RANK ENVELOPES AND FINITE-SAMPLE GEOMETRY IN SPLIT CONFORMAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Split conformal prediction (SCP) provides finite-sample, distribution-free marginal coverage through discrete calibration ranks. In small-calibration, high-coverage regimes, however, the rank grid may be too coarse to represent the desired coverage level, causing SCP to become conservative or vacuous. We develop rank-indexed quantile envelopes that characterize this finite-sample resolution through order-statistic geometry. The envelopes partition the rank space into forced-in, ambiguous, and forced-out regions and yield explicit finite-sample coverage bounds for prediction rules constrained by this geometry. The resulting construction separates rank decisions that are fixed by finite-sample calibration from those that remain unresolved and can therefore be handled by different efficiency mechanisms. When randomization is acceptable, envelope mixing uses the unresolved region to attain exact marginal target coverage whenever the target lies within the attainable coverage range. When historical, prior, or model-based information about the score distribution is available, reference-guided rules use this information only within the unresolved region while retaining explicit distribution-free coverage control. The same construction extends naturally to exchangeable strata, providing stratum-wise finite-sample guarantees when effective calibration sizes are small because coverage is required separately across classes or subgroups. The framework also clarifies the limits of deterministic improvement under finite rank resolution and recovers standard SCP as the ambiguity region vanishes. Experiments in regression and classification illustrate how randomization, auxiliary information, and stratification can improve informativeness in small-calibration regimes while making their corresponding coverage guarantees explicit.

1 INTRODUCTION

Providing reliable uncertainty quantification is important for deploying machine learning models in high-stakes domains such as medical diagnostics (Esteva et al., 2017), autonomous systems Bojarski et al. (2016), and legal auditing. Conformal prediction (CP) Vovk et al. (2005); Lei et al. (2018); Angelopoulos & Bates (2021); Saunders et al. (1999) has emerged as a widely used framework for this task, offering rigorous, distribution-free, finite-sample coverage guarantees. Given a test instance with an unknown label (e.g., an image), conformal prediction constructs a prediction set – a collection of plausible labels designed to contain the true label with a user-specified marginal coverage probability (e.g., 95%). To do so, it relies on a labeled holdout calibration set to compute nonconformity scores, which measure how well a model's prediction aligns with the observed outcome. These scores are then used to quantify uncertainty for future predictions. Crucially, the usual marginal coverage guarantee holds under exchangeability of the calibration and test examples, without requiring a parametric model for their distribution.

Despite its theoretical strengths, the practical informativeness of CP can be strongly constrained by the calibration sample size (m). In small-sample regimes, CP can yield highly conservative or even vacuous prediction sets, limiting its utility in applications requiring personalization or subgroup-specific guarantees Vadlamani et al. (2025); Lu et al. (2022); Gibbs et al. (2025). This issue is particularly relevant in medical diagnostics Banerji et al. (2023); Liu & Meng (2016) and class-conditional image classification Vovk et al. (2003); Ding et al. (2023), where only a small number of representative calibration examples may be available within a group. Because attainable rank-based

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

coverage levels lie on a discrete grid with spacing $1/(m+1)$, a *resolution bottleneck* arises when the target miscoverage level α is small relative to $1/(m+1)$. In such regimes, standard split conformal prediction (SCP) may be forced to use a vacuous endpoint in order to retain its distribution-free coverage guarantee. Although mathematically valid, such prediction sets may provide little useful information for downstream decisions. This finite-sample discreteness raises a broader question. Standard SCP simultaneously uses the calibration ranks to determine both the coverage guarantee and the resulting prediction-set size. Once a single empirical quantile is selected, there is little room to distinguish what is required for validity from how the remaining uncertainty should be handled for efficiency. In this work, we ask: *Can these two roles be separated more explicitly?* We address this question by shifting the analysis from a single empirical order statistic to the continuous unit interval of population quantile locations. This perspective allows us to characterize what finite-sample calibration determines, what remains unresolved, and how different prediction rules may operate within that unresolved region.

Our framework characterizes finite-sample uncertainty in rank-to-quantile mappings through *rank-indexed quantile envelopes* derived from order-statistic geometry. These envelopes partition the rank space into forced-in, ambiguous, and forced-out regions and induce explicit finite-sample coverage bounds for prediction rules constrained by the envelope geometry (Theorem 3.1). In particular, the ambiguity band (Proposition 3.8) identifies the rank decisions that are not determined by finite-sample calibration alone. This separation provides a common validity layer on top of which different efficiency mechanisms can be constructed. When randomization is acceptable, an envelope-mixing rule attains exact marginal target coverage whenever the target lies within the attainable coverage range (Theorem 3.9). When auxiliary information about the score distribution is available, a projected-quantile rule uses a reference distribution only within the unresolved region while inheriting explicit distribution-free coverage bounds (Proposition 3.10). The framework also clarifies when finite-sample rank constraints leave no room for deterministic improvement at the same target guarantee. The same geometric construction extends naturally to structured settings. In particular, rank envelopes can be constructed within exchangeable strata, yielding stratum-wise finite-sample guarantees when effective calibration sizes become small because prediction guarantees are required separately across classes or subgroups (Section 3.1.2). Thus, rather than treating small-sample discreteness as a phenomenon to be removed, our framework characterizes what the available calibration ranks support and provides principled mechanisms for operating within the remaining uncertainty. Experiments in regression and classification illustrate how randomization, auxiliary information, and stratification affect the resulting coverage–efficiency behavior in precisely the regimes where standard SCP becomes conservative or vacuous.

Our Contributions:

- **Finite-Sample Rank Geometry.** We introduce rank-indexed quantile envelopes based on finite-sample order-statistic geometry, separating rank decisions that are determined by calibration from those that remain unresolved.
- **Universal Finite-Sample Coverage Bounds.** We derive deterministic, distribution-free lower and upper coverage bounds for prediction rules constrained by the envelope geometry, explicitly characterizing the coverage region supported at finite calibration size.
- **Exact Coverage via Envelope Mixing.** We develop a randomized envelope-mixing rule that attains exact marginal target coverage whenever the target lies within the attainable envelope coverage range.
- **Reference-Guided Prediction.** We introduce a projected-quantile rule that incorporates prior, historical, or model-based information about the score distribution only within the ambiguity region, while inheriting the finite-sample distribution-free coverage bounds of the envelope construction.
- **Stratified Rank Envelopes.** We extend the construction to exchangeable strata, providing stratum-wise finite-sample guarantees in settings where class- or subgroup-specific calibration samples are small.
- **Empirical Evaluation in Limited-Resolution Regimes.** Through synthetic, regression, and classification experiments, we study how the proposed mechanisms improve informativeness or calibrate coverage in small-sample, high-coverage regimes where standard SCP is conservative or vacuous.

2 BACKGROUND

A central task in predictive inference is to quantify uncertainty in model predictions at new test points. Given covariates and responses $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ drawn from an unknown distribution P, the goal is to construct a prediction set $\widehat{C}(X_{m+1}) \subseteq \mathcal{Y}$ for a new input X_{m+1} such that, for a prescribed miscoverage level $\alpha \in (0, 1)$,

$$\mathbb{P}\Big\{Y_{m+1} \in \widehat{C}(X_{m+1})\Big\} \geq 1 - \alpha, \tag{1}$$

where $(X_{m+1}, Y_{m+1}) \sim P$ is an independent test example.

We assume access to a calibration dataset $\mathcal{D}_{\text{real}} = \{(X_i, Y_i)\}_{i=1}^{m}$, consisting of exchangeable samples from P. The calibration examples are exchangeable with the test point, which forms the basis of conformal inference. Conformal prediction has emerged as a widely used framework for achieving equation 1 without making distributional assumptions on P. Following the standard conformal prediction setup, we assume a fixed, pre-trained *nonconformity score* function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which measures how incompatible a candidate label y is with an input x. Typical examples include absolute residuals, probability scores, or negative log-likelihoods. For the calibration data, we define the scores $S_i = s(X_i, Y_i),\ i = 1, \ldots, m$, and denote by $S_{m+1} = s(X_{m+1}, Y_{m+1})$ the (unobserved) score of the test example.

2.1 SPLIT CONFORMAL PREDICTION (SCP)

In its simplest and most widely used form, SCP constructs prediction sets by thresholding the nonconformity score Vovk et al. (2005); Papadopoulos et al. (2002). Let $S_{(1)} \leq S_{(2)} \leq \cdots \leq S_{(m)}$ denote the ordered calibration scores. For a target miscoverage level α, split conformal prediction defines the threshold $\hat{q}_{\text{SCP}} := S_{(k)},\ k = \lceil (m+1)(1-\alpha) \rceil$, and outputs the prediction set $\widehat{C}_{\text{SCP}}(X_{m+1}) = \{y \in \mathcal{Y} :\ s(X_{m+1}, y) \leq \hat{q}_{\text{SCP}}\}$. Under the exchangeability assumption, the rank of S_{m+1} among $\{S_1, \ldots, S_m, S_{m+1}\}$ is uniformly distributed on $\{1, \ldots, m+1\}$, which implies that $\widehat{C}_{\text{SCP}}$ satisfies the marginal coverage guarantee equation 1. This validity holds for any choice of score function and for any underlying data-generating distribution P.

SCP obtains distribution-free validity from discrete calibration ranks, but this discreteness limits resolution when m is small. Since attainable rank-based coverage levels are spaced by $1/(m+1)$, the target quantile cannot be represented when α is too small relative to the calibration size. In particular, if $(1-\alpha)(m+1) > m$, the required rank exceeds the available calibration scores, and SCP must use the endpoint $S_{(m+1)} = \infty$ for unbounded scores. The resulting prediction set is vacuous, $\widehat{C}(x) = \mathcal{Y}$, despite remaining valid (Figure 1). More generally, finite rank resolution ties the coverage requirement to a single discrete cutoff, motivating a framework that explicitly characterizes which rank decisions are determined and which remain unresolved.

3 OUR FRAMEWORK

Our objective is to construct prediction sets that satisfy the finite-sample, distribution-free coverage guarantee of the form given in equation 1, while overcoming the resolution limitations induced by discrete empirical quantiles. To this end, we move from reasoning directly on empirical order statistics to reasoning on the *continuous unit interval*, which naturally represents quantile locations. This shift enables high-resolution inference even when the calibration sample size is small.

The key observation underlying our framework is that, under exchangeability, the rank of the test score among the calibration scores is uniformly distributed, but its corresponding quantile location is not known exactly at finite sample sizes. Instead, for each possible rank, the quantile location is constrained to lie within a range determined by finite-sample order-statistic geometry. We make this uncertainty explicit by constructing *rank-indexed envelopes*, which are intervals on the unit interval that certify where the true quantile location associated with a given rank can lie while remaining compatible with distribution-free validity.

Suppose $S_1, \ldots, S_m, S_{m+1} \sim Q$, a continuous distribution on $\mathbb{R}$. We define $V_i := Q(S_i), i = 1, \ldots, m+1$. Since this is a probability integral transform, we have $V_i \sim Unif([0,1]), i =$

$1, \ldots, m+1$. Let $V_{(r)}$ denote the r-th order statistic of $V_1, \ldots, V_{m+1}$. Classical results on order statistics imply $V_{(r)} \sim \mathrm{Beta}(r, m-r+2)$. This distribution characterizes the finite-sample uncertainty in the quantile location associated with rank r. The Beta distribution of $V_{(r)}$ induces a natural finite-sample uncertainty interval for each rank. For a confidence parameter $\beta \in (0,1)$, we define the *rank-indexed envelope* as

$$L_m^\beta(r) := F_{\mathrm{Beta}}^{-1}\left(\frac{\beta}{2};\, r,\, m-r+2\right), \quad U_m^\beta(r) := F_{\mathrm{Beta}}^{-1}\left(1-\frac{\beta}{2};\, r,\, m-r+2\right), \tag{2}$$

where $F_{\mathrm{Beta}}^{-1}(\cdot;\beta_1,\beta_2)$ denotes the quantile function of a $\mathrm{Beta}(\beta_1,\beta_2)$ distribution.

By construction,

$$\mathbb{P}\big\{L_m^\beta(r) \le V_{(r)} \le U_m^\beta(r)\big\} = 1-\beta. \tag{3}$$

The collection $\{[L_m^\beta(r), U_m^\beta(r)] : r = 1, \ldots, m+1\}$ forms an envelope in quantile space that explicitly captures finite-sample uncertainty in rank-to-quantile correspondence. (see Figure 2)

We define the rank of $s(x,y)$ among the multiset $\{S_1, \ldots, S_m, s(x,y)\}$ as

$$r_m(x,y) := 1 + \sum_{i=1}^{m} \mathbf{1}\{S_i < s(x,y)\} \in \{1, \ldots, m+1\}. \tag{4}$$

With the help of the above constructed envelopes in quantile space, for fixed β, we construct envelope prediction sets, for $\tau := 1-\alpha$, as follows:

$$\widehat{C}^-(x) := \{y :\ U_m^\beta(r_m(x,y)) \le \tau\}, \quad \widehat{C}^+(x) := \{y :\ L_m^\beta(r_m(x,y)) \le \tau\}. \tag{5}$$

The following theorem provides universal finite-sample coverage bounds for any prediction set (possibly randomized) $\widehat{C}(x)$ satisfying

$$\widehat{C}^-(x) \subseteq \widehat{C}(x) \subseteq \widehat{C}^+(x) \qquad \text{for all } x. \tag{6}$$

We call these sets *admissible prediction sets*.

Theorem 3.1 (Universal Finite-Sample Coverage Bounds) *Fix $\beta \in (0,1)$ and let $\tau := 1-\alpha$. Define the prediction sets $C^-(X_{m+1})$, $C^+(X_{m+1})$ and $C(X_{m+1})$ as in (5) and (6). Then, under exchangeability of $(S_1, \ldots, S_m, S_{m+1})$,*

$$\mathcal{LB}_m^\beta(\tau) \ \le\ \mathbb{P}\{Y_{m+1} \in \widehat{C}(X_{m+1})\} \ \le\ \mathcal{UB}_m^\beta(\tau), \tag{7}$$

where

$$\mathcal{LB}_m^\beta(\tau) = \frac{1}{m+1}\sum_{r=1}^{m+1} \mathbf{1}\{U_m^\beta(r) \le \tau\}, \quad \mathcal{UB}_m^\beta(\tau) = \frac{1}{m+1}\sum_{r=1}^{m+1} \mathbf{1}\{L_m^\beta(r) \le \tau\}. \tag{8}$$

Theorem 3.1 provides a non-parametric characterization of the coverage guarantees for the entire class of admissible prediction sets. Crucially, these bounds are *universal*, they depend exclusively on the calibration size m and the user-specified parameters α and β, remaining invariant to the underlying score distribution P. The interval $[\mathcal{LB}_m^\beta(\tau), \mathcal{UB}_m^\beta(\tau)]$ defines the feasible coverage region, formalizing the *validity layer* of our framework. This result implies that as long as a prediction set is constrained within the rank-indexed envelopes, its coverage properties are governed by the finite-sample geometry of the rank distribution. The distance between the upper and lower bounds represents the *resolution uncertainty* inherent in small samples.

A natural question is therefore whether one can select β so that the lower bound itself already certifies the target coverage, i.e., $\mathcal{LB}_m^\beta(\tau) \ge \tau$, which guarantees that every prediction set constructed within the envelope attains the desired coverage level. The following proposition provides a closed-form answer to this question by identifying the smallest value of β for which this condition holds, and by characterizing precisely when such a choice of β is feasible at finite sample sizes.

