

Stochastic Grouping Conformal Prediction for Effective Subgroup Reliability

Meihui Zhong¹, Wenxin Tai¹, Ting Zhong^{1,3}, Fan Zhou^{1,2,3*}

¹ University of Electronic Science and Technology of China

² Intelligent Digital Media Technology, Key Laboratory of Sichuan Province

³ Kashi Institute of Electronics and Information Industry

Abstract

Conformal prediction offers a distribution-free coverage guarantee, making it especially attractive for clinical applications. Standard conformal prediction, however, provides such guarantees only at the population level, and its prediction sets can exhibit coverage disparities across clinically important subgroups. A natural remedy is to calibrate within predefined groups. However, this can require access to sensitive subgroup attributes and is prone to a worst-group bottleneck: protecting the most difficult subgroup can inflate prediction sets for all, increasing cognitive burden on decision makers. To this end, we propose Stochastic Grouping Conformal Prediction (SGCP), a conformal framework for subgroup-reliable uncertainty quantification. It learns a stochastic grouping map that allows each sample to draw calibration information from others with similar calibration behavior, yielding a local score law that boosts reliability across subpopulations. We prove that SGCP retains the standard coverage guarantee. Experiments on synthetic and real-world benchmarks show that it consistently reduces subgroup coverage gaps while achieving smaller or comparable prediction set sizes relative to existing baselines.

1 Introduction

Conformal prediction (CP) has emerged as an attractive framework for medical AI, because it converts any predictive model into a prediction set with finite-sample coverage guarantees under minimal assumptions [17, 5, 16]. Consider training data $\{(X_i, Y_i)\}_{i=1}^n$ and a future test point (X_{n+1}, Y_{n+1}) , where each feature $X_i \in \mathcal{X}$ lies in the feature space and response $Y_i \in \mathcal{Y}$ lies in the label space. The goal is to construct a prediction set $\hat{C}_n(X_{n+1}) \subseteq \mathcal{Y}$ that contains the unknown test response Y_{n+1} with high probability. The appeal of distribution-free prediction is that this guarantee can be obtained without imposing strong assumptions on the underlying data distribution [24, 22, 1], namely,

$$\mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1})\} \geq 1 - \alpha. \quad (1)$$

While this guarantee is highly appealing, it has an important limitation. The coverage in Eq. (1) is averaged over the calibration data and the future test point, thus providing validity only at the population level [10, 21]. In clinical settings, this means that a nominal 90% coverage guarantee may hold on average over all patients, while still failing to provide equally reliable uncertainty estimates for particular demographic groups, disease subtypes, or other clinically meaningful subpopulations.

This concern is tied to fairness in healthcare, where the costs of unreliable uncertainty often fall disproportionately on underrepresented patient populations [9, 4]. Differences in prevalence, data quality, and feature distributions across populations can induce distinct error patterns and nonconformity score distributions [14]. Thus, population-level calibration is prone to masking subgroup failures

*Corresponding author: fan.zhou@uestc.edu.cn

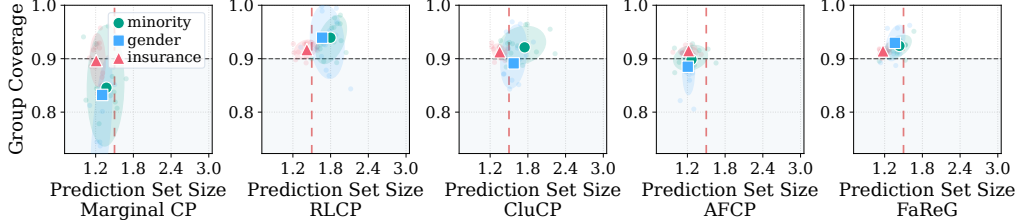


Figure 1: Tradeoff between subgroup coverage and prediction set size on MIMIC-IV [13] ($1 - \alpha = 0.9$). Results are shown for RLCP [12], CluCP [8], AFCP [29], and FaReG [27]. The red dashed line marks half of the total number of classes. Several fairness-aware CP methods improve sensitive-group coverage, but often do so with much larger prediction sets, making them less informative in practice.

by averaging under-coverage in some groups with over-coverage in others. In medical applications, this imbalance matters, as what we care about is not only whether uncertainty is valid on average, but whether it remains reliable for the subgroups for whom prediction errors are most costly.

