
Class conditional conformal prediction for multiple inputs by p-value aggregation

Jean-Baptiste Fermanian

IMAG, IROKO, Univ Montpellier, Inria, CNRS,
Montpellier, France

jean-baptiste.fermanian@inria.fr

Mohamed Hebiri

LAMA, Université Gustave Eiffel,
Paris, France

Mohamed.Hebiri@univ-eiffel.fr

Joseph Salmon

IMAG, IROKO, Univ Montpellier, Inria, CNRS,
Montpellier, France

joseph.salmon@inria.fr

Abstract

Conformal prediction methods are statistical tools designed to quantify uncertainty and generate predictive sets with guaranteed coverage probabilities. This work introduces an innovative refinement to these methods for classification tasks, specifically tailored for scenarios where multiple observations (multi-inputs) of a single instance are available at prediction time. Our approach is particularly motivated by applications in citizen science, where multiple images of the same plant or animal are captured by individuals. Our method integrates the information from each observation into conformal prediction, enabling a reduction in the size of the predicted label set while preserving the required class-conditional coverage guarantee. The approach is based on the aggregation of conformal p-values computed from each observation of a multi-input. By exploiting the exact distribution of these p-values, we propose a general aggregation framework using an abstract scoring function, encompassing many classical statistical tools. Knowledge of this distribution also enables refined versions of standard strategies, such as majority voting. We evaluate our method on simulated and real data, with a particular focus on Pl@ntNet, a prominent citizen science platform that facilitates the collection and identification of plant species through user-submitted images.

1 Introduction

As modern algorithms and the data they process become increasingly complex, the importance of quantifying their uncertainty has grown significantly. One systematic method to meet this requirement is (*split*) *conformal prediction*, which leverages a scoring function and labeled calibration data to generate a set of potential outputs for a new instance (Vovk et al., 2005). For the users, the size of this prediction set is intended to reflect the algorithm’s uncertainty. This approach requires the assumption that the calibration data and new observations are exchangeable, (*i.e.*, invariant by permutations) to get a prescribed coverage. In classification tasks, the conformal set usually comprises a subset of labels that, with a high degree of prescribed confidence, includes the true label of the new observation (see Angelopoulos et al., 2023; Fontana et al., 2023 for overviews of conformal prediction and Chzhen et al., 2021 for set valued classification).

In this work, we focus on a setting where each input to be classified (at testing) consists of multiple observations, each of which can be independently processed prior to aggregation. This methodology is commonly employed in machine learning-driven citizen science applications, such as Pl@ntNet

(Joly et al., 2014), where users are prompted to submit one or more images of a plant for identification. Similar multi-input samples are also prevalent in iNaturalist (Van Horn et al., 2018) or in eBird (Sullivan et al., 2009), the latter managing multi-input sounds of birds, rather than images. Although multiple images may be submitted, algorithms often process them independently and only aggregate the resulting (softmax) predictions. A naive conformal prediction step based on such aggregated scores severely disrupts the exchangeability. This results in excessively large prediction sets, rendering them impractical (see Appendix B).

Simultaneously, we aim at class conditional coverage calling which would need a sufficiently large number of calibration examples for each class to reliably estimate how aggregation behaves. In datasets with a highly imbalanced class distribution, such as those commonly found in citizen science datasets with many underrepresented classes (see, for example, Garcin et al., 2021), achieving this requirement is not feasible.

To overcome this limitation, we propose to perform conformal prediction at the level of individual observations, and to aggregate the resulting conformal prediction sets. We frame this problem as one of p-value aggregation, where the p-values are the conformal p-values computed for each individual observation. Using recent results on the distribution of these p-values, we develop a general framework for aggregating them using score functions, which encompasses classical techniques such as p-value correction, majority voting, or template matching.

Setting. We consider the scenario of multi-class classification with K classes. We have at disposal an already trained classifier $\mathcal{A} : \mathcal{X} \rightarrow [0, 1]^K$, a *calibration set* of n labeled points $\{(X_i, Y_i)\}_{i=1}^n$, and multiple observations $\{X_{n+j}\}_{j=1}^m$ of the *same* item from the (unknown) class Y_{n+1} . Our objective is to build a set $\mathcal{C}_\alpha$ depending on the classifier, the calibration set, and the new points satisfying for a fixed $\alpha \in (0, 1)$ and for all $y \in [K]$:

$$\mathbb{P}\left[y \in \mathcal{C}_\alpha\left(\mathcal{A}, \{(X_i, Y_i)\}_{i=1}^n, \{X_{n+j}\}_{j=1}^m\right) \middle| Y_{n+1} = y\right] \geq 1 - \alpha . \quad (1)$$

To get an informative set, $\mathcal{C}_\alpha$ is intended to be as small as possible.

Notation. For two integers $n \leq m$, we denote $[n : m] := \{n, \dots, m\}$ and $[n] := [1 : n]$. We set $s : \mathcal{X} \times [K] \rightarrow \mathbb{R}$ to be a score function and $S_i = s(X_i, Y_i)$ for $i \in [n]$ are the calibration scores. For a label $y \in [K]$, we denote $n_y = |\{i \in [n] : Y_i = y\}|$ the number of calibration points of class y (supposed fixed in most of our results) and $S_1^y, \dots, S_{n_y}^y$ the scores of these points. We denote $S_{n+i}(y) := s(X_{n+i}, y)$ the score of a label y at the test point X_{n+i} for $i \in [m]$.

Contributions. We develop a framework for constructing a conformal prediction set that is class conditionally valid (1), by aggregating multiple observations from the same class. This approach can be viewed as stemming from tests that assess the exchangeability between the observations and the points within each class. It has been made possible through the adaptation of recent findings on the joint distribution of conformal p-values, as discussed (Gazin et al., 2023). Our contributions are summarized below:

1. The prediction set is built from class conditional conformal p-values for which we give the exact distribution. Notably, our construction explicitly handles *exchangeability* and, in particular, *ties* between the conformal scores—a frequent occurrence with softmax outputs from neural networks, yet often overlooked in the literature. This probabilistic result serves as a key ingredient in the contributions that follow (Section 3).
2. We relate our method to the prediction set aggregation framework of Gasparin and Ramdas (2024), and propose an enhancement tailored to the setting where the aggregated sets are class-conditional conformal sets constructed from a shared calibration set (Section 4).
3. We propose a new efficient way to aggregate class-conditional conformal p-values by constructing a region of reject based on the construction of a score v on the p-values. (Section 5).
4. We test our methods on a classification on a synthetic dataset of mixture of Gaussian distributions and on LifeCLEF Plant Identification Task 2015 (Goëau et al., 2015). In all the experiments, the classifiers are neural networks with softmax output (Section 6).

2 Background and related works

2.1 Overview on classical conformal prediction

Conformal Prediction (CP) is a framework introduced by [Vovk et al. \(2005\)](#) to construct distribution-free prediction sets quantifying the uncertainty in the predictions of an algorithm. In this context, $\alpha \in (0, 1)$ represents the significance level, and $1 - \alpha$ denotes the targeted coverage probability. Formally, given a calibration set $\{(X_i, Y_i)\}_{i \in \llbracket n \rrbracket}$ of points in $\mathcal{X} \times \llbracket K \rrbracket$ and a trained classifier, a CP method constructs for a new point $X_{n+1} \in \mathcal{X}$ a *marginally valid* set $\mathcal{C}_\alpha$, i.e., a set containing the unobserved outcome $Y_{n+1} \in \llbracket K \rrbracket$ with high probability:

$$\mathbb{P}[Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1})] \geq 1 - \alpha . \quad (2)$$

The idea behind CP methods is to first build non-conformity scores $S_i := s(X_i, Y_i)$ on the calibration set, for a given score function $s : \mathcal{X} \times \llbracket K \rrbracket \rightarrow \mathbb{R}$ intended to quantify the error of the algorithm at a given point. Many score functions have been considered to catch different type of information, see e.g., [Angelopoulos et al. \(2023\)](#) for a review. A common choice in classification is to use (one minus) the estimated probability of the class (e.g., the softmax output of neural network). In this paper, we will focus on the case where an already trained algorithm is provided, which corresponds to the case of split CP. Then, for a score function s , the prediction set

$$\mathcal{C}_\alpha(x) := \left\{ y \in \llbracket K \rrbracket : s(x, y) \leq S_{(\lceil (1-\alpha)(n+1) \rceil)} \right\} , \quad (3)$$

is marginally valid (as defined in Eq. (2)) where $S_{(k)}$ is the k -th smallest score of $S_1, \dots, S_n, \infty$ ([Vovk et al., 2005](#)).

In certain applications, it is desirable to construct prediction sets that satisfy a *class conditional* coverage guarantee; that is, for every class y , the following condition holds:

$$\mathbb{P}[y \in \mathcal{C}_\alpha(X_{n+1}) \mid Y_{n+1} = y] \geq 1 - \alpha . \quad (4)$$

While this condition implies marginal coverage, the converse does not hold. In scenarios with highly imbalanced or heterogeneous classes, marginal split CP can lead to prediction sets with significant variability in class-wise coverage ([Sadinle et al., 2019](#); [Ding et al., 2023](#)). To enforce conditional coverage as specified in (4), a natural approach is to apply split CP separately for each class ([Vovk et al., 2005](#)). The resulting prediction set is given by

$$\mathcal{C}_\alpha^{cd}(x) = \left\{ y : s(x, y) \leq S_{(\lceil (1-\alpha)(n_y+1) \rceil)}^y \right\} , \quad (5)$$

which is conditionally valid, with $\{S_i^y\}_{i \in \llbracket n_y \rrbracket}$ being the scores of the n_y points with label y in the calibration set. A limit of this approach is that it requires to have at least $\frac{1}{\alpha} - 1$ points for each class ([Ding et al., 2023](#)). Otherwise, such a class would always be predicted which can lead to uninformative prediction set.

2.2 P-value point of view of conformal prediction

The conformal set (3) can be interpreted as the outcome of a family of two-sample tests, each assessing—for every $y \in \llbracket K \rrbracket$ —whether $s(X_{n+1}, y)$ is exchangeable with the scores of the calibration. The predicted classes $y \in \mathcal{C}_\alpha$ are those not rejected by a Wilcoxon-Mann-Whitney test (between a sample of size 1 and of size n , [Wilcoxon, 1945](#); [Mann and Whitney, 1947](#)). One can associate a so-called split conformal p-value ([Vovk et al., 2005](#)) to split CP, defined as

$$p_{n+1}^{mrg}(y) = \frac{1}{n+1} \left[\sum_{i=1}^n \mathbf{1}\{s(X_{n+1}, y) \leq S_i\} + 1 \right] , \quad (6)$$

such that the condition $s(X_{n+1}, y) \leq S_{(\lceil (1-\alpha)(n+1) \rceil)}$ in Eq. (3) is equivalent to $p_{n+1}(y) \geq \alpha$. For the true label Y_{n+1} , assuming the scores have *no ties*, the p-value of the true label $p(Y_{n+1})$ follows a uniform distribution over $\{i/(n+1) : i \in \llbracket n+1 \rrbracket\}$. This distribution is explicit and independent of the underlying data distribution, which allows for effective identification of unlikely candidate labels. The construction of this distribution-free quantity forms the basis of conformal prediction (see [Barber and Tibshirani, 2025](#) for a recent p-value point of view).