Proposition 3.2 (Choosing β to certify $\mathcal{LB}_m^\beta(\tau) \ge \tau$) *Let $k = \lceil (m+1)\tau \rceil$. Set*

$$\beta_{\min}^{\mathcal{LB}}(\tau, m) := 2\Big(1 - F_{\mathrm{Beta}}(\tau;\, k,\, m-k+2)\Big). \tag{9}$$

Then for any $\beta \in (0,1)$ satisfying $\beta \ge \beta_{\min}^{\mathcal{LB}}(\tau, m)$, we have $\mathcal{LB}_m^\beta(\tau) \ \ge\ \tau$. Moreover, $\beta_{\min}^{\mathcal{LB}}(\tau, m)$ is the smallest value of β (within $(0,1)$ interval) for which $\mathcal{LB}_m^\beta(\tau) \ge \tau$ holds.

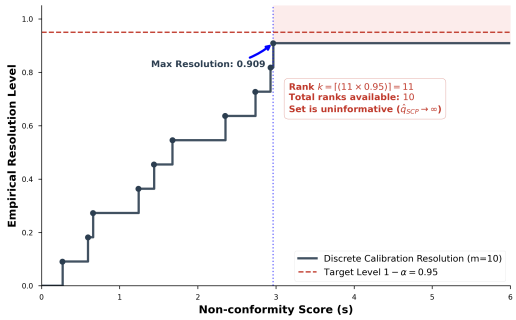


Figure 1: For $m = 10$ and $\alpha = 0.05$, the target level $1 - \alpha = 0.95$ exceeds the maximum resolution $1 - 1/(m+1) \approx 0.909$. To preserve valid coverage in this regime, SCP is forced to adopt an infinite threshold, resulting in a *vacuously valid* but uninformative prediction set $\hat{C}_{\mathrm{SCP}}(X_{m+1}) = \mathcal{Y}$.

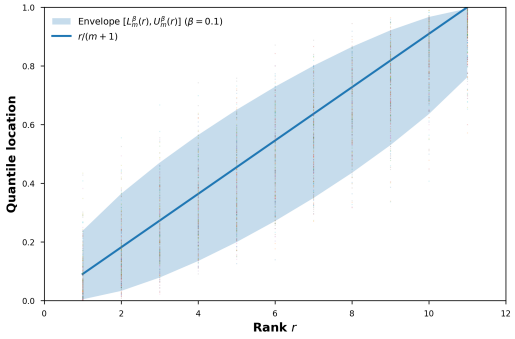


Figure 2: The shaded region illustrates the certified feasible region $[L_m^\beta(r), U_m^\beta(r)]$ for the latent quantile location of each discrete rank r. $(m = 10, \beta = 0.1)$

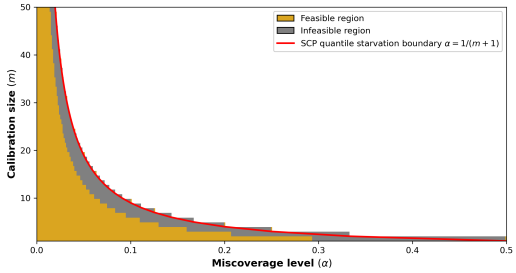


Figure 3: Feasibility of $\widehat{C}^-$ within the SCP quantile starvation regime. The red curve is the SCP quantile starvation boundary $\alpha = 1/(m+1)$ (equivalently $k = \lceil (m+1)\tau \rceil \leq m$). Gold indicates (m, α) for which there exists $\beta \in (0, 1)$ such that $\mathcal{LB}_m^\beta(\tau) \geq \tau$ (Corollary 3.3); gray is infeasible region.

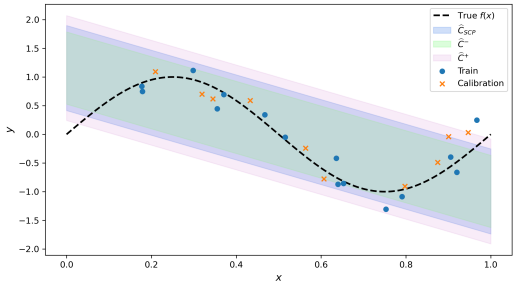


Figure 4: Toy regression example: The SCP band $\widehat{C}_{\mathrm{SCP}}$ (blue) is sandwiched between the certified lower and upper envelopes $\widehat{C}^-$ (green) and $\widehat{C}^+$ (purple), respectively. Training and calibration points are shown for reference.

Corollary 3.3 (Feasibility of $\mathcal{LB}_m^\beta(\tau) \geq \tau$ with $\beta \in (0,1)$) *With notation as in Proposition 3.2, there exists a $\beta \in (0,1)$ such that $\mathcal{LB}_m^\beta(\tau) \geq \tau$ if and only if $F_{\mathrm{Beta}}(\tau;\, k,\, m-k+2) \geq \frac{1}{2}$. Equivalently, the median of $\mathrm{Beta}(k, m-k+2)$ is at most τ.*

Figure 3 provides a geometric view of the feasibility condition in Corollary 3.3. Within the SCP quantile starvation regime, the figure identifies the region of (m, α) for which the certified lower set $\widehat{C}^-$ can already attain the target coverage. The gray region highlights parameter regimes where this condition fails, thereby delineating the precise settings in which additional efficiency mechanisms become necessary.

Our framework generalizes SCP, recovering it as a degenerate limit when the finite-sample uncertainty in the rank-to-quantile mapping vanishes.

Proposition 3.4 (SCP as a special case) *Let $k = \lceil (m+1)\tau \rceil$. Suppose β is chosen such that*

$$U_m^\beta(r) \leq \tau \quad \textit{for all } r \leq k, \qquad \textit{and} \qquad L_m^\beta(r) > \tau \quad \textit{for all } r \geq k+1. \tag{10}$$

Then the envelope prediction sets coincide, $\widehat{C}^-(x) = \widehat{C}^+(x)$, and the resulting prediction set reduces to the classical SCP set $\widehat{C}_{SCP}(x) = \{y :\ r_m(x,y) \leq k\}$.

In other words, when the safety envelopes leave no ambiguous ranks, the feasible coverage region collapses to a single rank threshold, and the framework reduces to SCP.

Corollary 3.5 *As $m \to \infty$, all admissible prediction sets reduce to SCP.*

Figure 4 illustrates the relationship between SCP and the proposed envelope-based prediction sets in a small-sample regression setting. The SCP band $\widehat{C}_{\mathrm{SCP}}$ lies between the inner and outer envelopes, $\widehat{C}^{-}(x) \subseteq \widehat{C}_{\mathrm{SCP}}(x) \subseteq \widehat{C}^{+}(x), \quad \forall x$. These envelopes define deterministic lower and upper bounds on all admissible prediction sets. In small-sample regimes where SCP becomes overly conservative or vacuous, the envelopes separate, revealing an *ambiguity region* that can be exploited to improve efficiency while preserving finite-sample guarantees.

We provide a short tutorial in Appendix C that illustrates all steps of our framework through a toy example.

3.1 Efficiency Layer

While Theorem 3.1 establishes the feasible region of admissible sets, the *efficiency layer* seeks to optimize informativeness within these certified bounds. Standard SCP is constrained to a single empirical order statistic; in contrast, our framework identifies an *ambiguity band* of ranks compatible with validity. By making principled selections within this band, we can recover resolution lost to discretization and improve efficiency without compromising the underlying frequentist guarantees.

Let $\mathrm{Size}(\widehat{C}(x))$ denote a notion of prediction set size (e.g., length for regression intervals, cardinality for classification, or more generally a measure $\mu(\widehat{C}(x))$). We measure efficiency by the expected size $\mathcal{E}(\widehat{C}) := \mathbb{E}\big[\mathrm{Size}(\widehat{C}(X_{m+1}))\big]$.

We first record a basic implication of the sandwich structure.

Proposition 3.6 (Deterministic size sandwich) *Assume $\mathrm{Size}(\cdot)$ is monotone with respect to set inclusion (i.e., $A \subseteq B \Rightarrow \mathrm{Size}(A) \le \mathrm{Size}(B)$). If equation 6 holds for all x, then*

$$\mathbb{E}\big[\mathrm{Size}(\widehat{C}^{-}(X_{m+1}))\big] \;\le\; \mathbb{E}\big[\mathrm{Size}(\widehat{C}(X_{m+1}))\big] \;\le\; \mathbb{E}\big[\mathrm{Size}(\widehat{C}^{+}(X_{m+1}))\big].$$

Proposition 3.6 provides universal, distribution-free *efficiency envelopes*: regardless of the data distribution, any valid construction within our framework cannot be smaller than $\widehat{C}^{-}$ nor larger than $\widehat{C}^{+}$.

3.1.1 Comparing SCP to the envelope sets

To compare SCP with our framework, recall that SCP uses the deterministic rank cutoff $k = \lceil (m+1)\tau \rceil$ and includes candidates whose rank does not exceed k.

Proposition 3.7 (When SCP lies between envelope sets) *Let $k = \lceil (m+1)\tau \rceil$ and suppose $L_m^{\beta}(k) \le \tau \le U_m^{\beta}(k)$. Then for all x, $\widehat{C}^{-}(x) \subseteq \widehat{C}_{\mathrm{SCP}}(x) \subseteq \widehat{C}^{+}(x)$, and hence, by Proposition 3.6, $\mathbb{E}\big[\mathrm{Size}(\widehat{C}^{-}(X_{m+1}))\big] \le \mathbb{E}\big[\mathrm{Size}(\widehat{C}_{\mathrm{SCP}}(X_{m+1}))\big] \le \mathbb{E}\big[\mathrm{Size}(\widehat{C}^{+}(X_{m+1}))\big]$.*

Condition $L_m^{\beta}(k) \le \tau \le U_m^{\beta}(k)$ exactly states that the target quantile level τ lies inside the envelope at the SCP cutoff rank k. When this holds, SCP is one admissible set among many; the efficiency layer can then improve upon SCP by making more informed choices within admissible sets.

3.1.2 Mechanisms for efficiency improvements

Proposition 3.8 *Fix β and τ. For each x, define the forced-in set $\widehat{C}^{-}(x) = \{y : U_m^{\beta}(r_m(x,y)) \le \tau\}$, the forced-out region $\widehat{R}^{\mathrm{out}}(x) = \{y : L_m^{\beta}(r_m(x,y)) > \tau\}$, and the ambiguity band $\widehat{A}(x) = \{y : L_m^{\beta}(r_m(x,y)) \le \tau < U_m^{\beta}(r_m(x,y))\}$. Then $\widehat{C}^{+}(x) = \widehat{C}^{-}(x) \cup \widehat{A}(x)$ and $\widehat{A}(x)$ is the* unique *region where admissible sets may differ: any admissible $\widehat{C}$ must coincide with $\widehat{C}^{-}(x)$ on $\widehat{A}(x)^c$. Consequently, all efficiency gains over SCP can only arise from decisions on $\widehat{A}(x)$.*

Equivalently, our framework induces two rank cutoffs $r_-(\beta,\tau) := \max\{r : U_m^\beta(r) \le \tau\}$, $r_+(\beta,\tau) := \max\{r : L_m^\beta(r) \le \tau\}$, so that $\widehat{C}^-(x) = \{y : r_m(x,y) \le r_-\}$, $\widehat{C}^+(x) = \{y : r_m(x,y) \le r_+\}$. In particular, the ambiguity band is $\{r_- + 1, \dots, r_+\}$. The framework now enables multiple efficiency improvement strategies that are unavailable in classical SCP:

(i) Exact coverage via randomized inclusion. When τ lies strictly between $L_m^\beta(r)$ and $U_m^\beta(r)$ for some ranks, randomized inclusion over the ambiguous ranks can achieve exact target coverage while controlling expected set size. This avoids systematic over-coverage inherent to discrete SCP.

Theorem 3.9 (Exact coverage via envelope mixing) *Assume $\mathcal{LB}_m^\beta(\tau) \le \tau < \mathcal{UB}_m^\beta(\tau)$. Define a randomized prediction set $\widehat{C}_\lambda$ by*

$$\widehat{C}_\lambda(x) := \begin{cases} \widehat{C}^+(x), & \text{with probability } \lambda, \\ \widehat{C}^-(x), & \text{with probability } 1-\lambda, \end{cases} \qquad \text{where } \lambda := \frac{\tau - \mathcal{LB}_m^\beta(\tau)}{\mathcal{UB}_m^\beta(\tau) - \mathcal{LB}_m^\beta(\tau)}. \tag{11}$$

Then, under exchangeability, $\mathbb{P}\{Y_{m+1} \in \widehat{C}_\lambda(X_{m+1})\} = \tau$. Moreover, for any monotone size functional Size$(\cdot)$, $\mathbb{E}[\text{Size}(\widehat{C}_\lambda(X_{m+1}))] = (1-\lambda)\,\mathbb{E}[\text{Size}(\widehat{C}^-(X_{m+1}))] + \lambda\,\mathbb{E}[\text{Size}(\widehat{C}^+(X_{m+1}))]$.

(ii) Safe reference/prior anchoring. A reference model for score geometry (e.g., a parametric family for the score CDF, or a learned reference distribution) can be used to select inclusion decisions *within* the envelope, improving efficiency when the reference is informative. Crucially, the envelope constraints firewall validity: even if the reference is misspecified, coverage remains certified by Theorem 3.1.

Proposition 3.10 *Fix (m, β, τ). Let $\tilde{Q}$ be any continuous reference distribution on $\mathbb{R}$ with CDF $\Phi_{\tilde{Q}}$. Let the reference quantile coordinate be $u_{\tilde{Q}}(s) := \Phi_{\tilde{Q}}(s)$ and define its projection onto the feasible envelope interval by $\widetilde{u}(x,y) := \min\left\{U_m^\beta(r_m(x,y)), \max\{L_m^\beta(r_m(x,y)), u_{\tilde{Q}}(s)\}\right\}$. Define the* projected-quantile prediction set *by the inclusion rule $y \in \widehat{C}_{\text{proj}}(x) \iff \widetilde{u}(x,y) \le \tau$. Then $\widehat{C}_{\text{proj}}$ is an admissible set.*

Corollary 3.11 (Inherited universal coverage bounds) *Under the conditions of Proposition 3.10, $\widehat{C}_{\text{proj}}$ inherits the universal finite-sample coverage bounds of Theorem 3.1.*

(iii) Stratified envelopes and conditional efficiency. Our framework generalizes naturally to heterogeneous settings via stratification. By constructing rank-indexed envelopes within exchangeable strata, we obtain localized feasible regions that improve conditional efficiency, particularly for underrepresented groups. This stratification enables higher-resolution inference tailored to specific subgroups while strictly preserving finite-sample validity within each stratum (see Appendix F.2 for empirical evaluation).

Collectively, these mechanisms identify the ambiguity band induced by finite-sample discretization as the primary source of efficiency loss in standard CP, and provide a systematic framework to harvest this latent resolution while strictly preserving distribution-free, finite-sample validity.

4 Simulation Study

The goal of this section is to empirically validate the theoretical properties of rank-indexed envelopes, specifically characterizing the resolution-efficiency trade-off in small-sample regimes. We evaluate how the framework exploits the *ambiguity band* to provide informative sets where SCP fails. Implementation details and additional experiments, including the projected-quantile strategy ($\hat{C}_{\text{proj}}$) and stratified envelopes, are provided in Appendix F.