We empirically observe exactly this tension, as shown in Figure 1. Standard conformal prediction achieves the desired overall coverage, but still exhibits lower coverage on several clinically important subpopulations, suggesting that marginal validity may ‘average out’ group-specific uncertainty. A natural response is to calibrate separately within each group. However, methods [20, 27] that target finer grained group reliability often become increasingly conservative, resulting in larger prediction sets. Such enlarged sets may be less informative in practice and may even hinder downstream decision-making by imposing excessive cognitive burden [28, 7, 6]. One may go even further and calibrate at the individual-sample level, as in localized conformal prediction [12, 20]. Yet our results show that localization, while sometimes improving coverage on subgroups, often incurs substantially larger prediction sets and, more importantly, greater instability across random seeds and data splits.

In sum, these observations pose two coupled challenges. First, the heterogeneity that matters for calibration is often latent: it may *not be* directly observable at inference, nor align with pre-specified sensitive groups. Second, the groups that most need tailored calibration are often those with *limited* sample sizes, making fine grained localized calibration conservative and statistically unstable [8]. More broadly, we seek a practical middle ground between global calibration, which can wash out subgroup-specific uncertainty, and calibration strategies that rely too heavily on known groups or local neighborhoods, which can become conservative or unstable when calibration data are limited.

Motivated by this gap, we propose **stochastic grouping conformal prediction (SGCP)**, a fairness-aware conformal prediction method that improves subgroup reliability without relying on known sensitive attributes. Rather than asking coverage to hold for each individual test point, we relax the requirement to a weaker but still meaningful sense – coverage over samples that exhibit similar calibration behavior. In other words, exact 90% coverage may be unattainable for one patient in isolation, yet it may still be possible to achieve more reliable coverage on patients whose scores reflect similar uncertainty patterns and calibration difficulty.

To achieve this, SGCP learns a latent calibration structure that organizes samples according to uncertainty-relevant score behavior. Atop this structure, it learns a *stochastic grouping map* that determines how each sample should draw calibration information from a small set of latent calibration components. This yields a sample-adaptive local score law: calibration is tailored to each sample, but estimated through structured sharing rather than from a small group or neighborhood alone. Statistically, this design strikes a balance between flexibility and stability. It is flexible enough to capture heterogeneous score behavior across samples, yet structured enough to be reliably estimated from finite data while preserving conformal validity. As a result, SGCP can improve subgroup reliability beyond global calibration without relying on sparse subgroup calibration data. Empirically, across fairness-sensitive clinical and public prediction tasks, SGCP improves coverage on clinically meaningful subgroups while maintaining valid overall coverage and efficient prediction sets.

Our contributions are summarized as follows. First, we propose SGCP, a novel fairness-aware conformal prediction method. It learns a local score law through stochastic grouping, thereby approximating the score distribution of test points. This improves subgroup reliability without relying on rigid predefined groups. Second, we prove that SGCP retains finite-sample marginal validity. We further show that subgroup coverage gaps decrease when the learned local score law better captures

the score behavior of meaningful subpopulations. Third, through experiments on synthetic and medical datasets, we demonstrate that SGCP achieves a favorable tradeoff among overall coverage, subgroup reliability, and prediction set efficiency. We expect SGCP to enable more effective and trustworthy decision-making in high-risk medical applications. Code and supplementary materials are included in the anonymous supplementary upload.

2 Preliminaries

We consider multiclass conformal classification with an exchangeable dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y} = \{1, \dots, C\}$. Let $f_\theta : \mathcal{X} \rightarrow \Delta^{C-1}$ be a base classifier, with $f_\theta^y(x)$ denoting the predicted probability of class y . We use the nonconformity score $s(x, y) = 1 - f_\theta^y(x)$. For an observed example (X_i, Y_i) , we write $s_i = s(x_i, y_i)$, and it measures how inconsistent a point is with the true label. Given a training set $\mathcal{D}_{\text{tra}}$ and a calibration set $\mathcal{D}_{\text{cal}}$, we are interested in constructing, for a test point X_{n+1} , a confidence set $\hat{C}(x) \subseteq \mathcal{Y}$ such that $\mathbb{P}\{Y_{n+1} \in \hat{C}(X_{n+1})\} \geq 1 - \alpha$, for a miscoverage level α , without distributional assumptions beyond exchangeability.