Similarly, for class-conditional prediction sets (5), a class-conditional p-value can be defined as $p_{n+1}^{cd}(y) = \frac{1}{n_y+1} [\sum_{i=1}^{n_y} \mathbf{1}\{s(X_{n+1}, y) \leq S_i^y\} + 1]$. As in the marginal case, this set can be interpreted as arising from a test of exchangeability between the test score and the calibration scores for the corresponding class.

2.3 Joint control of the p-values.

In our setting, where multiple images $X_{n+1}, \dots, X_{n+m}$ with the same label Y_{n+1} are observed, our approach relies on constructing vectors of conformal p-values $\mathbf{p}(y) = (p_1(y), \dots, p_m(y))$, where $p_j(y)$ is the class-conditional p-value associated with X_{n+j} for class y . One question is to find a set $E \subset [0, 1]^m$ that contains the vector $\mathbf{p}(Y_{n+1})$ with high probability:

$$\mathbb{P}[\mathbf{p}(Y_{n+1}) \in E] \geq 1 - \alpha. \quad (7)$$

For $m = 1$, conformal prediction chooses $E = [\alpha, 1]$, but, for m larger, a wide range of choices become possible. This question has been extensively studied across various contexts and for multiple purposes. In the realm of multiple testing, numerous methods have been developed to control the false discovery proportion (FDP), often leading to criteria interpretable as statistical envelope: For instance Bonferroni (1936); Rüger (1978); Rüschendorf (1982); Vovk et al. (2022) address scenarios involving arbitrarily dependent p-values.; Fisher (1925); Pearson (1934); Simes (1986); Benjamini and Hochberg (1995) focus on settings with independent p-values. Additionally, Sarkar (1998); Benjamini and Yekutieli (2001) provide results under positive dependence assumptions while the exchangeable case has been recently considered by Gasparin et al. (2025). Blanchard et al. (2020) (see also Li et al., 2024) introduce the notion of *template*, which most closely aligns with the envelope perspective we adopt. In the framework of conformal prediction, controls of the form (7) have been employed across various applications, including outlier detection using conformal scores (Bates et al., 2023; Gazin et al., 2023), batch prediction (Gazin et al., 2024; Lee et al., 2024), aggregation of prediction intervals (Gasparin and Ramdas, 2024), and conformal prediction for ranking (Fermanian et al., 2025). All these methods can be summarized as constructing a score function v of the p-values such that $\mathbf{p} \in E \iff v(\mathbf{p}) \geq q_E$, where q_E is a threshold specific to each method. Table 1 provides an overview of these methods, the associated score functions v , and a concise summary of the required dependencies among the p-values to ensure the guarantee (7).

	Set $E \subset [0, 1]^m$	$v(\mathbf{p})$	Dependence
Bonferroni	$\{\mathbf{p} : \forall j, p_j \geq \alpha/m\}$	$\min p_j$	–
Simes	$\{\mathbf{p} : \forall j, p_{(j)} \geq j\alpha/m\}$	$\min_j m p_{(j)}/j$	PRDS
Template	$\{\mathbf{p} : \forall j, p_{(j)} \geq t_j(\lambda_\alpha)\}$	$\min_j t_j^{-1}(p_{(j)})$	–
Majority vote	$\{\mathbf{p} : \sum_{j=1}^m \mathbf{1}_{\{p_j \geq \alpha/2\}} \geq m/2\}$	$\hat{F}_m^{-1}(1/2)$	–
	$\{\mathbf{p} : \forall k, \sum_{j=1}^k \mathbf{1}_{\{p_j \geq \alpha/2\}} \geq k/2\}$	$\min_k \hat{F}_k^{-1}(1/2)$	Exchangeability

Table 1: Examples of sets satisfying (7) and the associated score function. $\hat{F}_k$ denotes the empirical CDF of $(p_1, \dots, p_k)$. Template method refers to Blanchard et al. (2020) where t_j denotes a family of functions and $\lambda_\alpha > 0$ is a parameter. Majority vote refers to the methods analyzed by Gasparin and Ramdas (2024). PRDS denotes Positive Regression Dependence on a Subset (Benjamini and Yekutieli, 2001).

3 Joint distribution of the conditional p-values

The present section introduces fundamental mathematical tools that will be required to derive conformal predictors (based on aggregation or on p-values) with the right validity in Section 4 and Section 5. We first present some key lemma about a uniform distribution over the ranks. We then link this distribution to the joint distribution of conditional p-values associated to each observation X_{n+i} . Finally, we derive the distribution of some key statistics.

Key lemma. Let $A_{n,m}$ for $n, m \in \mathbb{N}^*$, be the set of ordered discrete p-values:

$$A_{n,m} := \left\{ a \in \left\{ \frac{i}{n} : 0 \leq i \leq n \right\}^m : a_1 \leq \dots \leq a_m \right\}. \quad (8)$$

Lemma 3.1. Let $n, m \in \mathbb{N}^*$ and $P \sim \mathcal{U}(A_{n,m})$, where $\mathcal{U}$ denotes the uniform distribution. Then

- $n \cdot P(j) \sim \text{BetaBin}(n, j, m - j + 1)$ for $j \in \llbracket m \rrbracket$.
- $\sum_{j=1}^m \mathbf{1}_{\{n \cdot P(j) \geq \ell\}} \sim \text{BetaBin}(m, n + 1 - \ell, \ell)$ for $\ell \in \llbracket 0 : n \rrbracket$.

Remark 3.2. The beta-binomial distribution $\text{BetaBin}(m, a, b)$ is the distribution of a binomial random variable $\mathcal{B}(m, Q)$ where Q is drawn independently as a beta distribution $B(a, b)$. If a and b are positive integers, then for $k \in \llbracket 0 : m \rrbracket$,

$$\mathbb{P}[\text{BetaBin}(m, a, b) = k] = \frac{\binom{k+a-1}{k} \binom{m-k+b-1}{m-k}}{\binom{m-1+a+b}{m}}.$$

In this case, the distribution $\text{BetaBin}(m, a, b)$ is a negative hypergeometric distribution which can be easily simulated using Polya's urns. For $a = 0$ (resp. $b = 0$), the distribution is defined as a Dirac in 0 (resp. in m).

Joint distribution of the conditional p-values. We focus here on the joint distribution of the ordered vector of p-values $\mathbf{p}_\uparrow(y) = (p_{(1)}(y), \dots, p_{(m)}(y))$, where for $j \in \llbracket m \rrbracket$ and $y \in \llbracket K \rrbracket$

$$p_j(y) = \frac{1}{n_y} \sum_{i=1}^{n_y} \mathbf{1}\{s(X_{n+j}, y) \leq S_i^y\}. \quad (9)$$

The guarantees we obtain on this vector are under the condition that the scores of the new points evaluated in the true label are exchangeable with the scores of the same class:

Assumption 3.3. For all $y \in \llbracket K \rrbracket$ and conditionally to $Y_{n+1} = y$, the vector of scores $(S_1^y, \dots, S_{n_y}^y, S_{n+1}(y), \dots, S_{n+m}(y))$ is exchangeable.

This hypothesis is fundamental to our work. It assumes that the scores of the observed images behave similarly to the calibration scores of the same class. This assumption is more discussed in Section E.

Remark 3.4. The number of multi-inputs m is formally assumed to be fixed throughout the paper. However, the method remains directly applicable when m varies. In that case, if Assumption 3.3 holds conditionally on m , then our results remain valid conditionally on m .

The distribution of the unordered vector of p-values has been deeply studied by [Gazin et al. \(2023\)](#); [Gazin \(2024\)](#) in the general case where the p-values are the marginal ones, the test points do not all share the same class, and the scores have no ties. The following proposition can be seen as an extension of these results conditionally to the class, which includes the possibility of *ties* between scores. To this purpose, we introduce a randomized version $\mathbf{p}_\uparrow^{\text{rd}}$ of the vector of conformal p-values $\mathbf{p}$.

Theorem 3.5. Under Assumption 3.3, for $y \in \llbracket K \rrbracket$ and $j \in \llbracket m \rrbracket$, let us introduce

$$p_j^{\text{rd}}(y) = \frac{1}{n_y} \sum_{i=1}^{n_y} \left[\mathbf{1}\{S_{n+j}(y) < S_i^y\} + \mathbf{1}\{S_{n+j}(y) = S_i^y\} \mathbf{1}\{U_j \leq U_i^y\} \right], \quad (10)$$

where $(U_i)_{i \in \llbracket m \rrbracket}$ and $(U_i^y)_{i \in \llbracket n_y \rrbracket}$ for $y \in \llbracket K \rrbracket$ are i.i.d. uniform random variables on $[0, 1]$. Then the ordered vector of randomized p-values $\mathbf{p}_\uparrow^{\text{rd}}(y) = (p_{(1)}^{\text{rd}}(y), \dots, p_{(m)}^{\text{rd}}(y))$, conditionally to $Y_{n+1} = y$, is uniformly drawn from $A_{n_y, m}$:

$$\mathbf{p}_\uparrow^{\text{rd}}(Y_{n+1}) \mid (Y_{n+1} = y) \sim \mathcal{U}(A_{n_y, m}). \quad (11)$$

Moreover, for $j \in \llbracket m \rrbracket$, we have $n_y \cdot p_{(j)}^{\text{rd}}(Y_{n+1}) \mid (Y_{n+1} = y) \sim \text{BetaBin}(n_y, j, m - j + 1)$.

The proof is given in Appendix A.3. As a minor remark, notice that when scores have no ties, the original $\mathbf{p}$ is preserved. More important, the advantage of the above result is that the knowledge of the exact distribution of the conformal p-values will allow us to estimate more precisely the quantiles of statistics constructed from this vector. As a consequence, the resulting prediction sets will be more informative (smaller sets) than those based on bounds for these quantiles.

Remark 3.6. The definition (10) of the conformal p-values differs a bit from the usual definition (different normalization and no additional +1, can be compared with (6)) but provides an equivalent tool. With this definition, the ordered vector of p-values has a symmetric distribution around the diagonal: $1 - (p_{(m)}(Y_{n+1}), \dots, p_{(1)}(Y_{n+1}))$ has the same distribution as $\mathbf{p}_\uparrow^{\text{rd}}(Y_{n+1})$.

Other important law derivations. Combining this first result with Lemma 3.1, we can derive the distribution of critical quantities such as the empirical coverage or the marginal distribution of the ordered vector.