4.1 Half-normal scores

We compare SCP with the inner envelope $\widehat{C}^-$ (certified when feasible; Corollary 3.3) and the mixing set $\widehat{C}_\lambda$ under i.i.d. half-normal scores $S_i \sim |N(0,1)|$. We run 1000 Monte Carlo trials over coverage

levels $\tau \in \{0.97, 0.98, 0.99\}$ and calibration sizes $m \in \{5, 10, 15, 20, 25, 50, 100, 500\}$. Results are summarized in Table 1. In small-sample regimes ($m \leq 25$), where SCP becomes near-vacuous, both proposed methods yield substantial reductions in prediction set size. The lower set $\widehat{C}^-$ achieves the largest gains but frequently undercovers, reflecting its operation at the extreme boundary of the admissible region. In contrast, the mixing rule $\widehat{C}_\lambda$ preserves coverage near the nominal level by interpolating within the *ambiguity band*, delivering consistent efficiency gains without sacrificing validity. As m increases, the admissible region collapses and all methods converge to SCP, in line with the theory.

4.2 Heteroskedastic scores

We next evaluate robustness under heteroskedasticity using a scale-mixture model $S_i = \sigma_i |N(0,1)|$ with $\log \sigma_i \sim N(0, 0.8^2)$. This preserves exchangeability while inducing heavy-tailed variability in the scores. As shown in Table 1, discretization effects in SCP become more pronounced in this setting. The lower set $\widehat{C}^-$ continues to achieve large nominal gains but exhibits severe undercoverage, especially at high τ. In contrast, $\widehat{C}_\lambda$ remains stable, maintaining coverage while providing meaningful efficiency improvements. These results demonstrate that the proposed framework can exploit latent finite-sample resolution even under substantial heterogeneity, provided the efficiency layer is calibrated within the admissible region.

Table 1: Efficiency gain vs. SCP (in %) and achieved empirical coverage for synthetic score experiments. Each entry reports **Gain** (mean ± SE) **/ mean Coverage** over 1000 trials. Efficiency gain is computed as $100 \times \big(1 - \mathbb{E}[\text{Size}(\widehat{C})]/\mathbb{E}[\text{Size}(\widehat{C}_{\text{SCP}})]\big)$.

Experiment	m	Target coverage τ					
		0.97		0.98		0.99	
		$\widehat{C}^-$	$\widehat{C}_\lambda$	$\widehat{C}^-$	$\widehat{C}_\lambda$	$\widehat{C}^-$	$\widehat{C}_\lambda$
Half-normal	5	68.07±0.35/0.8405	12.61±0.84/0.9728	68.29±0.37/0.8349	8.10±0.71/0.9799	79.26±0.25/0.6662	2.61±0.45/0.9889
	10	62.46±0.32/0.9099	20.37±0.95/0.9695	62.15±0.32/0.9122	13.89±0.84/0.9784	62.44±0.33/0.9089	5.64±0.58/0.9913
	15	59.00±0.31/0.9376	27.50±0.95/0.9718	59.20±0.31/0.9365	19.12±0.90/0.9789	58.96±0.30/0.9390	9.76±0.70/0.9903
	20	56.92±0.31/0.9506	35.49±0.90/0.9695	56.66±0.30/0.9525	24.26±0.91/0.9801	56.19±0.31/0.9536	12.37±0.76/0.9892
	25	54.81±0.29/0.9618	42.28±0.77/0.9705	54.52±0.29/0.9635	29.18±0.88/0.9803	54.94±0.30/0.9611	14.18±0.78/0.9889
	50	0.00/0.9817	0.00/0.9817	0.00/0.9814	0.00/0.9814	50.07±0.28/0.9788	23.70±0.82/0.9894
	100	0.00/0.9697	0.00/0.9697	0.00/0.9791	0.00/0.9791	0.00/0.9901	0.00/0.9901
	500	0.00/0.9703	0.00/0.9703	0.00/0.9793	0.00/0.9793	0.00/0.9895	0.00/0.9895
Heteroskedastic	5	88.87±0.27/0.8291	17.77±1.13/0.9673	88.08±0.31/0.8348	10.89±0.92/0.9818	94.83±0.12/0.6586	2.29±0.46/0.9916
	10	84.19±0.33/0.9119	27.08±1.26/0.9709	83.77±0.36/0.9107	18.97±1.13/0.9784	83.84±0.36/0.9124	10.28±0.88/0.9889
	15	80.07±0.39/0.9420	39.13±1.30/0.9700	81.41±0.34/0.9372	27.81±1.24/0.9780	81.44±0.35/0.9380	11.56±0.91/0.9908
	20	78.83±0.40/0.9518	49.59±1.25/0.9699	78.90±0.38/0.9520	33.38±1.25/0.9808	78.91±0.39/0.9515	16.99±1.05/0.9889
	25	77.32±0.38/0.9620	62.41±1.02/0.9691	77.65±0.40/0.9595	39.26±1.26/0.9789	78.18±0.37/0.9589	18.29±1.07/0.9902
	50	0.00/0.9808	0.00/0.9808	0.00/0.9811	0.00/0.9811	70.71±0.43/0.9797	36.33±1.15/0.9901
	100	0.00/0.9708	0.00/0.9708	0.00/0.9808	0.00/0.9808	0.00/0.9905	0.00/0.9905
	500	0.00/0.9709	0.00/0.9709	0.00/0.9805	0.00/0.9805	0.00/0.9898	0.00/0.9898

Note: Gains are reported relative to SCP length, which overcovers in these settings. For sufficiently large m, rank resolution improves and all methods coincide with SCP, yielding zero gain.

5 Real Data Analysis

To assess the practical utility of our framework, we evaluate the inner envelope $\widehat{C}^-$, *envelope-mixing* set $\widehat{C}_\lambda$ and the *projected-quantile* set $\widehat{C}_{\text{proj}}$ (using a scale-aligned half-normal reference) against SCP. We focus on small-sample, high-coverage regimes ($\tau \geq 0.97$), where rank discretization most severely compromises SCP resolution. Full experimental details, hyperparameter settings, and comprehensive results are provided in Appendix G, along with additional real data examples using image classification on CIFAR-10 dataset (Appendix G.3).

5.1 California Housing Data

We evaluate on the California Housing dataset ($n = 20{,}640$), predicting median house values using a `RandomForestRegressor` trained on $n_{\text{train}} = 2000$ samples, with absolute residuals as nonconformity scores. We report *efficiency gain* (percentage reduction in average set size relative to

Table 2: Mean gain (%) relative to SCP on California Housing. Reported values are Monte Carlo means over $T = 500$ trials.

$\tau = 0.97$	m=10	m=20	m=30	m=40	m=50
$\widehat{C}^-$	50.00	34.68	29.95	0.00	0.00
$\widehat{C}_{\text{proj}}$	46.74	34.11	29.89	0.00	0.00
$\widehat{C}_\lambda$	17.13	23.15	27.22	0.00	0.00
$\tau = 0.99$	m=10	m=20	m=30	m=40	m=50
$\widehat{C}^-$	50.85	34.18	28.11	22.27	17.08
$\widehat{C}_{\text{proj}}$	43.06	31.46	26.99	21.71	16.93
$\widehat{C}_\lambda$	6.45	5.58	10.36	8.30	7.05

Table 3: Mean gain (%) relative to SCP on DIABETES. Reported values are Monte Carlo means over $T = 500$ trials.

$\tau = 0.97$	m=10	m=15	m=20	m=25	m=30
$\widehat{C}^-$	32.08	23.39	18.43	13.99	10.69
$\widehat{C}_{\text{proj}}$	20.99	17.18	14.46	11.42	9.04
$\widehat{C}_\lambda$	9.72	12.81	12.20	11.34	10.03
$\tau = 0.99$	m=10	m=15	m=20	m=25	m=30
$\widehat{C}^-$	28.96	23.04	18.62	13.75	11.33
$\widehat{C}_{\text{proj}}$	7.86	6.26	6.42	4.45	3.80
$\widehat{C}_\lambda$	3.38	3.91	3.44	3.85	3.59

SCP) over $T = 500$ trials for calibration sizes $m \in \{10, \ldots, 50\}$ (Table 2). In low-sample regimes ($m \leq 30$), SCP often yields vacuous intervals. Here, the lower set $\widehat{C}^-$ achieves gains exceeding 50%, while the projected set $\widehat{C}_{\text{proj}}$ closely matches this performance, indicating that reference anchoring effectively recovers resolution. The mixing rule $\widehat{C}_\lambda$ is more conservative but remains uniformly more efficient than SCP while preserving coverage. As m increases, the ambiguity region shrinks and all methods converge to SCP.

5.2 DIABETES DATA

We replicate the study on the DIABETES dataset ($n = 442$) using a `GradientBoostingRegressor` with calibration sizes $m \in \{10, \ldots, 30\}$. This smaller, higher-noise setting exacerbates SCP's tendency to produce vacuous intervals at high coverage levels. As shown in Table 3, the results reflect the theoretical safety–efficiency trade-off. The lower set $\widehat{C}^-$ achieves the largest gains (e.g., $\approx 32\%$ at $m = 10$) but can undercover in sparse regimes. In contrast, the mixing rule $\widehat{C}_\lambda$ maintains near-nominal coverage while delivering consistent efficiency improvements over SCP. The projected set $\widehat{C}_{\text{proj}}$ provides a balanced alternative, producing substantially tighter sets than SCP without the severe undercoverage of $\widehat{C}^-$. Additional coverage and standard error results are reported in Appendix G.2.

6 CONCLUSION

We introduced a framework for conformal inference that decouples distribution-free *validity* from statistical *efficiency* by explicitly modeling finite-sample uncertainty in rank-to-quantile mappings. By characterizing this uncertainty through analytical envelopes derived from the distribution of order statistics, we identify the feasible region of prediction sets that satisfy coverage guarantees. This perspective resolves the resolution limits of standard conformal prediction in small-sample, high-coverage regimes, enabling principled efficiency gains without sacrificing validity.

The resulting framework supports a range of constructions, including exact-coverage mixing and reference-anchored prediction, and recovers standard SCP as a limiting case. More broadly, it provides a modular foundation for extending conformal inference to settings involving heterogeneity, prior information, and structured uncertainty, while retaining finite-sample guarantees.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on the conceptual framing and research methodology, assist in refining mathematical claims and their proof structure, and support the interpretation and presentation of experimental results. We have not used generative AI tools to generate synthetic datasets, implement the proposed methods, clean or reformat datasets, translate research material, or conduct qualitative or thematic data analysis; the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to draft and edit portions of the manuscript for clarity and concision, suggest manuscript structure, titles and keywords, summarize and organize relevant literature, and assist with brainstorming and identifying potential gaps in presentation. All AI-assisted mathematical statements and methodological suggestions were independently checked by the authors against the underlying derivations and established literature, and all descriptions of experimental results were verified against the actual outputs of our experiments. The authors reviewed and revised all AI-assisted text and take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops methodology for uncertainty quantification and does not involve human subjects or the collection of sensitive personal data. The proposed methods may be relevant to high-stakes applications, where prediction sets should be interpreted according to their stated finite-sample guarantees and limitations, particularly in small-calibration regimes.

REPRODUCIBILITY STATEMENT

We provide complete methodological details, proofs, experimental settings, and implementation choices needed to reproduce the results. Code and scripts for all simulations and real-data experiments will be released upon publication.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Christopher RS Banerji, Tapabrata Chakraborti, Chris Harbron, and Ben D MacArthur. Clinical ai tools must convey predictive uncertainty for each individual patient. *Nature medicine*, 29(12): 2996–2998, 2023.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Meshi Bashari, Roy Maor Lotan, Yonghoon Lee, Edgar Dobriban, and Yaniv Romano. Synthetic-powered predictive inference. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=ujg9XKwqWT.

Pankaj Bhagwat, Linglong Kong, and Bei Jiang. CBMA: Improving conformal prediction through bayesian model averaging. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BKSeNw2HIr.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Amirmohammad Farzaneh, Sangwoo Park, and Osvaldo Simeone. Quantile learn-then-test: Quantile-based risk control for hyperparameter optimization. *IEEE Signal Processing Letters*, 2024.

Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

Chenyin Gao, Shu Yang, Mingyang Shan, Wenyu Ye, Ilya Lipkovich, and Douglas Faries. Improving randomized controlled trial analysis via data-adaptive borrowing. *Biometrika*, 112(2): asae069, 2025.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkaf008, 2025.

Marius Hobbhahn, Agustinus Kristiadi, and Philipp Hennig. Fast predictive uncertainty for classification with bayesian deep networks. In *Uncertainty in artificial intelligence*, pp. 822–832. PMLR, 2022.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Keli Liu and Xiao-Li Meng. There is individualized treatment. why not individualized inference? *Annual Review of Statistics and Its Application*, 3(1):79–111, 2016.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 12008–12016, 2022.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pp. 345–356. Springer, 2002.

C Saunders, A Gammerman, and V Vovk. Transduction with confidence and credibility. In *Proceedings of the 16th international joint conference on Artificial intelligence-Volume 2*, pp. 722–726, 1999.

Jake Snell, Thomas P Zollo, Zhun Deng, Toniann Pitassi, and Richard Zemel. Quantile risk control: A flexible framework for bounding the probability of high-loss predictions. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=p6jsTidUkPx.

Jake C Snell and Thomas L Griffiths. Conformal prediction as bayesian quadrature. In *International Conference on Machine Learning*, pp. 56068–56084. PMLR, 2025.

John W Tukey. Non-parametric estimation ii. statistically equivalent blocks and tolerance regions–the continuous case. *The Annals of Mathematical Statistics*, pp. 529–539, 1947.

U.S. Food and Drug Administration. Considerations for the design and conduct of externally controlled trials for drug and biological products: Guidance for industry. https://www.fda.gov/media/164960/download, 2023. Accessed: 2026-05-02.

Aditya T Vadlamani, Anutam Srinivasan, Pranav Maneriker, Ali Payani, and Srinivasan Parthasarathy. A generic framework for conformal fairness. *arXiv preprint arXiv:2505.16115*, 2025.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Abraham Wald. An extension of wilks' method for setting tolerance limits. *The Annals of Mathematical Statistics*, 14(1):45–55, 1943.

Samuel S Wilks. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941.

Yao Zhang and Emmanuel J Candès. Posterior conformal prediction. *arXiv preprint arXiv:2409.19712*, 2024.

Ke Zhu, Shu Yang, and Xiaofei Wang. Enhancing statistical validity and power in hybrid controlled trials: A randomization inference approach with conformal selective borrowing. *arXiv preprint arXiv:2410.11713*, 2024.

APPENDIX

A RELATED WORK

Foundations of Predictive Inference. The construction of distribution-free prediction sets traces back to foundational results in non-parametric statistics Wilks (1941); Wald (1943); Tukey (1947). Modern CP Saunders et al. (1999); Vovk et al. (2005) provides a rigorous framework for such inference using exchangeable non-conformity scores Papadopoulos et al. (2002); Angelopoulos & Bates (2021). While cross-validation variants like Jackknife+ Barber et al. (2021); Fontana et al. (2023) improve data efficiency in small samples, they remain constrained by the discrete resolution of empirical ranks.

Distributional Priors and Quantile Functions. Recent work has explored utilizing quantile functions within frequentist CP Snell et al. (2023); Farzaneh et al. (2024) or adopting Bayesian perspectives on uncertainty Hobbhahn et al. (2022); Bhagwat et al. (2025). Most relevant to our work is Snell & Griffiths (2025), which frames CP as Bayesian quadrature. They demonstrate that Dirichlet spacings provide a posterior distribution over risk that is valid regardless of prior choice. Our framework shares this conceptual foundation but shifts the focus from posterior mean risk to *quantile-indexed envelopes*. This enables a more general geometry that incorporates prior knowledge while maintaining deterministic, finite-sample guarantees.