In fairness-sensitive applications, the central issue is not merely whether the prediction set is valid on average, but rather how coverage and set size are distributed across heterogeneous subpopulations. Ideally, one desire distribution-free *conditional* coverage – that is,

$$\mathbb{P}\{Y \in \hat{C}(x) \mid X = x\} \geq 1 - \alpha \quad P_X\text{-a.e. } x. \quad (2)$$

which would mitigate uneven coverage across patient demographics, disease patterns, and other meaningful groups. However, non-trivial distribution-free conditional coverage is impossible, since finite samples cannot ensure tight, valid prediction sets for arbitrary conditioning events [10, 12].

We typically classify existing approaches into two broad types. **(i) Group-wise calibration.** A natural response to subgroup miscoverage is to calibrate separately across groups [17, 25], a type we term *Group-wise calibration* CP. If $G(x) \in \{1, \dots, K\}$ denotes a group label, this yields a group-specific threshold $\hat{q}_k = \text{Quantile}\left(\{s_i : (x_i, y_i) \in \mathcal{D}_{\text{cal}}, G(x_i) = k\}, 1 - \alpha\right)$, and outputs

$$\hat{C}_{\text{GW}}(x) = \{y \in \mathcal{Y} : s(x, y) \leq \hat{q}_{G(x)}\}. \quad (3)$$

In principle, Eq. (3) reduces the mismatch induced by a single global threshold. In practice, however, one issue is that the relevant groups are often not known a priori, and may not align with pre-specified sensitive attributes. Another is that finer grouping fragments the calibration set, leaving some groups with too few samples for calibration, and making the empirical quantiles $\hat{q}_k$ statistically unstable [8].

This group-wise perspective also includes methods that learn or adapt groups before calibration. AFCP [29], for example, considers groups defined by sensitive attributes and makes the group selection itself adaptive. Let $\phi(x, \mathcal{A})$ denote the projection of x onto a selected sensitive-attribute set $\mathcal{A}$. For instance, a subgroup “males with a bachelor’s degree” can be written as $\phi(x, \{k, \ell\}) \in [M_k] \times [M_\ell]$. For each test point, AFCP selects either one sensitive attribute or none, $\hat{A}(X_{n+1}) \in \{\emptyset, \{1\}, \dots, \{K\}\}$, and targets $\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \phi(X_{n+1}, \hat{A}(X_{n+1}))\right] \geq 1 - \alpha$. This reduces to marginal coverage when $\hat{A}(X_{n+1}) = \emptyset$. Operationally, AFCP combines a marginal conformal set with an equalized-coverage set for each candidate label under the selected attribute. Although less conservative than exhaustive calibration over all attribute combinations, AFCP remains tied to discrete sensitive attributes. Its repeated leave-one-out calibration on pseudo-test points can also be computationally expensive and noisy in sparse-group regimes.

FAREG [27] further improves group discovery by learning groups in a latent representation space rather than enumerating them from raw input features. This expands the class of detectable subgroups, including those defined by nonlinear feature combinations. However, FAREG still frames fairness as subgroup discovery: it identifies explicit subgroups with coverage disparity and enforces coverage over them. It is therefore most suitable when heterogeneity is well captured by identifiable subgroups. It may, however, become conservative in some cases.