Corollary 3.7. *Under Assumption 3.3, let $B_y = \sum_{j=1}^m \mathbf{1}\{y \in \mathcal{C}_\alpha^{cd}(X_{n+j})\}$ be the number of class conditional sets containing the label y . Then for $y \in \llbracket K \rrbracket$ and $\ell \in \llbracket m \rrbracket$:*

$$\mathbb{P}[B_{Y_{n+1}} = \ell | Y_{n+1} = y] \geq \mathbb{P}[\beta = \ell], \quad (12)$$

where β follows the distribution $\text{BetaBin}(m, (n_y + 1) - k_{y,\alpha}, k_{y,\alpha})$ with $k_{y,\alpha} = \lfloor (n_y + 1)\alpha \rfloor$. Moreover, if the scores have no ties, Eq. (12) is an equality.

The distribution of empirical coverage for conformal prediction sets has been first studied by Vovk (2012) and Angelopoulos et al. (2023). For the exchangeable case with no ties, we will refer to Marques F. (2025) who obtains a similar result. This result will be used to enhance the aggregation of conformal prediction sets by majority voting in our setting.

4 Refinement of majority voting for multiple inputs

Thanks to these probabilistic results, we are in position to build new conformal predictors tailored for scenarios where multiple predictive observations of a single instance are available. As this problem can be seen as the aggregation of the prediction set associated to each instance, we focus in this section to the adaptation to our setting of the general majority voting methods proposed by Gasparin and Ramdas (2024). They propose to keep the labels appearing in at least one half of the sets, in our setting it consists to consider the *majority voting* set $\mathcal{C}_{\alpha,m}^M = \{y : \sum_{j=1}^m \mathbf{1}\{y \in \mathcal{C}_\alpha^{cd}(X_{n+j})\} \geq m/2\}$. These authors also propose *exchangeable majority* voting sets defined as $\mathcal{C}^{Me} = \cap_{k=1}^m \mathcal{C}_{\alpha,k}^M$ usable in our setting as the sets $\mathcal{C}_\alpha^{cd}(X_{n+j})$ are exchangeable. Both methods achieve a class-conditional coverage of $1 - 2\alpha$, replacing α by $\alpha/2$ leads to the correct coverage (see Appendix C).

If $\mathcal{C}_j = \mathcal{C}_\alpha^{cd}(X_{n+j})$, then thanks to Proposition 3.7, the distribution of the number of sets containing the true label Y_{n+1} is known. The threshold $m/2$ chosen in the majority vote method can then be improved: we replace it by the quantile of a Beta-Binomial distribution.

Proposition 4.1. *Let $\alpha \in (0, 1)$, assume Assumption 3.3, let*

$$\mathcal{C}_\alpha^{BB} := \left\{ y : \sum_{i=1}^m \mathbf{1}\{y \in \mathcal{C}_\alpha^{cd}(X_{n+i})\} \geq q_\alpha^{BB}(y) \right\}, \quad (13)$$

where $q_\alpha^{BB}(y)$ is the α -quantile of the distribution $\text{BetaBin}(m, \lceil (n_y + 1)(1 - \alpha) \rceil, \lfloor (n_y + 1)\alpha \rfloor)$. Then, for all $y \in \llbracket K \rrbracket$:

$$\mathbb{P}[Y_{n+1} \in \mathcal{C}_\alpha^{BB} | Y_{n+1} = y] \geq 1 - \alpha.$$

For each class, the set $\mathcal{C}_\alpha^{BB}$ requires computing or storing quantiles of a Beta-Binomial distributions, which depends on the number of calibration points belonging to that class. To avoid performing this computation for each class individually, for example if the number of classes is too large, one can approximate the Beta-Binomial distribution by a Binomial distribution (see Lemma F.1). This approximation induces a loss in coverage as presented in the following proposition.

Proposition 4.2. *Let $\alpha \in (0, 1)$, assume Assumption 3.3 holds, let*

$$\mathcal{C}_\alpha^{Bin} := \left\{ y : \sum_{j=1}^m \mathbf{1}\{y \in \mathcal{C}_\alpha^{cd}(X_{n+j})\} \geq q_\alpha^{Bin} \right\}, \quad (14)$$

where q_α^{Bin} is the α -quantile of the binomial distribution $\mathcal{B}(m, 1 - \alpha)$. Then for all $y \in \llbracket K \rrbracket$:

$$\mathbb{P}[Y_{n+1} \in \mathcal{C}_\alpha^{Bin} | Y_{n+1} = y, n_y] \geq 1 - \alpha - \frac{(m-1) \min(1, \alpha m)}{n_y + 1}, \quad a.s. \quad (15)$$

Moreover, if the labels $Y_1, \dots, Y_{n+1}$ are i.i.d., then

$$\mathbb{P}[Y_{n+1} \in \mathcal{C}_\alpha^{Bin}] \geq 1 - \alpha - \frac{K(m-1) \min(1, \alpha m)}{n+1}. \quad (16)$$

The loss of coverage remains negligible when the number of repetitions is small ($m \ll n/K$ for the marginal one). In the setting of Pl@ntNet data, the number of repetitions is effectively low, but some classes have also a small numbers of examples in the calibration set which makes this approximation only usable for large classes or when only a marginal coverage is targeted.

Remark 4.3. The possibility to use a binomial distribution quantile has been already proposed by [Gasparin and Ramdas \(2024\)](#) to merge independent prediction sets. The assumption of independence is a bit restrictive and this result proves that it can effectively be used for prediction sets correlated by the calibration (so not independent) for small enough number of sets m .

5 p-value aggregation methods

Voting methods presented above aggregate the conformal sets constructed for each observation point by only considering the number of sets that contain a given label. Although this procedure can be interpreted as an aggregation of the conformal p-values and as forming an envelope around them (see Figure 1), we observe, however, that some information is lost in the process. In this section, we exploit the structure of the ordered vector of p-values to aggregate the observations. These methods decisively outperform the previously discussed voting strategies.

5.1 General method

The method consists in building a (non-)conformity score function $v : [0, 1]^m \times \llbracket K \rrbracket \rightarrow \mathbb{R}$ directly on the p-values vector. Then, exploiting the result in Theorem 3.5, we calibrate this new score by simulating samples $P_1^y, \dots, P_T^y$ of the distribution $\mathcal{U}(A_{n_y, m})$ for each class, using for example Algorithm 1. Here, T can be seen as a budget for approximating the score's distribution by Monte Carlo sampling. The method is summarized in the following result which gives a guarantee on the conditional coverage.

Proposition 5.1. *Let $v : [0, 1]^m \times \llbracket K \rrbracket \rightarrow \mathbb{R}$ and $T \in \mathbb{N}$, for $y \in \llbracket K \rrbracket$ let $P_t^y \stackrel{i.i.d.}{\sim} \mathcal{U}(A_{n_y, m})$ and $V_t^y := v(P_t^y, y)$ for $t \in \llbracket T \rrbracket$. Then, for $\alpha \in (0, 1)$, if Assumption 3.3 is satisfied, the set*

$$\mathcal{C}_\alpha^v = \left\{ y : v(\mathbf{p}_\uparrow^{\text{rd}}(y), y) \geq V_{(\lfloor (T+1)\alpha \rfloor)}^y \right\} \quad (17)$$

is conditionally valid: $\mathbb{P}[y \in \mathcal{C}_\alpha^v | Y_{n+1} = y] \geq 1 - \alpha$ for all $y \in \llbracket K \rrbracket$.

This result is a direct application of conformal prediction using the score function v , and follows as a consequence of Theorem 3.5. While the choice of the score function v is discussed below, it is worth noting that for a given v , we retain the low scores in order to remain within the standard framework of envelope methods: as shown in Table 1, the scores associated with envelopes tend to reject p-values corresponding to low scores.

Remark 5.2. Recall that the guarantees provided by conformal prediction in (2) and (4) hold with high probability, including with respect to the calibration set. Formally, the sample $(P_t^y)_{t,y}$ should be redrawn at each execution. However, it is also possible to fix T sufficiently large and store the value $V_{(\lfloor (T+1)\alpha \rfloor)}^y$ for each class, interpreting it as a quantile estimate of the score distribution.

5.2 Construction of score functions v

We propose in the following different envelopes - associated to a score function $v : [0, 1]^m \rightarrow \mathbb{R}$ on the p-values—that can either be used for the vector of p-values $\mathbf{p}$ or its randomized version $\mathbf{p}_\uparrow^{\text{rd}}$ as in the above result.

Quantile envelope. The first envelope that we consider relies on quantiles of the beta-binomial distribution. For a clear presentation, let us first remark that as presented in Theorem 3.5, the marginal distributions of the sorted vector of p-values $\mathbf{p}_\uparrow^{\text{rd}}(Y_{n+1})$ are known. Indeed, conditionally to $Y_{n+1} = y$ and n_y , the random variable $n_y \cdot p_j^{\text{rd}}(y)$ follows the distribution $\text{BetaBin}(n_y, j, n_y - j + 1)$.

To capture the geometry of the trajectory of p-values, we propose to choose an envelope of the trajectories constructed as the vector of quantiles of level λ of each marginal (see Figure 1). Consider $\{F_j\}_j$ a general family of cdf and let us remark that

$$\{\mathbf{p} : \forall j : p_{(j)} \geq F_j^{(-1)}(\lambda)\} = \{\mathbf{p} : \min_j F_j(p_{(j)}) \geq \lambda\}.$$

The envelope of quantile of marginal associated to the *quantile* score v_Q is defined as

$$v_Q(\mathbf{p}, y) := \min_{j \in \llbracket m \rrbracket} F_{j, n_y}(n_y \cdot p_{(j)}(y)), \quad \text{for } \mathbf{p} \in [0, 1]^m, y \in \llbracket K \rrbracket, \quad (18)$$

where $F_{j, n}$ is the cdf of the distribution $\text{BetaBin}(n, j, m - j + 1)$. This kind of envelope, constructed as the true quantile of the marginal with a specific level, appears for instance in [Genovese and Wasserman \(2006\)](#); [Blanchard et al. \(2020\)](#) and in [Fermanian et al. \(2025\)](#) in an empirical version.

ℓ^q -Area envelope. A simple idea to reject the classes whose p-values are too small could be supported by the fact that the area of the trajectory is too small. For $q > 0$, we define the *area* score: $v_{q\text{-Area}}(\mathbf{p}, y) := \|\mathbf{p}\|_q$, where $\|\cdot\|_q$ for $q \geq 1$ is the ℓ_q -norm in $\mathbb{R}^m$. For $q = 1$, the procedure that consists in rejecting a trajectory or not can be viewed as a variant of the Wilcoxon-Mann-Whitney test ([Wilcoxon, 1945](#); [Mann and Whitney, 1947](#)).