Sample Efficiency and Synthetic Data. A growing body of literature targets CP resolution in data-scarce regimes, particularly in fair or group-conditional settings Lu et al. (2022); Gibbs et al. (2025); Zhang & Candès (2024); Vadlamani et al. (2025). However, these methods often sacrifice tight finite-sample bounds or rely on restrictive distributional assumptions. To improve resolution, Bashari et al. (2026) introduced *Synthetic-Powered Predictive Inference* (SPI), which augments calibration sets with samples from a generative model. Furthermore, all other existing methods that rely on synthetic data are not reproducible since the synthetic data generation mechanisms depend on seeds, and hence can lead to different prediction sets for the same conditions (test point). Our analytical results bypass this issue completely and lead to deterministic prediction sets (except for envelope-mixing set).

Synthetic-Powered Predictive Inference (SPI). SPI (Bashari et al., 2026) improves conformal prediction by augmenting the calibration set with synthetic samples drawn from a generative model, thereby increasing effective rank resolution. In contrast, our framework does not rely on synthetic data or generative modeling. Instead, we provide an analytical characterization of the finite-sample uncertainty in rank-to-quantile mappings via the distribution of order statistics, which induces a *feasible region* of admissible prediction sets.

Conceptually, SPI approximates this feasible region through sampling, while our approach characterizes it exactly. This distinction has both practical and theoretical implications: our method avoids model misspecification and additional exchangeability assumptions required for synthetic data, and yields deterministic, distribution-free coverage bounds without approximation error. The reference-anchored construction in our framework can be viewed as an analytical counterpart to synthetic augmentation, where external information refines efficiency without affecting validity.

Smoothed and Randomized Conformal Prediction. Smoothed conformal prediction (Fontana et al., 2023) and related randomized variants mitigate the discreteness of rank-based quantiles by interpolating between adjacent ranks. These methods perform *local smoothing* of the empirical distribution to achieve approximate or exact coverage.

Our framework generalizes this idea by identifying the full set of admissible ranks induced by finite-sample uncertainty. When the admissible region collapses to two adjacent ranks, our envelope-mixing rule recovers standard randomized conformal prediction. However, in general the admissible region may span multiple ranks, and our approach exploits this global structure to achieve greater efficiency gains. Thus, smoothing methods can be viewed as special cases corresponding to local interpolation within a much larger feasible region.

B Proofs of Main Results

B.1 Proof of Theorem 3.1

Let R denote the rank of S_{m+1} among $(S_1, \ldots, S_m, S_{m+1})$. By exchangeability and continuity, R is uniform on $\{1, \ldots, m+1\}$.

We first bound the coverage of the extremal envelope sets. Consider the set $\widehat{C}^-$. By construction, the true label Y_{m+1} is included whenever the rank of the true test score satisfies $U_m^\beta(R) \le \tau$. Hence

$$\mathbb{P}\{Y_{m+1} \in \widehat{C}^-(X_{m+1})\} = \mathbb{P}\{U_m^\beta(R) \le \tau\}.$$

Since $U_m^\beta(r)$ is deterministic given (m, β) and R is uniform,

$$\mathbb{P}\{U_m^\beta(R) \le \tau\} = \frac{1}{m+1} \sum_{r=1}^{m+1} \mathbf{1}\{U_m^\beta(r) \le \tau\} = \mathcal{LB}_m^\beta(\tau).$$

Similarly, for the set $\widehat{C}^+$ we have inclusion whenever $L_m^\beta(R) \le \tau$, yielding

$$\mathbb{P}\{Y_{m+1} \in \widehat{C}^+(X_{m+1})\} = \mathbb{P}\{L_m^\beta(R) \le \tau\} = \mathcal{UB}_m^\beta(\tau).$$

Finally, if $\widehat{C}$ satisfies the sandwich condition equation 6, then for every realization,

$$\mathbf{1}\{Y_{m+1} \in \widehat{C}^-(X_{m+1})\} \le \mathbf{1}\{Y_{m+1} \in \widehat{C}(X_{m+1})\} \le \mathbf{1}\{Y_{m+1} \in \widehat{C}^+(X_{m+1})\}.$$

Taking expectations gives

$$\mathcal{LB}_m^\beta(\tau) \le \mathbb{P}\{Y_{m+1} \in \widehat{C}(X_{m+1})\} \le \mathcal{UB}_m^\beta(\tau),$$

as claimed.

B.2 Proof of Proposition 3.2

Since $U_m^\beta(r)$ is non-decreasing in r, the condition $\mathcal{LB}_m^\beta(\tau) \ge \tau$ is equivalent to $U_m^\beta(k) \le \tau$ for $k = \lceil (m+1)\tau \rceil$. By definition,

$$U_m^\beta(k) = F_{\mathrm{Beta}}^{-1}\left(1 - \frac{\beta}{2};\, k,\, m-k+2\right),$$

and monotonicity of the Beta quantile function implies

$$U_m^\beta(k) \le \tau \quad \Longleftrightarrow \quad 1 - \frac{\beta}{2} \le F_{\mathrm{Beta}}(\tau;\, k,\, m-k+2),$$

which is equivalent to $\beta \ge 2(1 - F_{\mathrm{Beta}}(\tau;\, k,\, m-k+2))$. The minimality follows immediately.

B.3 Proof of Corollary 3.3

A solution $\beta \in (0,1)$ exists if and only if $\beta_{\min}^{\mathcal{LB}}(\tau, m) < 1$. By equation 9, this is equivalent to $2(1 - F_{\mathrm{Beta}}(\tau;\, k,\, m-k+2)) < 1$, i.e., $F_{\mathrm{Beta}}(\tau;\, k,\, m-k+2) \ge 1/2$.

B.4 Proof of Proposition 3.4

Recall the definitions (for $\tau = 1 - \alpha$):

$$\widehat{C}^-(x) := \{y : U_m^\beta(r_m(x,y)) \le \tau\}, \qquad \widehat{C}^+(x) := \{y : L_m^\beta(r_m(x,y)) \le \tau\}.$$

We will show that, under equation 10, both sets coincide with

$$\{y : r_m(x,y) \le k\}.$$

Let $y \in \mathcal{Y}$ and set $r := r_m(x,y) \in \{1, \dots, m+1\}$. If $r \le k$, then by equation 10 we have $U_m^\beta(r) \le \tau$, hence $y \in \widehat{C}^-(x)$. If instead $r \ge k+1$, then by equation 10 we have $L_m^\beta(r) > \tau$. Since always $L_m^\beta(r) \le U_m^\beta(r)$ for each r, it follows that

$$U_m^\beta(r) \ge L_m^\beta(r) > \tau,$$

so $U_m^\beta(r) \le \tau$ fails and thus $y \notin \widehat{C}^-(x)$. Therefore,

$$\widehat{C}^-(x) = \{y : r_m(x,y) \le k\}.$$

Again let $r := r_m(x,y)$. If $r \le k$, then by equation 10 we have $U_m^\beta(r) \le \tau$ and hence $L_m^\beta(r) \le U_m^\beta(r) \le \tau$, implying $y \in \widehat{C}^+(x)$. If $r \ge k+1$, then by equation 10 we have $L_m^\beta(r) > \tau$, so $L_m^\beta(r) \le \tau$ fails and thus $y \notin \widehat{C}^+(x)$. Therefore,

$$\widehat{C}^+(x) = \{y : r_m(x,y) \le k\}.$$

Combining the two characterizations yields

$$\widehat{C}^-(x) = \widehat{C}^+(x) = \{y : r_m(x,y) \le k\}.$$

Finally, note that the split conformal prediction set can be written exactly in this rank-threshold form with $k = \lceil (m+1)\tau \rceil$, hence the envelope construction reduces to split conformal prediction in this regime.

B.5 Proof of Theorem 3.9

By Theorem 3.1, $\mathbb{P}\{Y_{m+1} \in \widehat{C}^-(X_{m+1})\} = \mathcal{LB}_m^\beta$ and $\mathbb{P}\{Y_{m+1} \in \widehat{C}^+(X_{m+1})\} = \mathcal{UB}_m^\beta$. By total probability,

$$\mathbb{P}\{Y_{m+1} \in \widehat{C}_\lambda(X_{m+1})\} = (1-\lambda)\mathcal{LB}_m^\beta + \lambda \mathcal{UB}_m^\beta = \tau.$$

The size identity follows from linearity of expectation.

B.6 Proof of Proposition 3.8

The identity $\widehat{C}^+(x) = \widehat{C}^-(x) \cup \widehat{A}(x)$ follows by partitioning ranks into (i) $U_m^\beta(r) \le \tau$ (forced-in), (ii) $L_m^\beta(r) \le \tau < U_m^\beta(r)$ (ambiguous), and (iii) $L_m^\beta(r) > \tau$ (forced-out), and using $L_m^\beta(r) \le U_m^\beta(r)$. If $\widehat{C}^- \subseteq \widehat{C} \subseteq \widehat{C}^+$, then $\widehat{C}$ must include all forced-in labels and exclude all forced-out labels, leaving only $\widehat{A}(x)$ as the locus of choice.

B.7 Proof of Proposition 3.10

Fix x and let $r = r_m(x,y)$. If $U_m^\beta(r) \le \tau$ (forced-in), then by definition of projection, $\widetilde{u}(x,y) \le U_m^\beta(r) \le \tau$, hence $y \in \widehat{C}_{\mathrm{proj}}(x)$ for all such y; thus $\widehat{C}^-(x) \subseteq \widehat{C}_{\mathrm{proj}}(x)$.

If $L_m^\beta(r) > \tau$ (forced-out), then $\widetilde{u}(x,y) \ge L_m^\beta(r) > \tau$, hence $y \notin \widehat{C}_{\mathrm{proj}}(x)$ for all such y; thus $\widehat{C}_{\mathrm{proj}}(x) \subseteq \widehat{C}^+(x) = \{y : L_m^\beta(r) \le \tau\}$. Therefore,

$$\widehat{C}^-(x) \subseteq \widehat{C}_{\mathrm{proj}}(x) \subseteq \widehat{C}^+(x) \qquad \text{for all } x,$$

and $\widehat{C}_{\mathrm{proj}}$ can only differ from $\widehat{C}^-$ when $L_m^\beta(r) \le \tau \le U_m^\beta(r)$, i.e., on $\widehat{A}(x)$.

B.8 Proof of Corollary 3.11

The sandwich condition follows from Proposition 3.10, so $\widehat{C}_{\mathrm{proj}}$ is admissible by Proposition 3.8. The claimed coverage bounds then follow immediately by applying Theorem 3.1 to the admissible set $\widehat{C}_{\mathrm{proj}}$.

C Tutorial: How the Rank-Envelope Framework Operates

We now provide a self-contained tutorial illustrating how our rank-envelope-based framework operates in practice. The goal of this experiment is to make explicit how uncertainty in finite-sample ranks is propagated into certified prediction rules, and how this differs from SCP. Throughout, we consider a calibration sample of size m (for plots, we use $m = 30, \beta = 0.1$)

C.1 Step 1: Rank uncertainty on the unit interval

Let $V_{(r)}$ denote the r-th order statistic of $n = m+1$ i.i.d. $\mathrm{Unif}(0,1)$ variables. It is well known that

$$V_{(r)} \sim \mathrm{Beta}(r,\, m - r + 2), \qquad r = 1, \ldots, m+1.$$

Rather than treating the empirical rank r as a fixed proxy for the quantile location $r/(m+1)$, our framework explicitly models the *distributional uncertainty* of $U_{(r)}$.

For a user-specified confidence level $1-\beta$, we construct rank envelopes

$$L_m^\beta(r) := Q_{\mathrm{Beta}(r,m-r+2)}(\beta/2), \qquad U_m^\beta(r) := Q_{\mathrm{Beta}(r,m-r+2)}(1-\beta/2),$$

which provide a high-probability interval for the true quantile location associated with rank r.

Figure 5 visualizes these envelopes together with the reference line $r/(m+1)$. Overlaid simulated order statistics highlight the substantial variability of ranks in small samples and motivate the need for explicit rank uncertainty quantification.

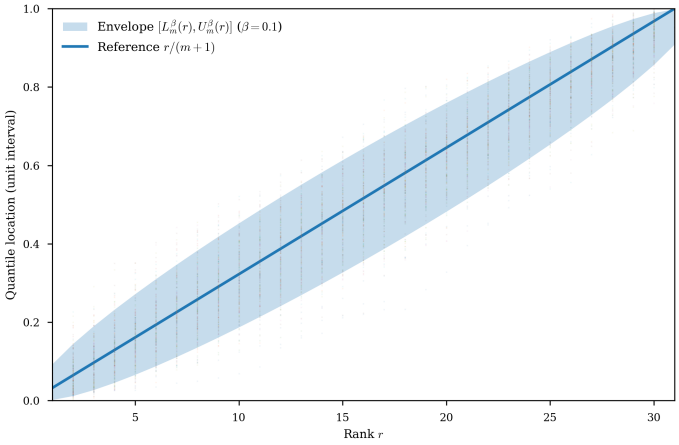


Figure 5: Rank envelopes $[L_m^\beta(r), U_m^\beta(r)]$ on the unit interval, illustrating uncertainty in finite-sample ranks.

C.2 Step 2: Forced-in, forced-out, and ambiguous ranks

Fix a target quantile level $\tau = 1-\alpha$. Using the rank envelopes, each rank r can be classified as:

- **Forced-in:** $U_m^\beta(r) \le \tau$, meaning the rank is guaranteed to lie below τ.
- **Forced-out:** $L_m^\beta(r) > \tau$, meaning the rank is guaranteed to exceed τ.
- **Ambiguous:** $L_m^\beta(r) \le \tau < U_m^\beta(r)$.

This decomposition reveals a central phenomenon: only a subset of ranks can be deterministically classified relative to τ under finite-sample uncertainty. Figure 6 illustrates these regions and highlights the induced uncertainty band in rank space.

We denote

$$r_- := \max\{r : U_m^\beta(r) \le \tau\}, \qquad r_+ := \max\{r : L_m^\beta(r) \le \tau\}.$$

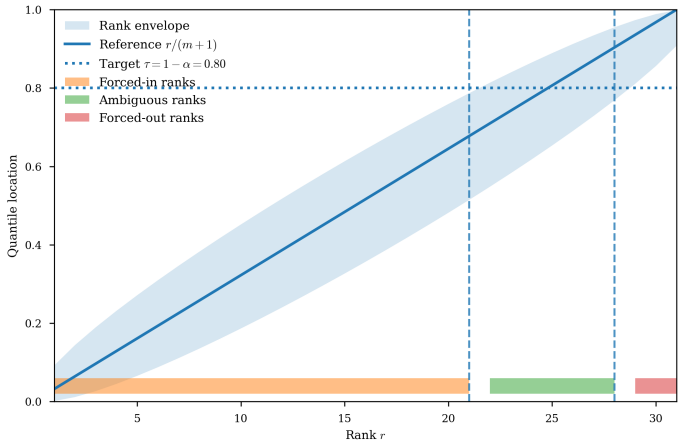


Figure 6: Forced-in, ambiguous, and forced-out ranks relative to a target quantile $\tau = 0.80$.