(ii) Localized calibration. Localized calibration, as in RLCP [12], avoids explicit group partitions by using nearby samples as the reference set for each test point. RLCP argues that heterogeneity relevant to calibration may be highly sample-specific. Let $\mathfrak{X}_x = \{\mathcal{X} \in \mathfrak{X} : x \in \mathcal{X}\}$ denote

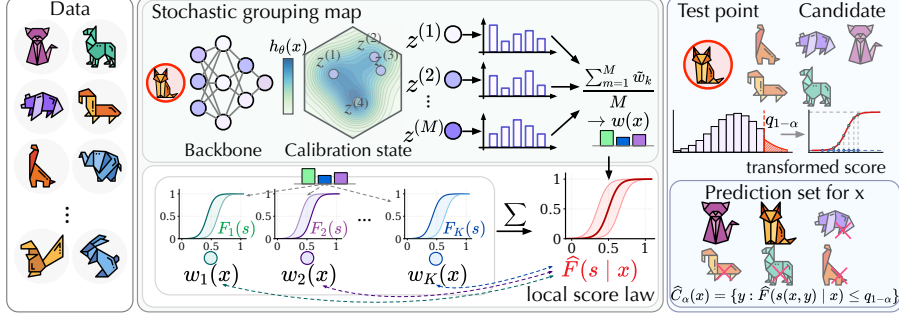


Figure 2: Overview of SGCP. SGCP learns a local score law by mapping x to a stochastic grouping map $w(x)$ over latent calibration components $\{F_k\}_{k=1}^K$. Candidate labels are transformed by this local law and kept when their calibrated scores fall below the conformal threshold.

the candidate neighborhoods containing x . A localized threshold can be written as $\hat{q}_{\text{loc}}(x) = \sup_{\mathcal{X} \in \mathcal{X}_x} \text{Quantile}(\{s_i : (x_i, y_i) \in \mathcal{D}_{\text{cal}}, x_i \in \mathcal{X}\}, 1 - \alpha)$, with prediction set

$$\hat{C}_{\text{loc}}(x) = \{y \in \mathcal{Y} : s(x, y) \leq \hat{q}_{\text{loc}}(x)\}. \quad (4)$$

Compared with group-wise calibration, Eq. (4) avoids rigid group boundaries and adapts to the region around the test point. Yet the supremum over candidate neighborhoods requires the prediction set to cover the worst local subset containing x , which may be overly conservative and yield unnecessarily large sets, particularly when calibration samples are scarce. Moreover, locality is typically defined via feature space similarity, which need not align with the actual error structure most relevant for calibration. RLCP may introduces additional randomness into the prediction intervals, leading to inconsistent outputs across independent runs. Other related works are discussed in Appendix F.

The preceding discussion suggests that the key challenge is to improve subgroup reliability without losing statistical stability or producing overly large prediction sets. We therefore seek a method that can capture calibration heterogeneity while promoting informative prediction sets.

3 Method

3.1 From subgroup reliability to local score laws

Our starting point is that conformal calibration is ultimately governed by the distribution of the nonconformity score. Accordingly, we take the sample-specific score distribution $F_x^*(s) := \mathbb{P}(S \leq s \mid X = x)$, where S denotes the nonconformity score, as the oracle object for reliable calibration to each subgroup. If we could estimate F_x^* accurately, then each sample could be calibrated according to its own local score law. That said, the difficulty is that directly learning an unconstrained conditional distribution F_x^* for every x is data-hungry and unstable in finite samples, especially when the relevant heterogeneity is latent and high-dimensional. A single test point rarely provides enough local calibration evidence for reliable estimation. This motivates the central question behind SGCP: can other samples help a target point estimate its local score distribution without resorting to a single global calibration? SGCP addresses this by replacing the unconstrained family $\{F_x^* : x \in \mathcal{X}\}$ with a structured mixture family. Concretely, SGCP approximates the oracle local score law by

$$F_x^*(s) \approx \hat{F}(s \mid x) = \sum_{k=1}^K w_k(x) F_k(s), \quad (5)$$

where $F_1, \dots, F_K$ are *latent calibration components* and $w(x)$ is a sample-related *stochastic grouping map*. Thus, each sample has its own effective local score law, but that law is not estimated from scratch. It is represented as a weighted combination of latent calibration components learned from related samples. This makes SGCP more expressive than discrete group-wise calibration, while still allowing samples with similar calibration behavior to share evidence through the components. Posterior averaging further stabilizes the resulting local score law (see Appendix B for details). Figure 2 illustrates how stochastic grouping induces a local score law for each sample.