ℓ^q -envelope. As asymptotically, the vector $\mathbf{p}_{\uparrow}^{\text{rd}}(Y_{n+1})$ converges to the identity vector $\text{Id} = (j/(m+1))_{j=1}^m$, our last score only considers

the distance of p-value vector to this vector. In particular, $\mathbb{E}[p_{(j)}^{\text{rd}}(Y_{n+1})] = \frac{j}{m+1}$ (cf. Theorem 3.5) which motivates the score v_{ℓ^q} defined as: $v_{\ell^q}(\mathbf{p}, y) := -\|\mathbf{p} - \text{Id}\|_q$. The resulting set (17) will then keep p-values close to Id . For $q = \infty$, bounds of the exact quantile and the asymptotic distribution are provided respectively by [Gazin et al. \(2023\)](#) and [Gazin \(2024\)](#).

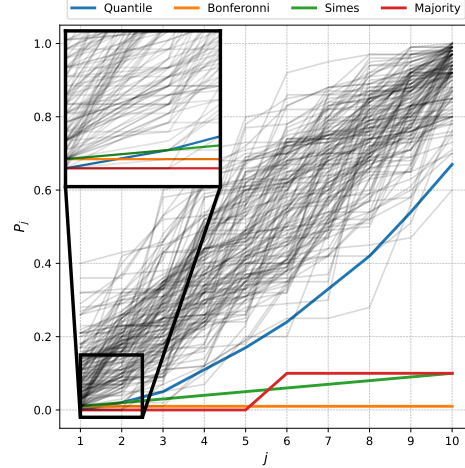


Figure 1: Simulation of 150 samples of $P \sim \mathcal{U}(A_{n,m})$ for $m = 10$ and $n = 100$ (black lines). Quantile is the envelope associated to v_Q , and Majority to the majority vote.

6 Experiments

A key requirement of a conformal approach is its ability to accurately quantify uncertainty. For an efficient classifier, this should lead to small prediction sets; for a less accurate one, the conformal set should still contain the correct label, at the cost of a larger set. To this purpose, we compare the methods, by comparing the size of the prediction sets returned for a fixed coverage.

Multiple observations of the same instance are particularly valuable in challenging tasks where a single observation does not permit informative classification. As we will demonstrate, our aggregation techniques lead to a reduction in prediction set size in such cases resulting in a gain of interpretability. Throughout our experiments, we report the accuracy of the base classifiers. The score used for all the following experiments is the *softmax score* which is one minus the softmax output of the class. Additional results using the APS score ([Romano et al., 2020](#)) are presented in Section D.2.

The tested prediction sets are the *Majority* $\mathcal{C}^M$ —Eq. (22), *Exchangeable Majority* $\mathcal{C}^{Me}$ —Eq. (23), *BetaBinomial* $\mathcal{C}^{BB}$ —Eq. (13), *Binomial* $\mathcal{C}^{Bin}$ —Eq. (14), and the set $\mathcal{C}^v$ —Eq. (17) with the score functions v_Q (*Quantile*), $v_{1\text{-Area}}$ (*Wilcoxon*), $v_{2\text{-Area}}$ (ℓ_2 Area) and v_{ℓ_2} (ℓ_2).

Code for reproducing our experiments is available at https://github.com/jeanbaptistefermanian/Class_Conditional_CP_for_Multi_Inputs.

6.1 Synthetic data

We choose as distribution of the points, a mixture of Gaussian distributions with randomly drawn centers. Formally, $Y \sim \sum_{y=1}^K p(y)\delta_y$ and $X|(Y = y) \sim \mathcal{N}(\mu_y, \sigma^2 I_d)$, where $K = 10$, $d = 6$, the

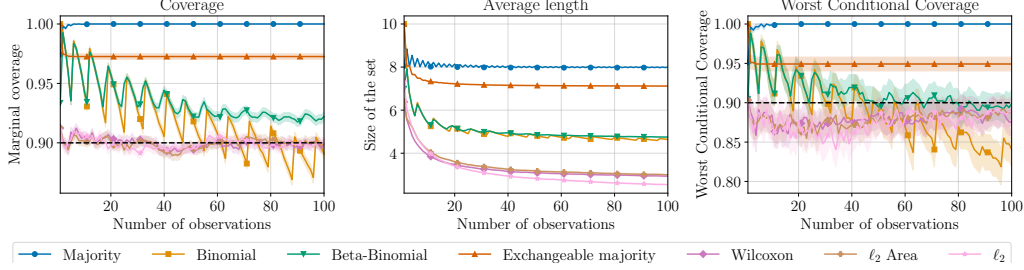


Figure 2: Synthetic data: marginal coverage, average length and minimum of the class conditional coverages in function of the number of observations for $\alpha = 0.1$, computed on 5000 repetitions. The confidence region is $\pm (\text{empirical std})/\sqrt{5000}$.

classes are a bit imbalanced $p(y) \propto (1 + 3(y - 1)/K)^{-1}$ and $\sigma^2 = 3$. The centers of the classes μ_y are drawn independently as $\mathcal{N}(0, I_d)$. The strength of the noise makes this problem difficult with only one observation as the distance between two centers clusters is relatively close to the distance of a point to its cluster’s center : $\mathbb{E}[\|X - \mu_Y\|] \simeq \mathbb{E}[\|\mu_y - \mu_z\|]$. The model is a ReLu neural network with three hidden layers of width 15. It is trained with 2000 observations and achieves a top 1 accuracy of 0.49. We apply our conformal procedure with a calibration set of size $n = 1000$ and repeat 5000 times our procedure for a number of observations $m \in \{1, \dots, 100\}$.

We observe in Figure 2 that the methods based on the p-values perform best as compared to majority vote methods. For instance, for $m = 20$ and ℓ_2 -envelope method, we observe a significant improvement with a set of average size of 3.6 as compared to an original value of 6.6. They provide the smallest sets while maintaining a marginal coverage above $1 - \alpha$, and exhibit conditional coverage close to $1 - \alpha$. The majority vote methods appear to be overly conservative. The proposed refinement of the majority vote improves on the two original versions. As announced in Proposition 4.2, for small m , the binomial quantile approximates the beta-binomial one well. The marginal coverage is not significantly affected as long as m/n remains small. However, in the third figure, the conditional coverage for some classes is impacted by this approximation when m is large as the numbers of points of some classes in the calibration set are smaller than m . The over-coverage observed with the Beta-Binomial approach arises from the potential non-existence of a quantile at the exact level $1 - \alpha$. This issue can be addressed by introducing additional randomization.

6.2 The LifeCLEF Plant Identification Task 2015

We apply our method to LifeCLEF Plant Identification Task 2015¹ (Goëau et al., 2015) dataset, a dataset of 113,205 images of plants spread between 1K species (classes), with high heterogeneity. We split equally between train, calibration and test sets the original full dataset thanks to a stratified approach to maintain the original classes distribution. The classifier considered is ResNet50, trained on Pl@ntNet-300K Garcin et al. (2021)², for which we have fine tuned the last layer on our training set. It achieves a top-1 accuracy of 43% (and top-5 of 63%, see Section D.1 for others top-k accuracy). As a pre-processing step, we restrict ourselves to the classes (species) for which we have at least 20 examples in the calibration set, hence keeping $K = 688$ classes. This is needed for having informative class conditional CP set. We group randomly images of the same classes to simulate multi-view observations of the same plant with a number of repetitions of $m \in \{1, \dots, 10\}$.

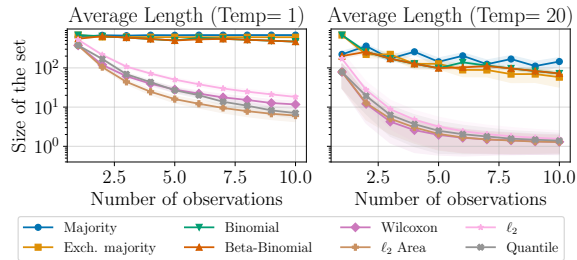


Figure 3: LifeCLEF Plant Identification Task 2015: average length in log-scale for $\alpha = 0.1$ and two different choice of temperature in function of the number of observations.

¹<https://www.imageclef.org/lifeclef/2015/plant>

²see code and models at <https://github.com/plantnet/PlantNet-300K/>

In Figure 3, the prediction set lengths of various methods for two temperature settings (Temp=1 and Temp=20), all methods achieve a valid coverage of at least 0.9, with detailed results deferred to Section D.2. The temperature refers to the standard re-scaling of the output of the network before the application of the softmax function. Our interpretation is that, due to the numerical approximation of the softmax function, a low temperature can lead to *ties* between scores (many are then numerically 0 or 1), which may hinder the performance of conformal procedures (see, *e.g.*, Dabah and Tirer, 2024 for a discussion on the relationship between temperature scaling and conformal prediction). Specifically, we observe that when using the raw network outputs without temperature adjustment (Temp=1), the majority vote struggles and tends to predict nearly all classes. In contrast, methods based on conformal p-values consistently produce narrower prediction sets after just a few observations and are less sensitive to this parameter. Our refinements to the majority vote method do lead to improvements over the original approach, but they are still outperformed by the p-value aggregation methods.

7 Conclusion and limitations

We introduced a new class-conformal prediction framework for the problem of classification when multiple observations of the same instance at prediction time are available. Assuming the scores of the new points evaluated in the true label are exchangeable with those of the same class, we developed an analysis of the distribution of the vector of the underlying ordered conformal p-values. We built two novel aggregation strategies of these p-values and showed how effective is the method that captures the best the structure of the conformal p-values.

Exchangeability of scores vector seems to be the bedrock of our strategy and more generally of conformal prediction (see Section E). We ensured this condition by sampling the multi-inputs randomly from the same class. It would be interesting to investigate the more realistic case of dependent multi-inputs, for which our method constitutes a foundational step.

Acknowledgments and Disclosure of Funding

The authors acknowledge Ulysse Gazin, Pierre Humbert and Etienne Roquain for helpful discussion. This work was funded by the French National Research Agency (ANR) through the grant Chaire IA CaMeLOt (ANR-20-CHIA-0001-01).

References

- Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.
- Rina Foygel Barber and Ryan J Tibshirani. Unifying different theories of conformal prediction. *arXiv preprint arXiv:2504.02292*, 2025.
- Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *Ann. Statist.*, 51(1):149–178, 2023.
- Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57(1):289–300, 1995.
- Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, pages 1165–1188, 2001.
- Gilles Blanchard, Pierre Neuvial, and Etienne Roquain. Post hoc confidence bounds on false positives using reference families. *Ann. Statist.*, 48(3):1281–1303, June 2020. doi: 10.1214/19-AOS1847.
- Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commerciali di firenze*, 8:3–62, 1936.
- Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, and Titouan Lorieul. Set-valued classification—overview via a unified framework. *arXiv preprint arXiv:2102.12318*, 2021.
- Lahav Dabah and Tom Tirer. On temperature scaling and conformal prediction of deep classifiers. *arXiv preprint arXiv:2402.05806*, 2024.

- Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *NeurIPS*, 36:64555–64576, 2023.
- Tiffany Ding, Jean-Baptiste Fermanian, and Joseph Salmon. Conformal prediction for long-tailed classification. *arXiv preprint arXiv:2507.06867*, 2025.
- Jean-Baptiste Fermanian, Pierre Humbert, and Gilles Blanchard. Transductive conformal inference for ranking. *arXiv preprint arXiv:2501.11384*, 2025.
- R.A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.
- Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.
- Camille Garcin, Alexis Joly, Pierre Bonnet, Antoine Affouard, Jean-Christophe Lombardo, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, and Joseph Salmon. Pl@ntnet-300k: a plant image dataset with high label ambiguity and a long-tailed distribution. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.
- Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote. *arXiv preprint arXiv:2401.09379*, 2024.
- Matteo Gasparin, Ruodu Wang, and Aaditya Ramdas. Combining exchangeable p-values. *Proc. Natl. Acad. Sci.*, 122(11):e2410849122, 2025.
- Ulysse Gazin. Asymptotics for conformal inference. *arXiv preprint arXiv:2409.12019*, 2024.
- Ulysse Gazin, Gilles Blanchard, and Étienne Roquain. Transductive conformal inference with adaptive scores. In *AISTATS*, 2023.
- Ulysse Gazin, Ruth Heller, Etienne Roquain, and Aldo Solari. Powerful batch conformal prediction for classification. *arXiv preprint arXiv:2411.02239*, 2024.
- Christopher R Genovese and Larry Wasserman. Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.*, 101(476):1408–1417, 2006.
- Hervé Goëau, Pierre Bonnet, and Alexis Joly. LifeCLEF Plant Identification Task 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation Forum*, volume CEUR Workshop Proceedings of *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, Toulouse, France, 2015.
- Alexis Joly, Hervé Goëau, Pierre Bonnet, Vera Bakić, Julien Barbe, Souheil Selmi, Itheri Yahiaoui, Jennifer Carré, Elise Mouysset, Jean-François Molino, Nozha Boujemaa, and Daniel Barthélémy. Interactive plant identification based on social image data. *Ecological Informatics*, 23:22–34, 2014.
- Yonghoon Lee, Eric Tchetgen Tchetgen, and Edgar Dobriban. Batch predictive inference. *arXiv preprint arXiv:2409.13990*, 2024.
- Jinzhou Li, Marloes H Maathuis, and Jelle J Goeman. Simultaneous false discovery proportion bounds via knockoffs and closed testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 86(4):966–986, 2024.
- Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, pages 50–60, 1947.
- Paulo C. Marques F. Universal distribution of the empirical coverage in split conformal prediction. *Statistics & Probability Letters*, 219:110350, 2025. ISSN 0167-7152.
- Karl Pearson. On a new method of determining "goodness of fit". *Biometrika*, 26(4):425–442, 1934.
- Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *NeurIPS*, 33:3581–3591, 2020.
- Bernhard Rüger. Das maximale signifikanzniveau des tests: "lehne h_0 ab, wenn k unter n gegebenen tests zur ablehnung führen". *Metrika*, 25:171–178, 1978.

- Ludger Rüschendorf. Random variables with maximum sums. *Ann. Appl. Probab.*, 14(3):623–632, 1982.
- Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *J. Amer. Statist. Assoc.*, 114(525):223–234, 2019.
- Sanat K Sarkar. Some probability inequalities for ordered mtp 2 random variables: a proof of the simes conjecture. *Ann. Statist.*, pages 494–504, 1998.
- R John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.
- Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, 142(10):2282–2292, 2009.
- Kanint Teerapabolarn. A bound on the binomial approximation to the beta binomial distribution. In *International Mathematical Forum*, volume 3, pages 1355–1358, 2008.
- Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018.
- Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.
- Vladimir Vovk, Bin Wang, and Ruodu Wang. Admissible ways of merging p-values under arbitrary dependence. *Ann. Statist.*, 50(1):351–375, 2022.
- Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

A Proofs

A.1 Preliminary results

This first result is derived from Theorem A.1 of [Gazin et al. \(2023\)](#).

Corollary A.1. *Let $S_1, \dots, S_{n+m}$ exchangeable random variables with no ties ($S_i \neq S_j$ almost surely). For $j \in \llbracket m \rrbracket$, let*

$$p_j = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{S_{n+j} \geq S_i\}, \quad (19)$$

and $\mathbf{p}_\uparrow = (p_{(1)}, \dots, p_{(m)})$ be the vector of ordered p -values. Then

$$\mathbf{p}_\uparrow \sim \mathcal{U}(A_{n,m}),$$

where $A_{n,m}$ is defined in (8).

Proof. This result is a direct consequence of Theorem A.1 of [Gazin et al. \(2023\)](#). According to this work, let $\mathbf{M} \in \llbracket 0, m \rrbracket^n$ be the histogram of $\mathbf{p}$, i.e., for $j \in \llbracket 0, n \rrbracket$:

$$\mathbf{M}_j = \left| \left\{ i : p_i = \frac{j}{n} \right\} \right|.$$

Then, $\mathbf{M} \sim \mathcal{U}(H_{n,m})$ where $H_{n,m} = \{h \in \llbracket 0, m \rrbracket^n : \sum_{i=1}^n h_i = m\}$ is the set of histogram ([Gazin et al., 2023](#), Theorem A.1). We can now remark that $H_{n,m}$ is in bijection with the set of ordered trajectories $A_{n,m} = \left\{ a \in \left\{ \frac{i}{n} : 0 \leq i \leq n \right\}^m : a_1 \leq \dots \leq a_m \right\}$. For example $\phi : H_{n,m} \rightarrow A_{n,m}$ defined for $h \in H_{n,m}$ by:

$$\phi(h) = \frac{1}{n} \left(\underbrace{0, \dots, 0}_{h_0 \text{ times}}, \dots, \underbrace{n, \dots, n}_{h_n \text{ times}} \right),$$

is such a bijection. Then $\mathbf{p}_\uparrow = \phi(\mathbf{M})$ is uniformly drawn into $A_{n,m}$. $\square$

A.2 Proof of Lemma 3.1

Distribution of the marginal. Let $P \sim \mathcal{U}(A_{n,m})$, $k \in \llbracket m \rrbracket$ and $j \in \llbracket 0 : n \rrbracket$. The probability of the event $\{P(j) = k/n\}$ is the probability that the sub-trajectories $P_{1:(j-1)} = (P(1), \dots, P(j-1))$ belong to $\{i/n, i \in \llbracket 0 : k \rrbracket\}^{j-1}$ and the sub-trajectories $P_{(j+1):m} = (P(j+1), \dots, P(m))$ belong to $\{i/n, i \in \llbracket k : n \rrbracket\}^{m-j}$. There are respectively $\binom{j+k-1}{j-1}$ and $\binom{n-k+m-j}{m-j}$ of such trajectories which are the numbers of non-decreasing sequences of $j-1$ (resp. $m-j$) elements of $\llbracket 0 : k \rrbracket$ (resp. $\llbracket k : n \rrbracket$). Then,

$$\begin{aligned} \mathbb{P}[P(j) = k/n] &= \frac{\#\{P \in A_{n,m} : nP_{1:(j-1)} \in \llbracket 0 : k \rrbracket^{j-1}, nP(j) = k, nP_{(j+1):m} \in \llbracket k : n \rrbracket^{m-j}\}}{\#A_{n,m}} \\ &= \frac{\#A_{k,j-1} \#A_{n-k,m-j}}{\#A_{n,m}} \\ &= \frac{\binom{j+k-1}{k} \binom{n+m-j-k}{n-k}}{\binom{n+m}{n}} = \mathbb{P}[\text{BetaBin}(n, j, m-j+1) = k]. \end{aligned}$$

Distribution of the empirical cdf. The event $\left\{ \sum_{j=1}^m \mathbf{1}_{n \cdot P(j) \geq \ell} = k \right\}$ is satisfied if the first $m-k$ values of P are strictly below ℓ/n and the k last ones are above ℓ/n . That means that the sub-trajectories $P_{1:(m-k)} = (P(1), \dots, P(m-k))$ belong to $\{i/n, i \in \llbracket 0 : \ell-1 \rrbracket\}^{m-k}$ and the sub-trajectories $P_{(m-k+1):m} = (P(m-k+1), \dots, P(m))$ belong to $\{i/n, i \in \llbracket \ell : n \rrbracket\}^k$. There are

respectively $\binom{\ell-1+m-k}{m-k}$ and $\binom{n-\ell+k}{k}$ of such trajectories . Then:

$$\begin{aligned} \mathbb{P}\left[\sum_{j=1}^m \mathbf{1}_{n \cdot P(j) \geq \ell} = k\right] &= \frac{\#\{P \in A_{n,m} : nP_{1:(m-k)} \in \llbracket 0 : \ell - 1 \rrbracket^{m-k}, nP_{(m-k+1):m} \in \llbracket \ell : n \rrbracket^k\}}{\#A_{n,m}} \\ &= \frac{\#A_{\ell-1,m-k} \#A_{n-\ell,k}}{\#A_{n,m}} \\ &= \frac{\binom{\ell-1+m-k}{m-k} \binom{n-\ell+k}{k}}{\binom{n+m}{m}} = \mathbb{P}[\text{BetaBin}(m, n+1-\ell, \ell) = k]. \end{aligned}$$

□

A.3 Proofs of Theorem 3.5 and Corollary 3.7

In this section we prove Theorem 3.5 and Theorem 3.7 simultaneously. Let us first remark that if the scores have no ties, under the Assumption 3.3, all the results are obtained by applying Corollary A.1 and Lemma 3.1 conditionally to $(Y_{n+1} = y)$ to the scores $(S_1^y, \dots, S_{n_y}^y, S_{n+1}(y), \dots, S_{n+m}(y))$.