C.3 Step 3: Distribution-free coverage bounds

The rank classification leads to distribution-free bounds on achievable coverage:

$$\mathcal{LB}_m^\beta(\tau) = \frac{1}{m+1}\sum_{r=1}^{m+1} \mathbf{1}\{U_m^\beta(r) \le \tau\}, \qquad \mathcal{UB}_m^\beta(\tau) = \frac{1}{m+1}\sum_{r=1}^{m+1} \mathbf{1}\{L_m^\beta(r) \le \tau\}.$$

These quantities provide certified lower and upper bounds on the fraction of ranks that can be safely included under uncertainty.

Figure 7 plots these bounds as functions of τ, demonstrating how rank uncertainty induces a non-trivial gap between $\mathcal{LB}_m^\beta(\tau)$ and $\mathcal{UB}_m^\beta(\tau)$ in small samples.

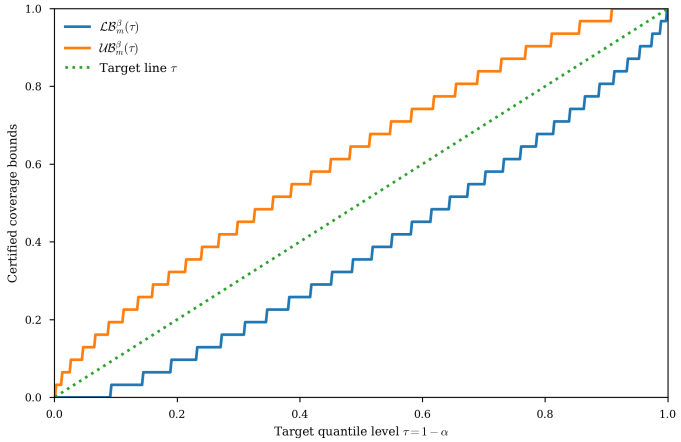


Figure 7: Distribution-free lower and upper coverage bounds induced by rank envelopes.

C.4 Step 4: Feasibility and minimal envelope width

Not all pairs (τ, β) yield meaningful guarantees. In particular, to ensure

$$\mathcal{LB}_m^{\beta}(\tau) \geq \tau,$$

β must be chosen sufficiently large. A closed-form expression yields

$$\beta_{\min}^{\mathrm{LB}}(\tau, m) = 2\left[1 - F_{\mathrm{Beta}(k, m-k+2)}(\tau)\right], \quad k = \lceil (m+1)\tau \rceil.$$

Figure 8 visualizes $\beta_{\min}^{\mathrm{LB}}(\tau, m)$ and highlights the feasible region where certified coverage is attainable. This feasibility constraint plays a central role in our method's design.

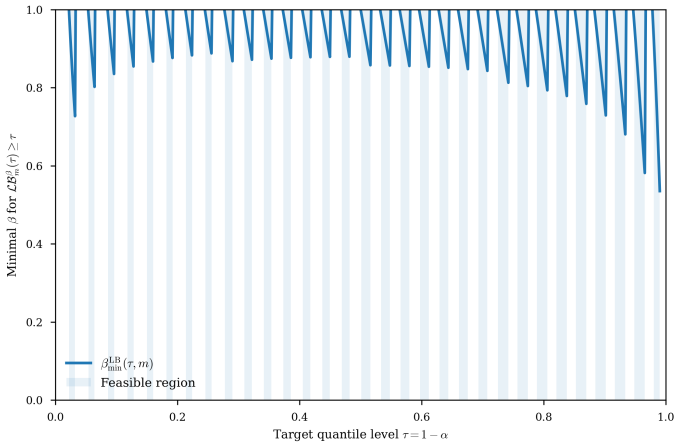


Figure 8: Minimal envelope confidence level required to certify $\mathcal{LB}_m^{\beta}(\tau) \geq \tau$.

C.5 Step 5: Comparison with split conformal cutoffs

Standard SCP uses the deterministic cutoff

$$k_{\mathrm{SCP}} = \lceil (m+1)\tau \rceil,$$

implicitly assuming negligible rank uncertainty. In contrast, our framework produces a *range* of admissible cutoffs $[r_-, r_+]$.

Figure 9 compares the SCP cutoff with the envelope-based conservative and liberal cutoffs across τ. The discrepancy is most pronounced in small samples, precisely where SCP is known to be unstable and overly conservative.

This tutorial highlights the core principle of our framework: rather than reasoning directly on empirical order statistics, we reason on quantile locations with explicit uncertainty envelopes. This separation of rank uncertainty from coverage targeting enables principled efficiency improvements while retaining finite-sample, distribution-free validity guarantees.

Remark (Handling Ties in Practice): Our theoretical development assumes continuous nonconformity scores to avoid ties and ensure well-defined rank distributions. In practice, ties may arise, particularly in classification settings or when scores are discretized.

A standard remedy is randomized tie-breaking. For example, one may add a small independent perturbation (e.g., $S_i' = S_i + \epsilon_i$ with $\epsilon_i \sim \mathrm{Uniform}(0, \delta)$ for small $\delta > 0$), or equivalently randomize uniformly within tied ranks. This restores a continuous ordering while preserving exchangeability, ensuring that the coverage guarantees of the envelope construction remain valid.

C.6 Computational Considerations

We briefly discuss the computational overhead of the proposed framework relative to SCP.

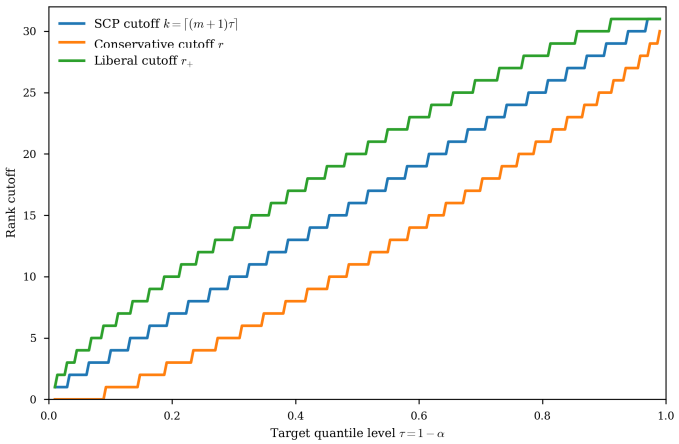


Figure 9: SCP cutoff versus envelope-based conservative and liberal rank cutoffs.

Precomputation of Rank Envelopes: For fixed calibration size m, target level τ, and envelope parameter β, the ambiguity band of admissible ranks and the corresponding coverage bounds can be precomputed using Beta quantiles. Specifically, the lower and upper envelope functions, as well as the induced feasible region, depend only on (m, τ, β) and not on the data itself. These quantities can therefore be tabulated once and reused across all predictions.

Prediction-Time Complexity: At prediction time, the procedure reduces to evaluating empirical quantiles at precomputed rank indices, similar to SCP. For randomized or reference-anchored constructions, the additional cost consists of simple interpolation or evaluation within the admissible region, which incurs negligible overhead compared to model inference.

Regression Settings: For real-valued outputs, threshold selection may involve evaluating envelope bounds or solving for quantile levels. In practice, these computations can be implemented using precomputed lookup tables or efficient numerical routines, with overall complexity linear in the calibration size.

Scalability: The framework introduces only a modest overhead beyond SCP and is well-suited to large-scale settings. Since the core quantities depend only on (m, τ, β), precomputation amortizes the cost across repeated use, making the method practical for deployment in high-throughput or real-time applications.

Overall, the additional computational cost arises primarily from one-time precomputation of envelope quantities. Once these are available, the method retains the simplicity and scalability of standard conformal prediction while enabling improved efficiency.

D LIMITATIONS

While the proposed framework provides a principled extension of conformal prediction, some limitations should be noted.

Dependence on Exchangeability: As with all conformal methods, our guarantees rely on the assumption of exchangeability between calibration and test data. Violations of this assumption (e.g., under distribution shift, temporal dependence, or covariate drift) may compromise coverage guarantees. Extending the envelope-based framework to non-exchangeable settings remains an important direction for future work.

Sensitivity to Reference Distributions: The efficiency of reference-anchored methods depends on the choice of the reference distribution. While validity guarantees are preserved regardless of this choice, a poorly specified reference may lead to suboptimal efficiency gains. In practice, simple

parametric fits or auxiliary data can provide reasonable choices, but systematic methods for selecting or learning reference distributions remain an open question.

Extensions to Structured Settings: While the framework naturally extends to stratified or group-wise conformal prediction, broader structured settings, such as time series, causal inference, or graph-based data, pose additional challenges. In these settings, both exchangeability and rank-based calibration may require modification, and integrating envelope-based constructions with existing structured conformal methods is an important direction for future work.

E Future Work

Our framework opens several directions for extending conformal inference beyond the standard i.i.d. setting, while preserving finite-sample guarantees.

Beyond Exchangeability: The current analysis relies on exchangeability, as in classical conformal prediction. Extending the envelope-based framework to non-exchangeable settings (e.g., time series, covariate shift, or distribution drift) is a natural direction. In particular, combining rank-envelope constructions with weighted or localized conformal methods may enable principled uncertainty quantification under structured dependence.

Adaptive Selection within the Admissible Region: The envelope characterization identifies a set of admissible prediction rules at finite sample sizes. A promising direction is to develop decision-theoretic or data-driven strategies for selecting optimal predictors within this region, for example by minimizing expected set size subject to coverage constraints, or by incorporating task-specific loss functions.

Structured and Conditional Inference: The geometric perspective naturally extends to stratified and group-conditional conformal prediction. This suggests applications to fairness-aware inference, where coverage guarantees must hold across subpopulations. Investigating how envelope-based methods can provide higher-resolution conditional guarantees remains an important direction.

Integration with Prior and Auxiliary Information: The reference-anchored construction provides a first step toward incorporating prior knowledge into conformal prediction. Future work could explore richer forms of inductive bias, including learned or hierarchical reference distributions, and study their impact on efficiency under model misspecification.

F Additional Experiments

F.1 Simulation Study with Synthetic scores

We begin our empirical evaluation with a controlled simulation study designed to characterize the performance of the proposed framework across the *resolution spectrum*. We specifically investigate the regime of *quantile starvation*, where the target miscoverage level α is small relative to the calibration sample size m, rendering standard split conformal prediction (SCP) uninformative.

We consider a synthetic model where non-conformity scores are drawn i.i.d. from a Half-Normal distribution, $S_i \sim |N(0,1)|$. This choice represents a canonical score distribution for regression tasks utilizing absolute residuals. We consider calibration sizes $m \in \{10, 15, 20, 30, 40, 50\}$ and target coverage levels $\tau = 1 - \alpha \in \{0.97, 0.99\}$. For each configuration, we report results averaged over $T = 1000$ Monte Carlo trials, using a test batch of size $n_{\text{test}} = 2000$ to evaluate empirical coverage.

We evaluate four strategies:

1. **SCP (Baseline):** Standard split conformal prediction using the $k = \lceil (m+1)\tau \rceil$ order statistic. In starvation regimes ($k > m$), the threshold is set to the vacuous upper bound of the score space.

2. **Inner Envelope ($\widehat{C}^-$):** The most aggressive valid set within our framework, and if feasible, utilizing the optimized safety parameter $\beta_{\min}$ (Prop. 3.2) to certify a coverage floor.

3. **Envelope Mixing ($\widehat{C}_\lambda$):** Our proposed randomized rule designed to achieve *exact* target coverage by interpolating between the inner and outer safety envelopes (Theorem 3.9).

4. **Analytical Reference Anchoring ($\widehat{C}_{\text{proj}}$):** A projected-quantile policy that incorporates an analytical Half-Normal reference $\mathcal{Q}$ to resolve the ambiguity band. To simulate a realistic deployment, we estimate the scale parameter of $\mathcal{Q}$ from the m calibration scores in each trial using Maximum Likelihood Estimation (MLE).

We measure the **Efficiency Gain (%)** relative to SCP, defined as $100 \times (1 - \mathbb{E}[\text{Size}(\widehat{C})]/\mathbb{E}[\text{Size}(\widehat{C}_{\text{SCP}})])$, and the achieved **Empirical Coverage**.

The results of the simulation are summarized in Table 4. Our analysis reveals three critical findings:

Resolution Recovery in Starvation Regimes. At $m = 10$ and $\tau = 0.99$, standard SCP enters a state of total resolution collapse ($k = 11$), producing vacuous intervals. In this regime, our proposed methods, specifically $\widehat{C}_{\text{proj}}$ and $\widehat{C}^-$, convert these infinite bounds into finite, high-utility prediction sets, achieving efficiency gains of 47.3% and 61.9% respectively. This confirms that analytical anchoring successfully "borrows" resolution from the continuous reference frame to navigate the starvation zone.

Safety-Efficiency Frontier. The results highlight a fundamental trade-off: while the inner envelope $\widehat{C}^-$ provides the most aggressive size reduction, it can undercover the target in ultra-low data regimes (e.g., 0.913 coverage for $\tau = 0.97$ at $m = 10$). In contrast, the envelope-mixing rule $\widehat{C}_\lambda$ demonstrates remarkable robustness, hitting the targets with high precision across all m. This validates the theoretical exactness of the mixing rule and positions it as the preferred policy for practitioners prioritizing calibrated reliability.

Asymptotic Convergence. As m increases, the rank-indexed envelopes collapse around the SCP threshold. In these regimes, the ambiguity band vanishes, and all methods converge to standard SCP (yielding 0% gain). This illustrates that our framework is a proper generalization of conformal inference: it provides a high-resolution boost when data is scarce but correctly reverts to the standard rank-based conformal rule as the sample size grows.

Table 4: Efficiency gain vs. SCP (in %) and achieved empirical coverage for synthetic score experiments ($|N(0,1)|$ residuals). Each entry reports **Gain $\pm$ SE / Coverage $\pm$ SE** over 1000 trials. Efficiency gain is defined as $100 \times \left(1 - \mathbb{E}[\text{Size}(\widehat{C})]/\mathbb{E}[\text{Size}(\widehat{C}_{\text{SCP}})]\right)$.

m	Method	Target coverage $\tau = 1 - \alpha$			
		0.97		0.99	
		Gain (%) $\pm$ SE	Coverage $\pm$ SE	Gain (%) $\pm$ SE	Coverage $\pm$ SE
10	$\widehat{C}^-$	62.21 ± 0.44	0.913 ± 0.004	61.96 ± 0.45	0.913 ± 0.004
	$\widehat{C}_{\text{proj}}$	55.12 ± 0.45	0.956 ± 0.002	47.35 ± 0.54	0.978 ± 0.001
	$\widehat{C}_\lambda$	22.26 ± 1.35	0.970 ± 0.003	5.96 ± 0.83	0.992 ± 0.001
15	$\widehat{C}^-$	58.93 ± 0.46	0.936 ± 0.003	59.39 ± 0.46	0.934 ± 0.003
	$\widehat{C}_{\text{proj}}$	54.03 ± 0.44	0.963 ± 0.002	48.39 ± 0.47	0.979 ± 0.001
	$\widehat{C}_\lambda$	27.54 ± 1.36	0.969 ± 0.002	10.30 ± 1.03	0.988 ± 0.002
20	$\widehat{C}^-$	56.80 ± 0.41	0.952 ± 0.002	56.87 ± 0.41	0.953 ± 0.002
	$\widehat{C}_{\text{proj}}$	53.14 ± 0.37	0.971 ± 0.001	47.98 ± 0.38	0.984 ± 0.001
	$\widehat{C}_\lambda$	34.40 ± 1.28	0.971 ± 0.002	11.07 ± 1.01	0.992 ± 0.001
30	$\widehat{C}^-$	53.43 ± 0.41	0.968 ± 0.001	53.35 ± 0.40	0.969 ± 0.001
	$\widehat{C}_{\text{proj}}$	51.63 ± 0.36	0.977 ± 0.001	46.78 ± 0.33	0.988 ± 0.001
	$\widehat{C}_\lambda$	49.74 ± 0.73	0.970 ± 0.001	17.70 ± 1.14	0.990 ± 0.001
50	$\widehat{C}^-$	0.00 ± 0.00	0.981 ± 0.001	50.24 ± 0.38	0.980 ± 0.001
	$\widehat{C}_{\text{proj}}$	0.00 ± 0.00	0.981 ± 0.001	46.10 ± 0.30	0.990 ± 0.000
	$\widehat{C}_\lambda$	0.00 ± 0.00	0.981 ± 0.001	25.97 ± 1.16	0.989 ± 0.001

F.1.1 Additional Experiment: Heteroskedastic Scores

We include an additional synthetic experiment to compare SCP, Randomized-SCP, and the envelope-mixing rule $\widehat{C}_\lambda$ in a bounded heteroskedastic setting. Scores are generated as

$$S_i = \min\{\sigma_i |Z_i|, B\}, \qquad Z_i \sim N(0,1), \qquad \log \sigma_i \sim N(0, 0.8^2),$$

with truncation level $B = 50$. We consider calibration sizes $m \in \{7, 13, 18, 23\}$, target coverage levels $\tau \in \{0.95, 0.96, 0.97, 0.98, 0.99\}$, and run 1000 trials with test size 2000.