The key question in Eq. (5) is therefore how to determine, for each sample x , which other samples should be regarded as ‘related’. While there are many ways to define relevance around a point, such as finding similar points through feature embedding or clustering, recall that the ideal object for subgroup reliable calibration is the sample-specific score distribution F_x^* in Eq. (5). Accordingly, the relevant notion of locality is not simply geometric proximity in the input space, but similarity in calibration relevant score behavior. So we want to identify samples that are governed by similar score behavior. This motivates us to propose the latent calibration state and stochastic grouping map. Together, they determine how each sample should transfer calibration evidence from the latent calibration components. We next formalize these two components and the resulting local score law.

3.2 Learning a differentiable local score law

Latent calibration state. Let $h_\theta(x) \in \mathbb{R}^d$ denote the feature representation produced by the backbone classifier, and let $f_\theta(x) \in \Delta^{C-1}$ denote the corresponding predictive probability vector. To capture heterogeneity in score behavior beyond predefined groups, we associate each sample with a latent calibration state z . We model this state via a posterior distribution

$$q_\phi(z | x) = \mathcal{N}(\mu_\phi(h_\theta(x)), \text{diag}(\sigma_\phi^2(h_\theta(x)))) \quad (6)$$

where $\mu_\phi(\cdot)$ and $\sigma_\phi(\cdot)$ are learned functions of the backbone representation. This stochastic formulation allows calibration behavior to remain ambiguous even when the observed feature vector is fixed. Samples with similar predictions may still display mixed or uncertain score behavior, and such ambiguity is naturally represented through posterior uncertainty in latent space.

Stochastic grouping map. For a latent draw $z \sim q_\phi(z | x)$, we obtain a simplex-valued membership vector $\tilde{w}(x, z) = (\tilde{w}_1(x, z), \dots, \tilde{w}_K(x, z)) \in \Delta^{K-1}$ through a learnable differentiable mapping from the latent state to the probability simplex. A single latent draw represents only one possible realization of the posterior calibration state and therefore only one possible decomposition of the sample’s local score law. What we seek, however, is a stable description of how the sample is associated with the latent calibration components. Accordingly, SGCP defines the *stochastic grouping map* by averaging memberships over the posterior distribution. Let $z^{(1)}, \dots, z^{(M)} \stackrel{\text{i.i.d.}}{\sim} q_\phi(z | x)$. We define the posterior-averaged membership of sample x on the k -th latent calibration component

$$w_k(x) = \mathbb{E}_{z \sim q_\phi(\cdot | x)}[\tilde{w}_k(x, z)] \approx \frac{1}{M} \sum_{m=1}^M \tilde{w}_k(x, z^{(m)}), \quad k = 1, \dots, K. \quad (7)$$

This yields the stochastic grouping map $w(x) = (w_1(x), \dots, w_K(x)) \in \Delta^{K-1}$, which determines how the latent calibration components are combined to form the learned local score law.

Aggregate local score law. Each latent calibration component $F_k : [0, 1] \rightarrow [0, 1]$ is a monotone mapping that represents one recurring calibration pattern in the score space. When needed, we write f_k for the corresponding density. Given the membership weight $w_k(x)$, SGCP defines the aggregate local score law at x as

$$\hat{F}(s | x) = \sum_{k=1}^K w_k(x) F_k(s), \quad (8)$$

with corresponding local density $\hat{f}(s | x) = \sum_{k=1}^K w_k(x) f_k(s)$. Equation (8) defines the central local score-law estimator in SGCP. It replaces the infeasible task of estimating an unconstrained F_x^* for every test point with a structured family generated by a small number of latent calibration components. In our implementation, each F_k is parameterized as a monotone differentiable mapping.

We next explain why local score-law normalization matters for subgroup reliability. If the local score law were known, the probability integral transform would map true-label nonconformity score onto a common percentile scale. The following proposition formalizes this property (proof in Appendix A.1).