Let us now consider the case where there are potentially ties between the scores. The proof uses some common idea with the proof of Proposition 1 of Fermanian et al. (2025). Let δ be the minimum non null distance between the new scores of the true class and those evaluated in the calibration set:

$$\delta := \min\left\{|s - s'|, s \neq s' \in \{S_i\}_{i \in \llbracket n \rrbracket} \cup \{S_{n+j}(Y_{n+1})\}_{j \in \llbracket m \rrbracket}\right\}.$$

If all the scores are equal, δ can be set to any non null value (for example 1). We now define some proxy scores $\tilde{S}$ by:

$$\tilde{S}_i^y := S_i^y + U_i^y \delta, \quad \text{for } y \in \llbracket K \rrbracket, i \in \llbracket n_y \rrbracket, \quad \tilde{S}_{n+j} := S_{n+j} + U_j \delta, \quad \text{for } j \in \llbracket m \rrbracket, \quad (20)$$

where U_i^y, U_j for $y \in \llbracket K \rrbracket, i \in \llbracket n_y \rrbracket$ and $j \in \llbracket m \rrbracket$ are i.i.d. uniform random variables in $[0, 1]$. The proxy scores $\tilde{S}$ satisfy Assumption 3.3 and have no ties. We can then apply Corollary A.1 to the proxy scores $(\tilde{S}_1^y, \dots, \tilde{S}_{n_y}^y, \tilde{S}_{n+1}(y), \dots, \tilde{S}_{n+m}(y))$ conditionally to $(Y_{n+1} = y)$ to get (11). Indeed, the class conditional p-values $\tilde{p}$ associated to the proxy scores $\tilde{S}$ is equal to p^{rd} :

$$\begin{aligned} \tilde{p}_j(y) &:= \frac{1}{n_y} \sum_{i=1}^{n_y} \mathbf{1}\{\tilde{S}_{n+j}(y) \leq \tilde{S}_i^y\} \\ &= \frac{1}{n_y} \sum_{i=1}^{n_y} \left[\mathbf{1}\{S_{n+j}(y) < S_i^y\} + \mathbf{1}\{S_{n+j}(y) = S_i^y\} \mathbf{1}\{\delta U_j \leq \delta U_i^y\} \right] \\ &= \frac{1}{n_y} \sum_{i=1}^{n_y} \left[\mathbf{1}\{S_{n+j}(y) < S_i^y\} + \mathbf{1}\{S_{n+j}(y) = S_i^y\} \mathbf{1}\{U_j \leq U_i^y\} \right] = p_j^{\text{rd}}(y). \end{aligned}$$

Then the distribution of the marginal $p_{(j)}^{\text{rd}}$ is a consequence of Lemma 3.1.

To get Corollary 3.7, we can remark that the proxy scores preserve the order as $S_{n+j}(y) < S_i^y \implies \tilde{S}_{n+j}(y) < \tilde{S}_i^y$. Thus for $j \in \llbracket m \rrbracket$ and $y \in \llbracket K \rrbracket$:

$$\begin{aligned} y \in \mathcal{C}_\alpha^{cd}(X_{n+j}) &\iff S_{n+j}(y) \leq S_{(\lfloor (n_y+1)(1-\alpha) \rfloor)}(y) \iff \sum_{i=1}^{n_y} \mathbf{1}\{S_{n+j}(y) \leq S_i^y\} \geq \lfloor (n_y+1)\alpha \rfloor \\ &\iff \sum_{i=1}^{n_y} \mathbf{1}\{S_{n+j}(y) < S_i^y\} + \sum_{i=1}^{n_y} \mathbf{1}\{S_{n+j}(y) = S_i^y\} \geq \lfloor (n_y+1)\alpha \rfloor \\ &\iff \sum_{i=1}^{n_y} \mathbf{1}\{\tilde{S}_{n+j}(y) < \tilde{S}_i^y\} \geq \lfloor (n_y+1)\alpha \rfloor \\ &\iff \sum_{i=1}^{n_y} \mathbf{1}\{\tilde{S}_{n+j}(y) \leq \tilde{S}_i^y\} \geq \lfloor (n_y+1)\alpha \rfloor \\ &\iff y \in \tilde{\mathcal{C}}_\alpha^{cd}(X_{n+j}), \end{aligned}$$

where $\tilde{\mathcal{C}}_\alpha^{cd}$ is the class conditional prediction set (5) constructed with the proxy scores $\tilde{S}$. We have therefore shown that $\tilde{\mathcal{C}}_\alpha^{cd}(X_{n+j}) \subset \mathcal{C}_\alpha^{cd}(X_{n+j})$, using that the randomization preserves the order and that the scores "tilde" have no ties. This implies that almost surely, the empirical coverage of the true scores is lower bounded by the empirical coverage of conformal sets constructed with the proxy scores, *i.e.*, :

$$\sum_{j=1}^m \mathbf{1}\{Y_{n+1} \in \mathcal{C}_\alpha^{cd}(X_{n+j})\} \geq \sum_{j=1}^m \mathbf{1}\{Y_{n+1} \in \tilde{\mathcal{C}}_\alpha^{cd}(X_{n+j})\}.$$

We conclude again using Lemma 3.1.

A.4 Proof of Proposition 4.1

The result is a direct consequence of Corollary 3.7. $\square$

A.5 Proof of Proposition 4.2

Let us denote $B_y = \sum_{j=1}^m \mathbf{1}\{y \in \mathcal{C}_\alpha^{cd}(X_{n+j})\}$. First observe that:

$$\{Y_{n+1} \notin \mathcal{C}_\alpha^{Bin}\} = \{B_{Y_{n+1}} < Q_\alpha(\mathcal{B}(m, 1 - \alpha))\} \subset \{B_{Y_{n+1}} < Q_\alpha(\mathcal{B}(m, 1 - \alpha_{Y_{n+1}}))\},$$

where $\alpha_{Y_{n+1}} = \frac{\lfloor (n_{Y_{n+1}} + 1)\alpha \rfloor}{n_{Y_{n+1}} + 1} < \alpha$ and $Q_\alpha(\mathbb{Q})$ denotes the quantile α of the distribution $\mathbb{Q}$. Let $Z_y \sim \mathcal{B}(m, 1 - \alpha_y)$ and $\beta_y \sim \text{BetaBin}(m, n_y + 1 - k_y, k_y)$ where $k_y = \lfloor \alpha(n_y + 1) \rfloor$. Then, for $y \in \llbracket K \rrbracket$:

$$\begin{aligned} \mathbb{P}[Y_{n+1} \notin \mathcal{C}_\alpha^{Bin} | Y_{n+1} = y] &\leq \mathbb{P}[B_y < Q_\alpha(\mathcal{B}(m, 1 - \alpha_y)) | Y_{n+1} = y] \\ &\leq \mathbb{P}[\beta_y < Q_\alpha(\mathcal{B}(m, 1 - \alpha_y))], \end{aligned}$$

using Corollary 3.7. It follows that:

$$\begin{aligned} \mathbb{P}[Y_{n+1} \notin \mathcal{C}_\alpha^{Bin} | Y_{n+1} = y] &\leq \mathbb{P}[Z_y < Q_\alpha(\mathcal{B}(m, 1 - \alpha_y))] \\ &\quad + d_{TV}(\mathcal{B}(m, 1 - \alpha_y), \text{BetaBin}(m, n_y + 1 - k_y, k_y)) \\ &\leq \alpha + \frac{(m-1) \min(1, \alpha m)}{n_y + 1}, \end{aligned} \tag{21}$$

where we have used Corollary F.2 in the last inequality to bound the total variation distance between the Binomial and the Beta-Binomial distributions. We have then obtained the conditional coverage guarantee.

Assume now that the labels $Y_1, \dots, Y_{n+1}$ are i.i.d.. Then $n_y = \sum_{i=1}^n \mathbf{1}\{Y_i = y\}$ is independent of Y_{n+1} and follows a Binomial distribution $\mathcal{B}(n, q(y))$ for $q(y) := \mathbb{P}[Y_1 = y]$. Then:

$$\mathbb{E}\left[\frac{1}{n_{Y_{n+1}} + 1}\right] = \mathbb{E}\left[\sum_{y=1}^K \frac{q(y)}{n_y + 1}\right] \leq \mathbb{E}\left[\sum_{y=1}^K \frac{q(y)}{q(y)(n+1)}\right] = \frac{K}{n+1},$$

where we have used Lemma F.3 for the inequality. That gives the marginal coverage after taking the expectation relatively to Y_{n+1} into (21). $\square$

B Failure of the naive approaches

We discuss in this section the challenges that naive approaches may face when dealing with the case of aggregated predictions.

B.1 Synthetic dataset

We first consider two naive methods that we view as natural. We illustrate them in the case of mean aggregation and evaluate their performance on the synthetic dataset presented in Section 6.1.

First method: directly using the calibration set. A first naive idea is to construct the prediction set using the original scores and to use it directly with the aggregating scores, *i.e.*, when the aggregation

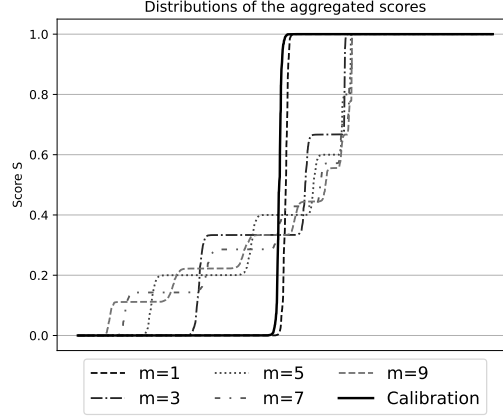


Figure 4: Distribution of the averaged softmax scores for synthetic data. 5000 averaged scores are computed and sorted for $m \in \{1, 3, 5, 7, 9\}$.

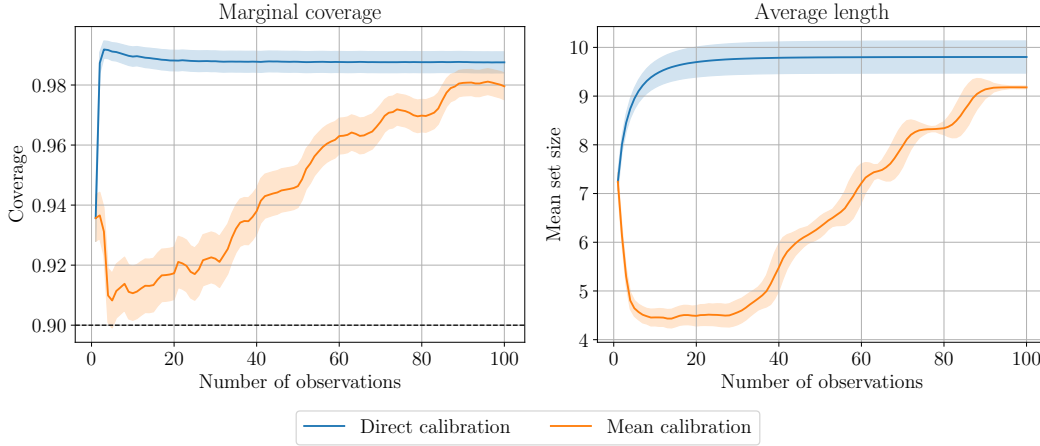


Figure 5: Empirical coverage and average lengths for class-conditional approach using two naive calibration sets for synthetic data. The calibration set is of size 5000, the coverage and the length are evaluated on 300 multi-inputs repeated 1000 times. The confidence region is $\pm (\text{empirical std})/\sqrt{5000}$

is the mean, we are considering the set $\left\{ y : S_m^{\text{agg}}(y) := m^{-1} \sum_{j=1}^m S_{n+j}(y) \leq S_{(\lceil (n_y+1)(1-\alpha) \rceil)}^y \right\}$. Formally, this set lacks any guarantee since the new (aggregated) score $S_m^{\text{agg}}(Y_{n+1})$ follows a different distribution than the calibration scores. In Figure 4, we compare the distribution of this averaged scores for $m = 1$ to 9 observations to the distribution of calibration scores. A clear change in distribution can be observed even for small value of m . In practice, this method produces conservative prediction sets, whose size increases with the number of repetitions, as it can be seen in Figure 5.

