\pm$.01	.0001 $\pm$.0000	1.991 $\pm$.014	.59 $\pm$.00	.051 $\pm$.001	.181 $\pm$.002	.089 $\pm$.001	.042 $\pm$.001	.007 $\pm$.001
TS (ours)	0.1	APS Const	.78 $\pm$.00	.28 $\pm$.00	.41 $\pm$.01	.0001 $\pm$.0000	1.479 $\pm$.018	.42 $\pm$.00	.010 $\pm$.000	.149 $\pm$.001	.069 $\pm$.001	.035 $\pm$.001	.006 $\pm$.000
TS (ours)	0.1	MSP Const	.78 $\pm$.00	.07 $\pm$.00	.21 $\pm$.08	.0001 $\pm$.0000	1.379 $\pm$.036	.33 $\pm$.00	.001 $\pm$.000	.068 $\pm$.005	.002 $\pm$.005	.025 $\pm$.005	.035 $\pm$.003
TS (ours)	0.1	APS Ada	.78 $\pm$.00	.37 $\pm$.01	.52 $\pm$.01	.0001 $\pm$.0000	1.783 $\pm$.026	.51 $\pm$.01	.037 $\pm$.001	.157 $\pm$.003	.072 $\pm$.001	.033 $\pm$.001	.001 $\pm$.001
TS (ours)	0.1	MSP Ada	.78 $\pm$.00	.39 $\pm$.04	.64 $\pm$.02	.0001 $\pm$.0000	1.886 $\pm$.062	.55 $\pm$.03	.044 $\pm$.005	.173 $\pm$.006	.081 $\pm$.005	.041 $\pm$.003	.005 $\pm$.002
TS (ours)	0.2	APS Const	.78 $\pm$.00	.30 $\pm$.00	.42 $\pm$.01	.0001 $\pm$.0000	1.510 $\pm$.015	.43 $\pm$.00	.013 $\pm$.000	.146 $\pm$.002	.071 $\pm$.001	.034 $\pm$.001	.005 $\pm$.001
TS (ours)	0.2	MSP Const	.78 $\pm$.00	.10 $\pm$.00	.30 $\pm$.02	.0001 $\pm$.0000	1.436 $\pm$.023	.35 $\pm$.00	.009 $\pm$.000	.090 $\pm$.002	.035 $\pm$.003	.001 $\pm$.002	.019 $\pm$.002
TS (ours)	0.2	APS Ada	.78 $\pm$.00	.37 $\pm$.01	.50 $\pm$.01	.0001 $\pm$.0000	1.793 $\pm$.027	.51 $\pm$.01	.038 $\pm$.001	.152 $\pm$.001	.073 $\pm$.001	.031 $\pm$.001	.000 $\pm$.001
TS (ours)	0.2	MSP Ada	.78 $\pm$.00	.37 $\pm$.01	.62 $\pm$.01	.0001 $\pm$.0000	1.813 $\pm$.017	.52 $\pm$.00	.040 $\pm$.001	.167 $\pm$.002	.086 $\pm$.001	.041 $\pm$.000	.005 $\pm$.000