The results in Figure 10 show three consistent patterns. First, SCP is highly conservative in the small-m/high-τ regime, often producing vacuous intervals with mean length 100. Second, Randomized-SCP mitigates this conservativeness by locally interpolating between adjacent ranks. Third, $\widehat{C}_\lambda$ matches or improves upon Randomized-SCP when the admissible rank region is wider than two ranks, since it mixes over the full feasible region rather than only adjacent ranks. For example, at $m = 18$ and $\tau = 0.95$, $\widehat{C}_\lambda$ achieves mean length 10.15 with coverage 0.9484, compared with Randomized-SCP length 13.41 and coverage 0.9503. At $m = 23$ and $\tau = 0.96$, $\widehat{C}_\lambda$ achieves length 11.35 with coverage 0.9593, compared with Randomized-SCP length 14.99 and coverage 0.9608. These results support the view that Randomized-SCP is recovered as a local special case, while envelope mixing can exploit larger ambiguity regions for additional efficiency gains.

F.2 Stratified envelopes and conditional efficiency under heterogeneous calibration.

In many real-world deployments, calibration data are composed of heterogeneous subgroups with varying noise scales. While standard SCP provides marginal coverage across the entire population, it often severely under-covers "harder" (high-variance) subgroups. Conversely, standard stratified CP (Mondrian CP) recovers conditional validity but suffers from extreme resolution collapse when a subgroup is underrepresented.

We simulate a two-group scenario with a total calibration size $m = 400$ and target miscoverage $\alpha = 0.05$. The population is highly imbalanced: a *major* group ($\sigma = 1.0$) comprises 95% of the calibration data ($m_{\text{major}} \approx 380$), while a *minor* group ($\sigma = 1.8$) comprises only 5% ($m_{\text{minor}} \approx 20$). We compare pooled calibration (SCP-all) against group-wise stratified calibration using both standard SCP and our proposed certified envelopes ($\widehat{C}^-$).

The conditional performance metrics are summarized in Table 5. Our analysis reveals a stark contrast in subgroup safety and efficiency:

- While SCP-all satisfies the 95% marginal target, it fails the conditional coverage guarantee for the high-variance minority, achieving only $\approx 75.2\%$ coverage. This underscores the danger of ignoring heterogeneity in small-sample subgroups.
- Stratified SCP restores validity for the minor group but is restricted by the low resolution of $m \approx 20$ samples. This results in significant over-coverage (97.1%) and unnecessarily wide intervals (≈ 8.64).
- Our stratified envelope approach ($\widehat{C}^-$ stratified) successfully restores conditional coverage to the target level (95.1%) while significantly improving resolution. By exploiting the analytical geometry of the order statistics, we reduce the average interval length for the minor group to 7.72, a 10.6% reduction in width compared to standard stratified CP.

F.3 Efficiency improvements using Reference distributions

We empirically evaluate the proposed *rank-envelope* framework and *projected-quantile* strategy against standard SCP in small-sample, high-coverage regimes. All experiments are conducted in the *score space* to isolate the effect of finite-sample rank resolution and ambiguity-band decision rules.

We consider a one-dimensional nonconformity score model with nonnegative scores

$$S = |Z|, \qquad Z \sim \mathcal{N}(0,1).$$

Table 5: Conditional Performance under Group Heterogeneity ($\alpha = 0.05$). We report means over $T = 500$ trials. Pooled methods fail the minority safety test, while our stratified framework maximizes resolution under valid conditional guarantees.

	Major Group ($m \approx 380$)		Minor Group ($m \approx 20$)	
Method	Coverage	Length	Coverage	Length
SCP (pooled)	0.961	4.16	0.752	4.16
SCP (stratified)	0.952	3.97	0.971	8.64
Ours (stratified)	**0.951**	**3.96**	**0.951**	**7.72**

Given a calibration set of size m, $\{S_i\}_{i=1}^m$, we form prediction sets for a new test score S_{m+1} by thresholding at a level q:

$$\widehat{C}(x) = \{y : \ s(x, y) \leq q\}.$$

In this toy setting, the induced prediction set corresponds to an interval of half-width q, hence we report

$$\text{Length}(\widehat{C}) = 2q,$$

with vacuous sets clipped at a large ceiling B for visualization. We estimate coverage by Monte Carlo on a large batch of independent test scores per trial.

Throughout, we focus on regimes with *high target coverage* $1 - \alpha$ and *small calibration size* m, where classical SCP exhibits strong discretization effects and may become vacuous when the target rank exceeds m.

We compare the following methods, all computed from the same calibration scores.

- SCP
- Inner Envelope set $\widehat{C}^-(X_{m+1})$
- Projected-quantile set $\widehat{C}_{\text{proj}}$

Experiment 1: Fixed Reference Distributions

In the first experiment, we fix the reference distribution Q *a priori* and compare SCP, $\widehat{C}^-$, and $\widehat{C}_{\text{proj}}$. We consider several reference families on $\mathbb{R}_+$:

- **Half-Normal:** $Q = \mathcal{N}^+(0, \sigma^2)$,
- **Gamma:** $Q = \text{Gamma}(a, \theta)$,
- **Log-Normal:** $Q = \text{LogNormal}(\mu, \sigma^2)$,
- **Folded Student-t:** $Q = |T|$ with $T \sim t_\nu(0, \sigma)$.

For each family, we report boxplots of empirical coverage and interval lengths over repeated trials, along with a "vacuous ceiling" to highlight trials where a method returns $+\infty$ and thus yields an uninformative prediction set. Representative results are shown in Figure 12.

Experiment 2: Data-Driven Reference Distributions

In the second experiment, we choose the reference distribution parameters *from the observed calibration scores* within each trial, yielding a data-adaptive reference $\widehat{Q}$. This reflects the intended role of the reference frame as an *efficiency prior* learned from the calibration data, while preserving validity through the sandwich constraint $\widehat{C}^- \subseteq \widehat{C}_{\text{proj}}(\widehat{Q}) \subseteq \widehat{C}^+$.

Concretely, for a given reference family:

- **Half-Normal:** estimate $\widehat{\sigma}$ by moment matching, $\widehat{\sigma} = \bar{S}\sqrt{\pi/2}$.
- **Gamma:** estimate $(\widehat{a}, \widehat{\theta})$ by method-of-moments, $\widehat{a} = \bar{S}^2/\widehat{\text{Var}}(S)$, $\widehat{\theta} = \widehat{\text{Var}}(S)/\bar{S}$.

- **Log-Normal:** estimate $(\widehat{\mu}, \widehat{\sigma})$ from $\log S$, $\widehat{\mu} = \overline{\log S}$, $\widehat{\sigma} = \mathrm{sd}(\log S)$.
- **Folded Student-t:** fix ν and estimate scale by matching the median of $|t_\nu|$ to $\mathrm{median}(S)$.

We then construct $\widehat{C}_{\mathrm{proj}}(\widehat{Q})$ using the projected-quantile rule. Results are shown in Figure 13.

Across all settings, SCP exhibits the expected discretization effects in high-coverage regimes, including vacuous sets when the target rank exceeds the calibration resolution. In contrast, $\widehat{C}^-$ provides a certified conservative baseline that remains admissible for all (m, α, β). The projected-quantile policy $\widehat{C}_{\mathrm{proj}}$ consistently improves efficiency over $\widehat{C}^-$ by exploiting reference information *only* within the ambiguity band, while inheriting the same universal finite-sample bounds by construction (Corollary 3.11). In particular, data-driven references $\widehat{Q}$ provide an adaptive mechanism for improving set sizes without altering the distribution-free validity layer of the framework.

F.4 Choosing the Reference Distribution in Practice

The reference-anchored construction introduces a distribution over nonconformity scores that influences efficiency while preserving distribution-free coverage guarantees. In this section, we provide practical guidance for selecting and implementing reference distributions.

Role of the Reference Distribution: The reference distribution affects only the efficiency layer of the method: it determines how prediction sets are selected within the admissible region defined by the envelope construction. Importantly, the validity guarantees are unaffected by this choice. However, a well-aligned reference distribution can substantially improve predictive resolution, while a poorly specified one may lead to suboptimal efficiency.

Recommended Choices: In practice, several simple and effective strategies can be used:

- **Parametric fitting.** Fit a simple parametric model (e.g., Gaussian, log-normal, or Gamma) to the empirical distribution of calibration scores. For non-negative scores, Beta or Gamma families often provide stable approximations.
- **Empirical smoothing:** Use a smoothed version of the empirical score distribution (e.g., via kernel density estimation or histogram smoothing). This preserves the observed structure while avoiding discretization artifacts.
- **Auxiliary data:** When additional unlabeled or historical data are available, their score distributions can be used to construct the reference, provided they are reasonably aligned with the calibration data.
- **Domain-informed priors:** In applications with prior knowledge (e.g., expected noise scales or tail behavior), parametric families can be chosen to reflect this structure.

Even under misspecification, the method remains valid but may lose efficiency. In practice, simple parametric or smoothed empirical references often perform well. When uncertainty about the reference is high, conservative choices (e.g., broader distributions) can reduce the risk of over-aggressive predictions.

The reference distribution should be viewed as a flexible tool for improving efficiency within a guaranteed-safe region. Simple, data-driven choices are typically sufficient, and the method is robust to moderate misspecification due to the separation of validity and efficiency layers.

Example: Small-Sample Clinical Trials. Small-sample clinical trials, such as rare disease or early-phase studies, often suffer from insufficient calibration data, causing SCP to produce vacuous or overly conservative prediction sets due to limited rank resolution. Our framework addresses this by leveraging external information, such as historical trials, registries, or real-world evidence, to construct a reference distribution over nonconformity scores, as increasingly supported by regulatory guidance (e.g., U.S. FDA (U.S. Food and Drug Administration, 2023)). Unlike existing borrowing methods (Zhu et al., 2024; Gao et al., 2025), which operate at the data or model level, our approach borrows *rank resolution*. The reference distribution improves efficiency by refining quantile selection within the admissible region, while the envelope construction preserves finite-sample,

distribution-free validity regardless of external data quality. This provides a principled mechanism for incorporating external information without requiring exchangeability, making it well-suited to settings where external data are informative but not fully aligned with the target population.

G Real Data Analysis

G.1 California Housing Data

We evaluate the proposed *projected-quantile* set $\widehat{C}_{\mathrm{proj}}$ and the *envelope-mixing* set $\widehat{C}_{\lambda}$ against standard split conformal prediction (SCP) in a real regression task. The experiment is designed to stress *small calibration regimes* and *high-coverage targets*, where SCP often becomes conservative or vacuous.

Dataset. We use the California Housing dataset (via `sklearn.datasets.fetch_california_housing`), which contains $n = 20{,}640$ observations with $d = 8$ real-valued covariates (block group-level features from the 1990 California census) and a continuous target Y representing the median house value (in units of \$100,000). Let $Y_{\max}$ and $Y_{\min}$ denote the maximum and minimum target values; we define the *vacuous* interval length as the empirical target range

$$\mathrm{Range}(Y) \;:=\; Y_{\max} - Y_{\min}.$$

Base predictor and score. We train a `RandomForestRegressor` with 200 trees (default `sklearn` settings, `n_jobs=-1`) on a random subset of size $n_{\mathrm{train}} = 2000$ (seed fixed at `SEED`=2026). Let $\widehat{f}$ denote the trained model. On the remaining data we form a *pool* of absolute residual scores

$$S \;=\; |Y - \widehat{f}(X)|.$$

In each trial we sample (without replacement) a calibration set of size m and a test set of size $n_{\mathrm{test}} = 1000$ from this residual pool.

Calibration sizes and target coverages. We consider calibration sizes

$$m \in \{10, 20, 30, 40, 50, 60\}$$

and coverage targets $\tau = 1 - \alpha$ with

$$\alpha \in \{0.07,\, 0.03,\, 0.01\} \quad \Longleftrightarrow \quad \tau \in \{0.93,\, 0.97,\, 0.99\}.$$

For each (m, α) we repeat the procedure for $T = 500$ Monte Carlo trials.

Methods compared. Each method produces a symmetric prediction interval for a test point, of the form

$$\widehat{C}(x) = \big[\widehat{f}(x) - q,\; \widehat{f}(x) + q\big],$$

so that the interval length is $\mathrm{Len} = 2q$ (or $\mathrm{Range}(Y)$ if vacuous).

1. **Split conformal prediction (SCP)**
2. **Inner envelope** $\widehat{C}^{-}$
3. **Envelope mixing** $\widehat{C}_{\lambda}$

Table 6 summarizes mean gains (relative to SCP), mean coverage, and Monte Carlo SEs. Across all settings, the certified baseline $\widehat{C}^{-}$ produces substantial length reductions compared to SCP, particularly in the smallest calibration regimes. For example, at $m = 10$ the certified baseline reduces average length by roughly 50% across $\tau \in \{0.93, 0.97, 0.99\}$ (with SE $\approx 1\%$), reflecting that SCP is frequently vacuous (length $\mathrm{Range}(Y)$) at high confidence.