Second method: aggregating the calibration scores. A second idea to obtain a calibration set exchangeable with the new score is, for a given number of observations m , to aggregate the calibration scores by class, *i.e.*, by averaging them in groups of m points, in order to mimic the distribution of aggregated scores. These new aggregated calibration scores will be exchangeable under Assumption 3.3 with the new aggregated score. However, by doing so, the calibration set is reduced to approximately n/m points (more precisely, $\sum_{y=1}^K \lfloor n_y/m \rfloor$). Recall that the number of points per class must be at least α^{-1} in order to obtain informative class-conditionally valid prediction sets, which no longer holds when m is large. As shown in Figure 5, this method yields valid prediction sets whose size decreases for small values of m , but increases again when the number of observations becomes too large. This strategy may be sufficient in cases involving a limited number of inputs, or for classes with a sufficient number of points, *i.e.*, when $m \ll n_y$. Note however that the methods based on

p-values aggregation outperform this strategy, with an average size quickly below 4 elements (see Figure 2).

Remark B.1. Figure 4 also highlights the occurrence of tied scores produced by the softmax output of a neural network. When examining the calibration curve, constructed from one minus the softmax outputs corresponding to the true classes, we observe that many of these values are either close or equal to 0 or 1 (plateaus at the left or the right). It indicates that the classifier often returns a Dirac, sometimes in the true class (scores of 0) but also often in a wrong class (scores of 1). This leads to score ties, therefore caused by a combination of classifier overconfidence and numerical approximation.

B.2 LifeClef dataset

Our approach assumes that, at the time of calibration, only individual data points are observed, rather than multi-inputs. If multi-inputs are available at calibration time, an alternative to our methods would be to apply a class-conditional conformal prediction approach to the multi-inputs. In this setting, each data point corresponds to a multi-input, and each calibration score is computed as the aggregated score of that multi-input. However, as with the "second method" presented above for the synthetic dataset, such an approach requires a sufficient number of multi-inputs per class, specifically at least α^{-1} . In the case of the LifeCLEF 2015 dataset, these conditions are not satisfied, and consequently, this method yields prediction sets that include nearly all classes. In comparison, the p-value aggregation methods does not suffer this issue as it is constructed from the calibration scores taken individually. To overcome this issue, it could be worthwhile to explore class-conditional conformal prediction methods specifically adapted to long-tailed distributions, such as [Ding et al. \(2023, 2025\)](#) but this remains out of the scope of this work. Nevertheless, we emphasize that if a sufficient number of multi-inputs is available, we can expect this method to behave better.

This is illustrated in Table 2, which reports the results of this naive method applied to the LifeCLEF dataset constructed as described in Section 6.2 in average over all sizes and for different sizes of multi-inputs. The aggregation strategy remains a simple average of the softmax outputs. However, as previously discussed, the large prediction set sizes are due primarily to the limited amount of available data, rather than to the quality of the aggregation itself.

m	1	3	5	7	9	In average
Coverage (%)	88.2 ± 0.5	96.3 ± 0.6	97.5 ± 0.7	95.5 ± 1.5	99.0 ± 1.0	91.9 ± 0.3
Length	623 ± 1	665 ± 1	673 ± 1	675 ± 1	677 ± 1	646 ± 1

Table 2: Evaluation of class-conditional conformal prediction using multi-inputs structure conditional of different size m of multi inputs and marginally over the different sizes. Recall that $K = 688$.

C Majority voting.

We present in this section the majority voting strategies proposed by [Gasparin and Ramdas \(2024\)](#) for aggregating conformal sets and the guarantee we get in our setting. The purpose of these methods is to aggregate marginally valid prediction sets $\mathcal{C}_1, \dots, \mathcal{C}_m$ coming from some blackbox method. We have compared to two of their methods that we consider of interest for our application. The *majority* vote set $\mathcal{C}^M$ contains the classes selected by at least one half of the sets. Formally:

$$\mathcal{C}^M = \left\{ y : \frac{1}{m} \sum_{j=1}^m \mathbf{1}\{y \in \mathcal{C}_j\} \geq 1/2 \right\}. \quad (22)$$

If the sets $\mathcal{C}_j$ have a coverage of $1 - \alpha$, this set has a marginal coverage of at least $1 - 2\alpha$. To be able to compare this method to our sets of class-conditional coverage $1 - \alpha$, we apply it to the sets $\mathcal{C}_j = \mathcal{C}_{\alpha/2}^{cd}(X_{n+j})$ of class-conditional coverage $1 - \alpha/2$. Then, the resulting majority vote set has a class conditional coverage of at least $1 - \alpha$, i.e., for all $y \in \llbracket K \rrbracket$:

$$\mathbb{P}[y \in \mathcal{C}^M | Y_{n+1} = y] = \mathbb{P}\left[\frac{1}{m} \sum_{i=1}^m \mathbf{1}_{y \in \mathcal{C}_{\alpha/2}^{cd}(X_{n+i})} \geq \frac{1}{2} \mid Y_{n+1} = y\right] \geq 1 - \alpha.$$

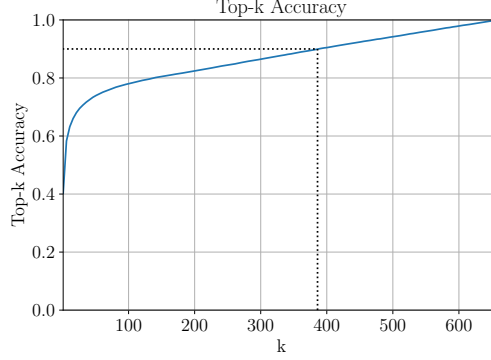


Figure 6: Top-k accuracy of the model for the LifeClef dataset. The dot line indicates the top-k accuracy of 0.9: at least $k = 380$ classes are needed to reach this level.

This method can be improved when the sets are exchangeable (Gasparin and Ramdas, 2024): in that situation the intersection $\mathcal{C}^{Me}$ of the majority vote sets with increasingly more "voters" defined by:

$$\mathcal{C}^{Me} = \left\{ y : \forall k \in \llbracket m \rrbracket : \frac{1}{k} \sum_{j=1}^k \mathbf{1} \left\{ y \in \mathcal{C}_{\alpha/2}^{cd}(X_{n+i}) \right\} \geq 1/2 \right\} \quad (23)$$

ensures a class conditional coverage of at least $1 - \alpha$. In fact, Assumption 3.3 guarantees that the sets $\mathcal{C}_{\alpha/2}^{cd}(X_{n+i})$ for $i \in \llbracket m \rrbracket$ are exchangeable given Y_{n+1} .

D Further experiments and technical details for Section 6.2

D.1 Technical details

The classifier considered is ResNet50, pre-trained on ImageNet-300K (Garcin et al. (2021)) (see code and models at <https://github.com/plantnet/PlantNet-300K/>), for which we have fine tuned the last layer on our training set. The system we used for this training phase is equipped with a 4x Intel Xeon Gold 6142 (64 cores/128 threads total @ 2.6-3.7 GHz, 88MB L3 cache) while the GPUs are 2x NVIDIA A10 (24GB VRAM each) and 2x NVIDIA RTX 2080 Ti (11GB VRAM each), for a total of 70GB GPU VRAM. Obtaining the softmax scores required about 10 hours on our GPU resources to fine tune this pre-trained model. The conformal method runs on a standard machine. The codes of the experiments are provided in the supplementary materials.

Figure 6 presents the top-k accuracy of the model for different values of k . To achieve an accuracy of 0.9, one needs to consider at least approximately $k = 380$ classes. This information can be compared with the average sizes of the conformal prediction sets shown in Figure 3. These sets tend to be significantly smaller for the methods we introduce, when a few additional observations are available. Even for $m = 1$ with a large temperature parameter (Temp=20), p-value-based methods produce smaller prediction sets, whose size is less than 100. This is likely due to the randomization involved in the p-value computation.

D.2 Additional experiments

In Figure 7, we report the marginal coverage of the prediction sets, as well as the average class-conditional coverage. Due to the small number of observations available for some classes (e.g., only 14 test points in the smallest class, which yields at most one multi-input sample of size $m = 10$ for this class), evaluating the conditional coverage per class becomes infeasible. In this case, the average class-conditional coverage serves as a proxy for the true conditional coverage. It is also worth noting that it closely aligns with the marginal coverage, which is expected given the limited number of test points.

In Figure 8, we present the results (coverage and size) of our methods applied using the APS score (Romano et al., 2020) instead of the softmax score used in the rest of our experiments. Let

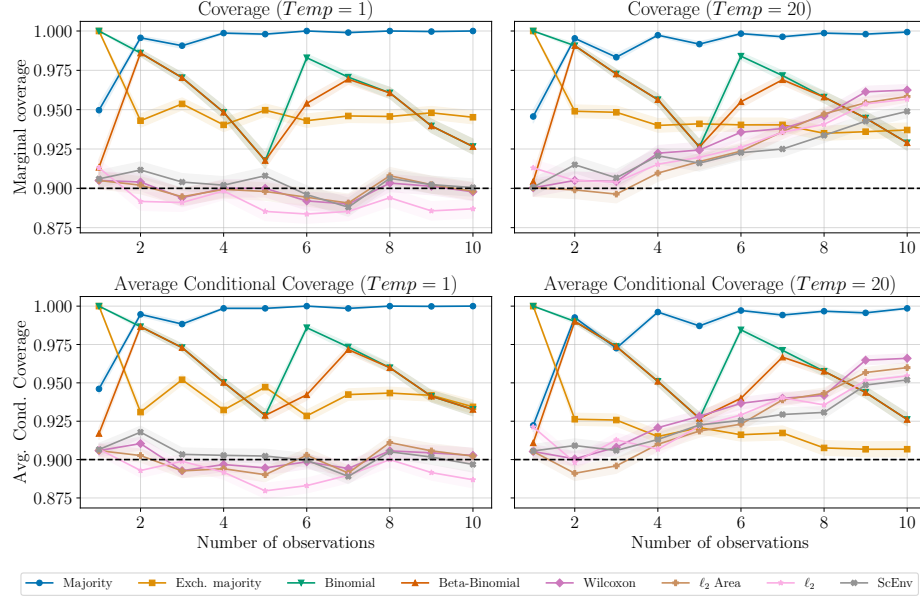


Figure 7: Coverage of the methods for the LifeClef dataset and different values of temperature.

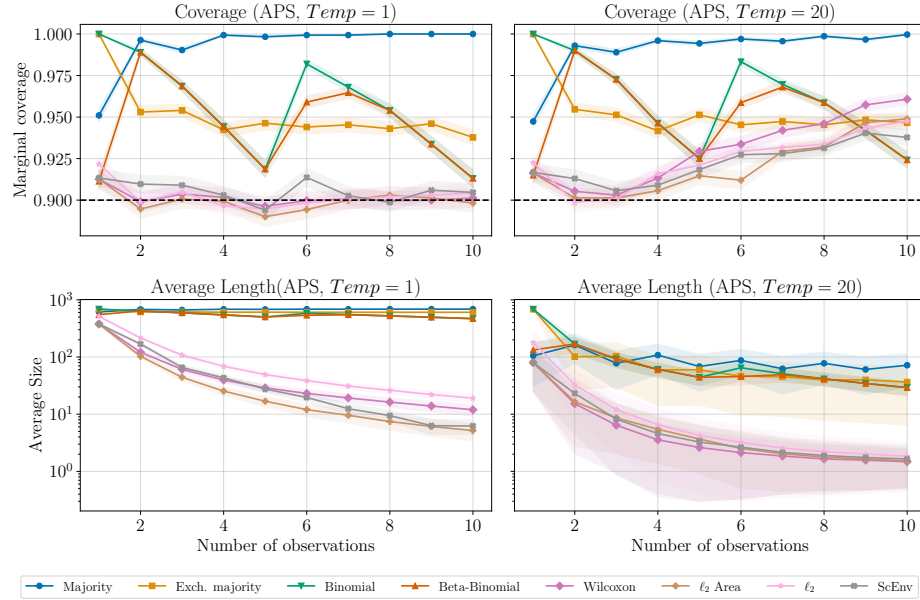


Figure 8: Coverage and size (in log scale) of the methods for the LifeClef dataset with APS score and different values of temperature.

$\hat{\pi} : \mathcal{X} \rightarrow [0, 1]^{|\mathcal{Y}|}$ be the classifier (softmax output of the neural network), it is defined as:

$$s_{\text{APS}}(x, y) = \sum_{y' \in \mathcal{Y}} \hat{\pi}(x)_{y'} \mathbf{1}_{\hat{\pi}(x)_{y'} > \hat{\pi}(x)_y} + U \hat{\pi}(x)_y, \quad (24)$$

where $U \sim \mathcal{U}(0, 1)$ is a uniform random variable. The results are similar to those obtained using the softmax score; the p-value aggregation methods offer again a substantial improvement in terms of size compared to the majority-based methods.

E Limitations of the assumption of exchangeability.

The methods proposed in this work rely on the assumption that the multi-input samples are exchangeable with data from the same class. In the context of Pl@ntNet, this means that multiple images of the same plant behave similarly to images of different plants from the same species. In the experiments presented above, this assumption is satisfied by construction of the dataset, as the multi-inputs are sampled within the same class.

The LifeCLEF Plant Identification Task 2015 dataset also features a multi-input structure, derived from images jointly submitted by users. However, we observed that the exchangeability assumption does not hold for the multi-inputs in this dataset. The correlation between images within the same input is strong enough to break Assumption 3.3, resulting in prediction sets that are no longer valid.

Our p-value-based methods implicitly rely on the idea that each new observation brings additional information, thereby enabling more refined class selection. When images within the same input are too similar, the resulting p-values tend to be overly close, which can lead to the rejection of the correct class.

In contrast, majority vote-based methods are more robust to this effect and typically yield prediction sets with appropriate coverage, although often at the cost of larger size. The Binomial and Beta-Binomial methods still improve upon majority vote strategies (with a reduction in size of around 100 for $m = 10$), but, like the p-value aggregation approach, they no longer offer theoretical guarantees when the exchangeability assumption is violated.

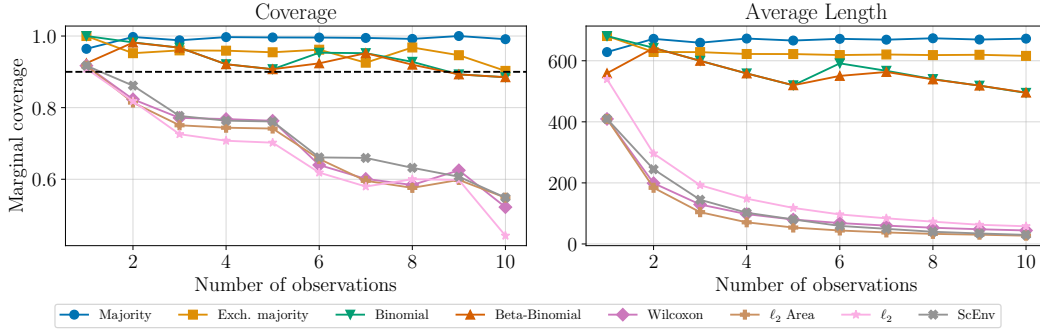


Figure 9: Empirical coverage and average lengths for the LifeClef dataset with multi-input breaking exchangeability.

The coverage and the average length of Figure 9 are evaluated for a number of multi-input observation from 4102 for $m = 1$ to 112 to $m = 10$. The temperature parameter is equal to 1. The loss of exchangeability causes a clear decrease in coverage of methods based on p-values aggregation.

F Probabilistic results

F.1 Total variation distance between Binomial and Beta-Binomial distributions

The following Lemma gives a bound on the total variation distance between a Binomial and a Beta-Binomial distribution.

Lemma F.1 (Teerapabolarn, 2008). *Let $m \in \mathbb{N}$, $\alpha, \beta \in \mathbb{R}_+$, if $p = \frac{\alpha}{\alpha + \beta}$ and $q = 1 - p$, then:*

$$d_{TV}\left(\mathcal{B}(m, p), \text{BetaBin}(m, \alpha, \beta)\right) \leq (1 - p^{m+1} - q^{m+1}) \frac{m(m-1)}{(m+1)(1 + \alpha + \beta)}.$$

Let us apply this bound to the conformal setting where BetaBin has parameters m , k_α and $n + 1 - k_\alpha$ with $k_\alpha = \lceil (n + 1)(1 - \alpha) \rceil$ and $\alpha \in (0, 1)$.

Corollary F.2. Let $k_\alpha = \lceil (n+1)(1-\alpha) \rceil$, then

$$d_{TV}\left(\mathcal{B}(m, k_\alpha/n+1), \text{BetaBin}(m, k_\alpha, n+1-k_\alpha)\right) \leq \frac{(m-1) \min(1, m\alpha)}{n+2}.$$

Proof. It is directly deduced from Lemma F.1. Just lower bound q by 0 and $p = \frac{k_\alpha}{n+1}$ by $1-\alpha$ to get that:

$$d_{TV}\left(\mathcal{B}(m, k_\alpha/n+1), \text{BetaBin}(m, k_\alpha, n+1-k_\alpha)\right) \leq (1 - (1-\alpha)^{m+1}) \frac{m}{m+1} \frac{(m-1)}{n+2}$$

Then let us just remark that:

$$\frac{(m-1)m}{m+1} (1 - (1-\alpha)^{m+1}) \leq \frac{(m-1)m}{m+1} \min(1, (m+1)\alpha) \leq (m-1) \min(1, \alpha m).$$

□

F.2 Technical lemma

Lemma F.3. Let $N \sim \mathcal{B}(n, p)$, then:

$$\mathbb{E}\left[\frac{1}{N+1}\right] = \frac{1}{(n+1)p} (1 - (1-p)^{n+1}).$$

Proof. By direct computation:

$$\begin{aligned} \mathbb{E}\left[\frac{1}{N+1}\right] &= \sum_{k=0}^n \frac{1}{k+1} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^n \binom{n+1}{k+1} \frac{1}{n+1} p^k (1-p)^{n-k} \\ &= \frac{1}{(n+1)p} \sum_{k=1}^{n+1} \binom{n+1}{k} p^k (1-p)^{n+1-k} \\ &= \frac{1}{(n+1)p} (1 - (1-p)^{n+1}). \end{aligned}$$

where the last equality holds because the sum is almost the full binomial expansion, except for the first term.

□

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The contributions are mostly methodological, and the methods introduced are well explained, with theoretical results as well as experiments showing their interest.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We have raised some issues about the exchangeability assumption that is not necessarily holding in practice. For instance for the LifeCLEF Plant Identification Task 2015 dataset ([Goëau et al., 2015](#)), we comment clearly that the assumption is only partially met. Other limit include the classwise setting that is possibly leading to poor results for long-tail label distribution. We discuss that in the conclusion-limitation section, and provide some results in Appendix E.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (*e.g.*, independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, *e.g.*, if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The proof are mostly in Appendix A to meet the size constraint, but are all given. The main assumption of the paper is Assumption 3.3 provided in the main paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The Python code is provided in the supplementary material, and the datasets are either simulated or open data like the the LifeCLEF Plant Identification Task 2015³ (Goëau et al., 2015). Overall the method is simple and could be reimplemented from scratch by an interested reader.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

³<https://www.imageclef.org/lifeclef/2015/plant>

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (*e.g.*, to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The Python code is provided in supplementary materials, and the datasets are either simulated or open data like the the LifeCLEF Plant Identification Task 2015⁴ (Goëau et al., 2015). Overall the method is simple and could be reimplemented from scratch by an interested reader.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (*e.g.*, for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (*e.g.*, data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experiments are rather straightforward, as the paper is mostly methodological, and the full details are provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

⁴<https://www.imageclef.org/lifeclef/2015/plant>

Justification: We provide errors bar on the average size of our sets and report 95% confidence interval (with Monte Carlo simulations of 5000 samples for synthetic data). For the coverage, as the variable are binary, the error bar where not included being less informative.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (*e.g.*, Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (*e.g.*, negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The experiments were short (less than one hour for the CPU experiments) once the scores are computed. Obtaining the softmax scores required about 10 hours on our GPU resources to fine tune a model pre-trained on the PI@ntNet-300K dataset⁵ (Garcin et al., 2021) for the the LifeCLEF Plant Identification Task 2015 experiments. The system is equipped with a 4x Intel Xeon Gold 6142 (64 cores/128 threads total @ 2.6-3.7 GHz, 88MB L3 cache) while the GPUs are 2x NVIDIA A10 (24GB VRAM each) and 2x NVIDIA RTX 2080 Ti (11GB VRAM each), for a total of 70GB GPU VRAM. These informations are provided in Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (*e.g.*, preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The contribution is centered on statistical methodology for classification, and respect all the code of Ethics.

⁵<https://github.com/plantnet/PlantNet-300K/>

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (*e.g.*, if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The direct contribution has the potential to benefit society in the mid to long term. Specifically, our methodological advancements could enhance citizen science platforms like Pl@ntNet, benefiting both the general public and ecologists by improving how these systems handle multi-inputs, as stated in introduction.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (*e.g.*, disinformation, generating fake profiles, surveillance), fairness considerations (*e.g.*, deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (*e.g.*, gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (*e.g.*, pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This is irrelevant for a methodological paper.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (*e.g.*, code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the datasets we used precisely. For the code, apart from standard Python code (numpy, pytorch, matplotlib, etc.) we rely mostly on our own code. It will be release under MIT License.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (*e.g.*, CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (*e.g.*, website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The research code with document is associated as zip file in the supplementary material.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.