The fitted-reference projected rule $\widehat{C}_{\mathrm{proj}}(\widehat{Q})$ yields gains close to the certified baseline for moderate targets ($\tau = 0.93$) and remains competitive at $\tau = 0.97$, but its gains diminish as τ approaches 0.99 (e.g., at $m = 10$, gain drops from 49.65% at $\tau = 0.93$ to 43.06% at $\tau = 0.99$). This is consistent

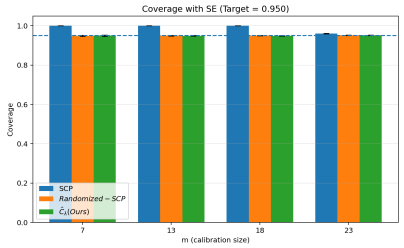
Table 6: California Housing: Mean Gain (%) relative to SCP and empirical Coverage (mean ± SE over $T = 500$ trials).

m	Method	$\tau = 0.93$	$\tau = 0.97$	$\tau = 0.99$
10	$\widehat{C}^{-}$	Gain: 50.60±1.06	50.00±1.11	50.85±0.96
		Cov.: 0.9066±0.0039	0.9069±0.0037	0.9095±0.0036
	$\widehat{C}_{\lambda}$	Gain: 39.05±1.31	17.13±1.25	6.45±0.84
		Cov.: 0.9281±0.0039	0.9674±0.0029	0.9879±0.0021
	$\widehat{C}_{\mathrm{proj}}$	Gain: 49.65±1.04	46.74±1.05	43.06±0.86
		Cov.: 0.9141±0.0035	0.9268±0.0028	0.9454±0.0020
	SCP	Gain: 0.00	0.00	0.00
		Cov.: 1.0000	1.0000	1.0000
20	$\widehat{C}^{-}$	Gain: 27.16±0.84	34.68±1.15	34.18±1.19
		Cov.: 0.9033±0.0029	0.9556±0.0020	0.9554±0.0020
	$\widehat{C}_{\lambda}$	Gain: 11.62±0.81	23.15±1.17	5.58±0.78
		Cov.: 0.9311±0.0026	0.9712±0.0018	0.9923±0.0011
	$\widehat{C}_{\mathrm{proj}}$	Gain: 25.09±0.80	34.11±1.13	31.46±1.10
		Cov.: 0.9120±0.0025	0.9585±0.0017	0.9653±0.0013
	SCP	Gain: 0.00	0.00	0.00
		Cov.: 0.9514±0.0021	1.0000	1.0000
30	$\widehat{C}^{-}$	Gain: 0.00	29.95±1.15	28.11±1.19
		Cov.: 0.9346±0.0021	0.9662±0.0014	0.9673±0.0014
	$\widehat{C}_{\lambda}$	Gain: 0.00	27.22±1.17	10.36±0.88
		Cov.: 0.9346±0.0021	0.9691±0.0014	0.9889±0.0010
	$\widehat{C}_{\mathrm{proj}}$	Gain: 0.00	29.89±1.15	26.99±1.14
		Cov.: 0.9346±0.0021	0.9665±0.0013	0.9712±0.0011
	SCP	Gain: 0.00	0.00	0.00
		Cov.: 0.9346±0.0021	1.0000	1.0000
40	$\widehat{C}^{-}$	Gain: 16.69±0.58	0.00	22.27±1.18
		Cov.: 0.9292±0.0017	0.9749±0.0011	0.9754±0.0011
	$\widehat{C}_{\lambda}$	Gain: 16.69±0.58	0.00	8.30±0.88
		Cov.: 0.9292±0.0017	0.9749±0.0011	0.9907±0.0009
	$\widehat{C}_{\mathrm{proj}}$	Gain: 16.69±0.58	0.00	21.71±1.15
		Cov.: 0.9292±0.0017	0.9749±0.0011	0.9771±0.0009
	SCP	Gain: 0.00	0.00	0.00
		Cov.: 0.9533±0.0015	0.9749±0.0011	1.0000
50	$\widehat{C}^{-}$	Gain: 0.00	0.00	17.08±1.17
		Cov.: 0.9430±0.0014	0.9806±0.0009	0.9807±0.0009
	$\widehat{C}_{\lambda}$	Gain: 0.00	0.00	7.05±0.89
		Cov.: 0.9430±0.0014	0.9806±0.0009	0.9915±0.0007
	$\widehat{C}_{\mathrm{proj}}$	Gain: 0.00	0.00	16.93±1.16
		Cov.: 0.9430±0.0014	0.9806±0.0009	0.9812±0.0008
	SCP	Gain: 0.00	0.00	0.00
		Cov.: 0.9430±0.0014	0.9806±0.0009	1.0000

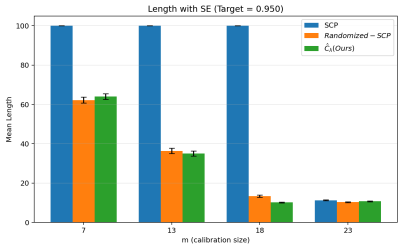
with the theorem: $\widehat{C}_{\mathrm{proj}}$ differs from $\widehat{C}^{-}$ only on the ambiguity band, and when τ is very high the projection constraint $\Phi_{\widehat{Q}}(s) \leq \tau$ becomes less restrictive, pushing the rule closer to the forced-in behavior.

The envelope-mixing rule $\widehat{C}_\lambda$ exhibits a different trade-off: it achieves *near-target* empirical coverage (often close to τ when SCP is vacuous), but typically with smaller efficiency gains than $\widehat{C}^-$ and $\widehat{C}_{\mathrm{proj}}$. For instance, at $m = 10$ its gains are 39.05%, 17.13%, and 6.45% for $\tau = 0.93, 0.97, 0.99$ respectively, while empirical coverages are 0.928, 0.967, and 0.988, close to the target levels. This behavior is expected: $\widehat{C}_\lambda$ randomizes between ranks r_- and r_+ to interpolate the attainable coverage interval $\left[\frac{r_-}{m+1}, \frac{r_+}{m+1}\right]$, which can substantially reduce conservatism when SCP is forced to be vacuous, but may also increase length relative to $\widehat{C}^-$ when r_+ is much larger than r_-.

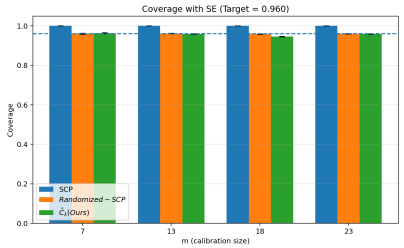
Overall, this real-data study supports two key messages: (i) *rank-resolution limitations* can make SCP vacuous in small-m/high-τ regimes, and the certified baseline $\widehat{C}^-$ substantially mitigates this effect; and (ii) *reference anchoring via projection* can preserve the certified structure while offering additional flexibility on the ambiguity band, whereas *envelope mixing* provides a principled way to tune achieved coverage toward the nominal target when discretization forces unavoidable gaps.



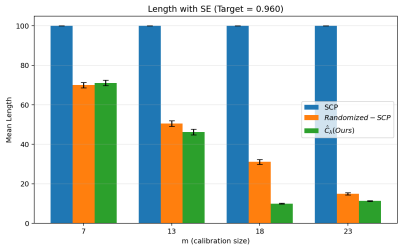
(a) Coverage ($\tau = 0.95$)



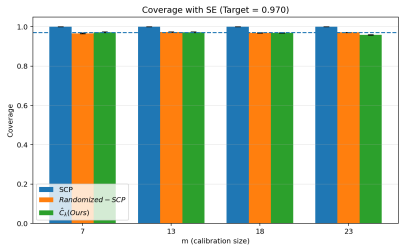
(b) Length ($\tau = 0.95$)



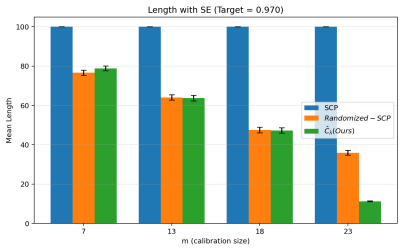
(c) Coverage ($\tau = 0.96$)



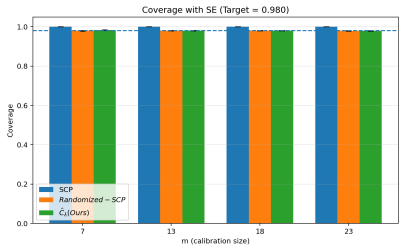
(d) Length ($\tau = 0.96$)



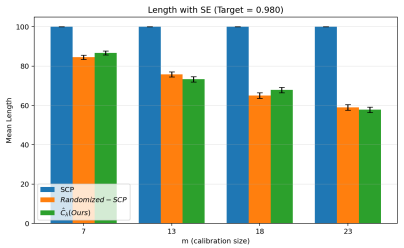
(e) Coverage ($\tau = 0.97$)



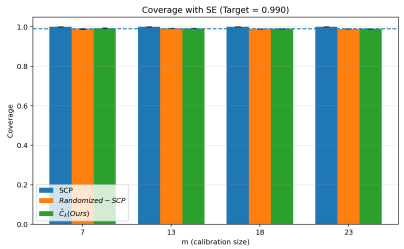
(f) Length ($\tau = 0.97$)



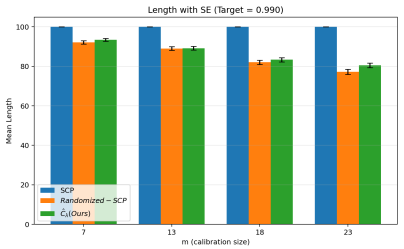
(g) Coverage ($\tau = 0.98$)



(h) Length ($\tau = 0.98$)



(i) Coverage ($\tau = 0.99$)



(j) Length ($\tau = 0.99$)

Figure 10: Additional heteroskedastic score experiment: Each row corresponds to a fixed target coverage level τ. Left: empirical coverage with standard errors. Right: mean prediction-set length with standard errors. SCP is often vacuous in the small-calibration regime, Randomized-SCP improves by local rank interpolation, and $\widehat{C}_\lambda$ can provide additional efficiency gains by mixing over the full admissible rank region.

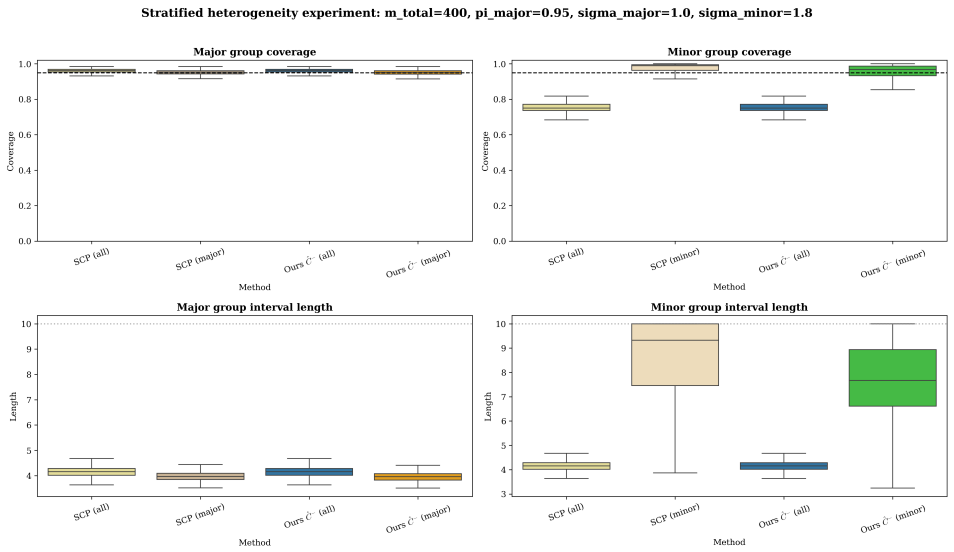
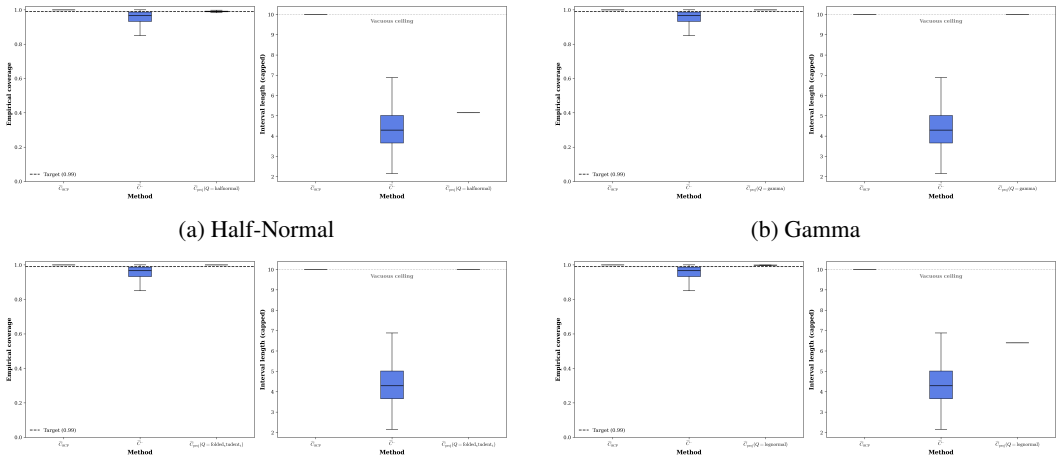


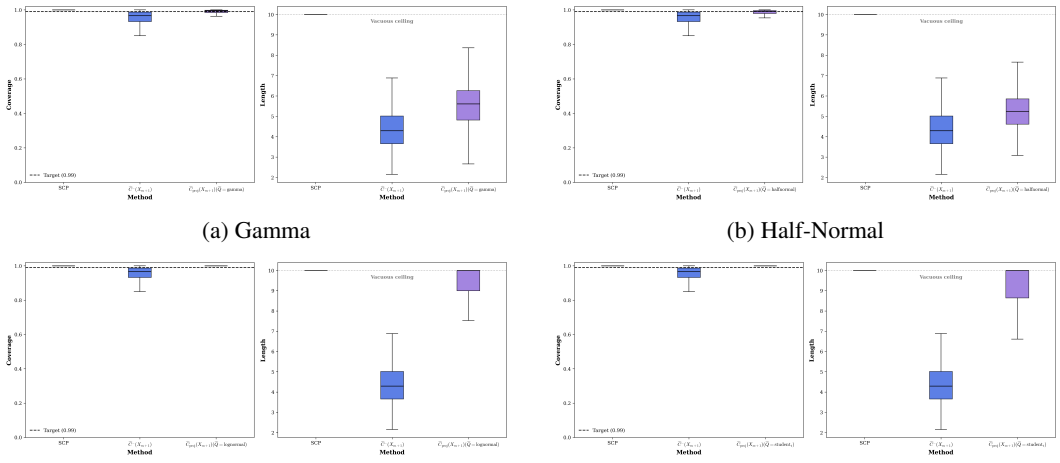
Figure 11: **Conditional Efficiency Gains via Stratified Envelopes. Top Row:** Pooled methods (khaki) fail to cover the high-variance minor group, whereas stratification (blues/oranges) restores validity. **Bottom Row:** For the underrepresented minor group, standard stratified SCP (wheat) suffers from discretization padding, producing overly conservative intervals. Our stratified envelopes (green) yield near-exact target coverage with significantly shorter intervals, demonstrating the resolution dividend for rare subgroups.



(a) Half-Normal (b) Gamma

(c) Folded Student-t (d) Log-Normal

Figure 12: **Projected-quantile conformal prediction with fixed references** ($m = 20$, $\alpha = 0.01$)**.** Boxplots of empirical coverage (left panels) and prediction set length (right panels) for SCP, the certified baseline $\widehat{C}^{-}$, and the projected-quantile rule $\widehat{C}_{\mathrm{proj}}(Q)$ under different fixed reference distributions Q. The dashed horizontal line denotes the target coverage $1 - \alpha$, and the dotted line indicates the vacuous ceiling B. Across all choices of Q, projection influences efficiency only within the ambiguity band, while distribution-free validity is preserved by the rank-envelope constraints.



(a) Gamma

(b) Half-Normal

(c) Log-Normal

(d) Folded Student-t

Figure 13: **Projected-quantile conformal prediction with fitted references** ($m = 20$, $\alpha = 0.01$)**.** Boxplots of empirical coverage and prediction set length for SCP, the certified baseline $\widehat{C}^{-}$, and the projected-quantile rule $\widehat{C}_{\mathrm{proj}}(\widehat{Q})$ under different fitted reference distributions $\widehat{Q}$. The reference is used only within the ambiguity band, while distribution-free validity is enforced by the rank-envelope constraints. Performance varies with reference quality, but coverage remains controlled across all choices.

G.2 DIABETES PROGRESSION DATASET

We repeat the small-calibration, high-coverage evaluation on the DIABETES dataset, comparing SCP, the certified baseline $\widehat{C}^{-}$, the projected-quantile set $\widehat{C}_{\mathrm{proj}}(\widehat{Q})$ with a fitted reference, and the envelope-mixing rule $\widehat{C}_{\lambda}$.

We use the DIABETES dataset (via `sklearn.datasets.load_diabetes`), containing $n = 442$ observations and $d = 10$ standardized covariates (age, sex, body mass index, blood pressure, and six serum measurements). The response Y is a quantitative measure of disease progression one year after baseline. We define the vacuous interval length as the empirical target range

$$\mathrm{Range}(Y) \;:=\; Y_{\max} - Y_{\min},$$

which equals 321 in our implementation (and therefore SCP becomes visibly "flat" when it is vacuous).

We fit a `GradientBoostingRegressor` with 200 estimators (seed 42) on a fixed training subset of size $n_{\mathrm{train}} = 300$. Let $\widehat{f}$ denote the trained model. On the remaining holdout pool we compute absolute residual scores

$$S \;=\; |Y - \widehat{f}(X)|.$$

We consider $m \in \{10, 15, 20, 25, 30\}, \tau \in \{0.97, 0.99\}$. For each (m, α) we run $T = 500$ trials, and within each trial we evaluate coverage on $n_{\mathrm{test}} = 100$ freshly sampled test residuals. We compare the methods as done in section G.1.

Table 7 summarizes performance for $\tau \in \{0.97, 0.99\}$. As in the California Housing experiment, SCP is frequently vacuous at high confidence when m is small, leading to a constant interval length of $\mathrm{Range}(Y) = 321$ and empirical coverages near 1.0.

The certified baseline $\widehat{C}^{-}$ delivers consistent length reductions relative to SCP, with gains decreasing smoothly as calibration size grows. The projected-quantile method $\widehat{C}_{\mathrm{proj}}(\widehat{Q})$ yields intermediate gains: it improves substantially over SCP but is generally less aggressive than $\widehat{C}^{-}$. For example, at $m = 10$ it achieves 21.0% gain at $\tau = 0.97$ and 7.9% at $\tau = 0.99$. This pattern matches the theorem-level behavior: $\widehat{C}_{\mathrm{proj}}$ differs from $\widehat{C}^{-}$ only on the ambiguity band, and when τ is extremely high the projection constraint becomes less influential, so improvements over $\widehat{C}^{-}$ are limited unless the fitted reference provides a strong additional cutoff. A key qualitative distinction appears in coverage. The certified baseline $\widehat{C}^{-}$ is intentionally conservative in this small-m regime: its observed coverages can fall below τ (e.g., at $m = 10$, coverage ≈ 0.909 for $\tau = 0.97$ and ≈ 0.914 for $\tau = 0.99$), reflecting the fact that it is a *forced-in* construction designed to avoid vacuity and control behavior uniformly across ranks, rather than to match the nominal target in finite samples. In contrast, envelope mixing $\widehat{C}_{\lambda}$ achieves near-nominal empirical coverage: for $m = 10$ it attains 0.972 at $\tau = 0.97$ and 0.9905 at $\tau = 0.99$, with small SEs. This is expected since $\widehat{C}_{\lambda}$ interpolates between the attainable rank coverages $r_{-}/(m+1)$ and $r_{+}/(m+1)$, effectively calibrating to the discretized coverage grid induced by small m. The projected-quantile method sits between these extremes: coverage is higher than $\widehat{C}^{-}$ but still often below τ at very small m (e.g., 0.9468 at $m = 10, \tau = 0.97$), consistent with the fact that it only tightens/relaxes decisions within the ambiguity band by comparing $\Phi_{\widehat{Q}}(s)$ to τ.

Overall, the Diabetes experiment reinforces the same message as California Housing: (i) in high-confidence, small-calibration regimes, SCP can become vacuous and overly conservative; (ii) the certified baseline $\widehat{C}^{-}$ provides large efficiency gains but may undershoot τ empirically; (iii) envelope mixing $\widehat{C}_{\lambda}$ targets nominal coverage more closely, trading away some efficiency; and (iv) projected-quantile anchoring $\widehat{C}_{\mathrm{proj}}(\widehat{Q})$ provides a principled middle ground whose benefit depends on how informative the fitted reference $\widehat{Q}$ is for the tail behavior of the score distribution.

Table 7: DIABETES: Efficiency gain (%) relative to SCP and empirical coverage. Each entry reports **Gain**$\pm$**SE** (top) and **Cov.**$\pm$**SE** (bottom) over $T = 500$ trials.

m	Method	$\tau = 0.97$	$\tau = 0.99$
10	$\widehat{C}^{-}$	32.08±1.01 0.9085±0.0039	28.96±1.08 0.9141±0.0040
	$\widehat{C}_{\lambda}$	9.72±0.87 0.9720±0.0029	3.38±0.55 0.9905±0.0019
	$\widehat{C}_{\mathrm{proj}}$	20.99±0.89 0.9468±0.0023	7.86±1.05 0.9678±0.0017
	SCP	0.00 1.0000	0.00 1.0000
15	$\widehat{C}^{-}$	23.39±0.97 0.9418±0.0028	23.04±1.00 0.9412±0.0028
	$\widehat{C}_{\lambda}$	12.81±0.87 0.9682±0.0025	3.91±0.57 0.9902±0.0015
	$\widehat{C}_{\mathrm{proj}}$	17.18±0.81 0.9606±0.0017	6.26±0.89 0.9748±0.0014
	SCP	0.00 1.0000	0.00 1.0000
20	$\widehat{C}^{-}$	18.43±0.93 0.9526±0.0024	18.62±0.96 0.9546±0.0022
	$\widehat{C}_{\lambda}$	12.20±0.86 0.9688±0.0021	3.44±0.55 0.9916±0.0012
	$\widehat{C}_{\mathrm{proj}}$	14.46±0.80 0.9636±0.0016	6.42±0.76 0.9768±0.0012
	SCP	0.00 1.0000	0.00 1.0000
25	$\widehat{C}^{-}$	13.99±0.87 0.9640±0.0019	13.75±0.87 0.9645±0.0019
	$\widehat{C}_{\lambda}$	11.34±0.83 0.9703±0.0019	3.85±0.54 0.9895±0.0014
	$\widehat{C}_{\mathrm{proj}}$	11.42±0.77 0.9701±0.0015	4.45±0.70 0.9797±0.0012
	SCP	0.00 1.0000	0.00 1.0000
30	$\widehat{C}^{-}$	10.69±0.88 0.9707±0.0015	11.33±0.85 0.9691±0.0017
	$\widehat{C}_{\lambda}$	10.03±0.85 0.9731±0.0015	3.59±0.51 0.9913±0.0011
	$\widehat{C}_{\mathrm{proj}}$	9.04±0.81 0.9740±0.0013	3.80±0.69 0.9805±0.0011
	SCP	0.00 1.0000	0.00 1.0000

Note: SCP often overcovers in this small-m/high-τ regime, hence gains are relative to SCP length.

G.3 IMAGE CLASSIFICATION: CIFAR-10

We evaluate our methods on the CIFAR-10 image classification benchmark, consisting of $K = 10$ object categories. We use a pretrained `cifar10_resnet20` model from Torch Hub, available at `https://github.com/chenyaofo/pytorch-cifar-models`, (Top-1 accuracy: 92.6%) and treat the CIFAR-10 test set as a fixed pool of nonconformity scores. This allows us to isolate the effect of small calibration size without retraining the model.

For each image x and label y, we define the nonconformity score $s(x, y) = 1 - p_y(x)$, where $p_y(x)$ is the predicted class probability. Scores are precomputed once and reused across trials. For each trial, we sample m calibration points and evaluate on a disjoint test subset. We consider calibration sizes $m \in \{5, 10, 15, 20, 25\}$ and target coverage levels $\tau \in \{0.95, 0.96, 0.97, 0.98, 0.99\}$, with $T = 500$ repetitions.

Results: Figure 14 summarizes empirical coverage and prediction set size across target levels. The results reveal a consistent pattern across all τ:

(i) Severe overcoverage of SCP in small-m regimes. For small calibration sizes ($m \leq 15$), standard SCP exhibits extreme overcoverage, often achieving coverage ≈ 1.0 regardless of the target level. This leads to vacuous prediction sets with maximal size (equal to the full label set).

(ii) Envelope mixing restores calibration with large efficiency gains. The proposed method $\widehat{C}_\lambda$ consistently achieves coverage close to the nominal target while significantly reducing set size. For example:

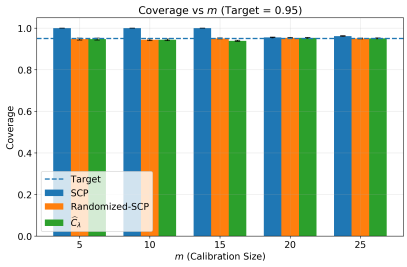
- At $\tau = 0.95$, $m = 10$: SCP achieves coverage 1.00 with set size 10, whereas $\widehat{C}_\lambda$ attains coverage 0.944 with set size 5.23 (a $\sim 48\%$ reduction).
- At $\tau = 0.99$, $m = 10$: SCP again yields coverage 1.00, while $\widehat{C}_\lambda$ achieves 0.989 with $\sim 10\%$ reduction in set size.

Across all configurations, $\widehat{C}_\lambda$ maintains near-nominal coverage while delivering substantial efficiency gains, especially in high-confidence regimes.

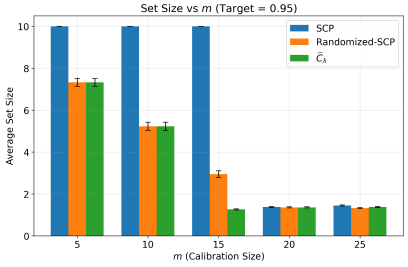
(iii) Randomized (Smoothed) SCP partially mitigates overcoverage: Randomized-SCP (Fontana et al., 2023) corrects the discreteness of rank thresholds via interpolation between adjacent ranks, restoring coverage close to the nominal level. In our framework, this corresponds to the special case where the admissible rank set collapses to two neighboring indices, in which case our envelope-mixing rule reduces exactly to randomized conformal prediction.

However, when the admissible region spans more than two ranks, our method performs mixing over the full feasible set rather than local interpolation. This enables additional efficiency gains beyond Randomized-SCP while still maintaining finite-sample validity. Empirically, we observe that Randomized-SCP improves over SCP, but $\widehat{C}_\lambda$ achieves further reductions in set size, particularly in small-sample, high-confidence regimes. Empirically, this distinction is visible in our CIFAR-10 experiments: for example, at $\tau = 0.99$ and $m = 10$, Randomized-SCP achieves near-nominal coverage with a moderate reduction in set size, whereas $\widehat{C}_\lambda$ attains comparable coverage while yielding further reduction in average set size. Similar improvements are observed across high-confidence regimes ($\tau \geq 0.97$), where the admissible region is wider.

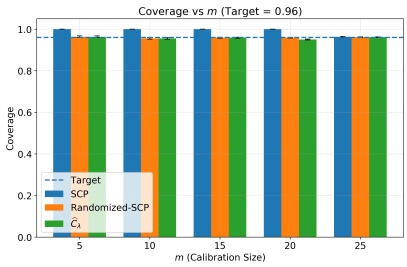
These results confirm the central premise of our framework: the inefficiency of SCP in small-sample regimes arises from unresolved rank uncertainty. By exploiting the admissible rank region, $\widehat{C}_\lambda$ recovers substantial predictive resolution while preserving distribution-free validity, providing a robust improvement over standard conformal prediction.



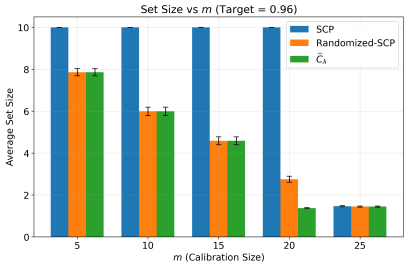
(a) Coverage ($\tau = 0.90$)



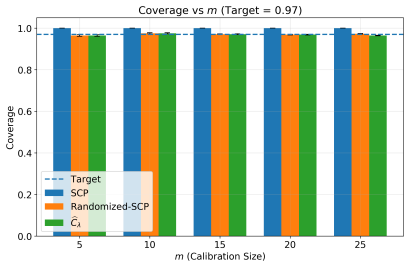
(b) Set size ($\tau = 0.90$)



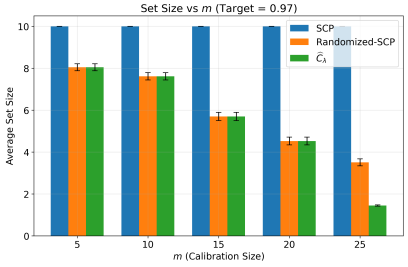
(c) Coverage ($\tau = 0.92$)



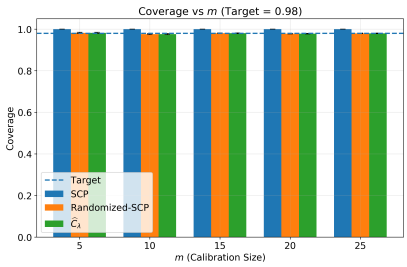
(d) Set size ($\tau = 0.92$)



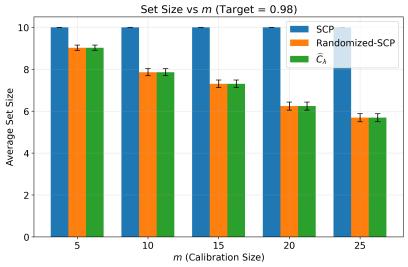
(e) Coverage ($\tau = 0.95$)



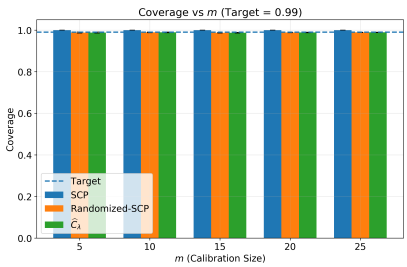
(f) Set size ($\tau = 0.95$)



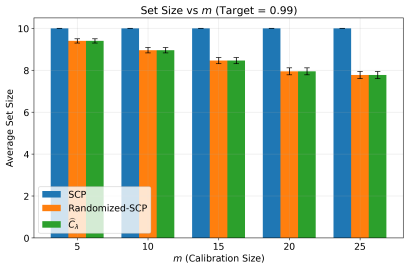
(g) Coverage ($\tau = 0.97$)



(h) Set size ($\tau = 0.97$)



(i) Coverage ($\tau = 0.99$)



(j) Set size ($\tau = 0.99$)

Figure 14: **CIFAR-10 classification experiments across target coverage levels.** Each row corresponds to a fixed target τ. Left: empirical coverage. Right: average prediction set size. Results are shown for SCP, Randomized-SCP, and $\widehat{C}_\lambda$ across calibration sizes m. Dashed lines denote the target coverage. The proposed method maintains coverage while achieving consistent efficiency gains in small-sample regimes.