Proposition 1. Let $V(x, y) := \hat{F}(s(x, y) | x)$ denote the transformed score. Suppose that the learned local score law coincides with the oracle local score law, namely, for P_X -almost every x , $\hat{F}(t | x) = F_x^*(t) := \mathbb{P}(S \leq t | X = x)$, $\forall t \in [0, 1]$. Assume furthermore that, for P_X -a.e. x , the conditional score distribution $S | X = x$ is continuous. Then, for any measurable subgroup $G \subseteq \mathcal{X}$ with $\mathbb{P}(X \in G) > 0$,

$$V(X, Y) | \{X \in G\} \sim \text{Unif}(0, 1).$$

Algorithm 1 Procedure for SGCP

Require: Training data $\mathcal{D}_{\text{tra}}$, calibration data $\mathcal{D}_{\text{cal}}$, test input X_{n+1} , nonconformity score $s(\cdot, \cdot)$, K latent calibration components, M posterior samples, hyperparameters λ, β , miscoverage level α

Ensure: Prediction set $\hat{C}_\alpha(X_{n+1})$

- 1: Split $\mathcal{D}_{\text{tra}}$ into $\mathcal{D}_0$ and $\mathcal{D}_1$
 - 2: **Fit SGCP:** on $\mathcal{D}_0$, learn f_θ , latent-state posterior $q_\phi(z | x)$ in Eq. (6), and stochastic grouping map $w(x)$ in Eq. (7); on $\mathcal{D}_1$, fit latent calibration components $\{F_k\}_{k=1}^K$ and local score law $\hat{F}(\cdot | x)$ in Eq. (8).
 - 3: **Calibration:** for each $(X_i, Y_i) \in \mathcal{D}_{\text{cal}}$, compute $s_i = s(X_i, Y_i)$ and $V_i = \hat{F}(s_i | X_i)$.
 - 4: **Quantile:** sort $\{V_i : (X_i, Y_i) \in \mathcal{D}_{\text{cal}}\}$ as $V_{(1)} \leq \dots \leq V_{(m)}$, where $m = |\mathcal{D}_{\text{cal}}|$, and set $\hat{q}_{1-\alpha} = V_{(\lceil (m+1)(1-\alpha) \rceil)}$, with fallback $\hat{q}_{1-\alpha} = 1$ if $\lceil (m+1)(1-\alpha) \rceil > m$.
 - 5: **Test transform:** for each candidate label $y \in \mathcal{Y}$, compute $V_{n+1}^y = \hat{F}(s(X_{n+1}, y) | X_{n+1})$.
 - 6: **Return:** $\hat{C}_\alpha(X_{n+1}) = \{y \in \mathcal{Y} : V_{n+1}^y \leq \hat{q}_{1-\alpha}\}$.
-

Proposition 1 explains why local score law learning can remove subgroup reliability inconsistency in the oracle case. In practice, SGCP learns this transform approximately, using the following objective.

Learning objective. We split the training data into $\mathcal{D}_{\text{tra}} = \mathcal{D}_0 \cup \mathcal{D}_1$. The split $\mathcal{D}_0$ is used to learn the backbone predictor, the latent-state posterior, and the stochastic grouping map, while $\mathcal{D}_1$ is used to fit the latent calibration components and the learned local score law. An additional independent calibration set $\mathcal{D}_{\text{cal}}$ is reserved for split conformal calibration. For each $(X_i, Y_i) \in \mathcal{D}_1$, we compute the true-label nonconformity score $s_i = s(X_i, Y_i)$, the membership weights $w_{ik} = w_k(X_i)$, and the component-wise transformed score $C_{ik} = F_k(s_i)$. Since F_k is monotone, C_{ik} can be viewed as the percentile position of s_i under the k -th latent calibration component.

The training objective encourages three complementary properties. First, within each component, the weighted distribution of transformed scores should be close to $\text{Unif}(0, 1)$. This yields the rank-uniformization loss $\mathcal{L}_{\text{rank}}$. Second, the induced local mixture density $\hat{f}(s_i | X_i) = \sum_{k=1}^K w_{ik} f_k(s_i)$ should explain the observed true-label scores, yielding the density-fitting loss $\mathcal{L}_{\text{null}}$. Finally, the stochastic grouping map should be informative: it should use the latent components diversely across the population while retaining sample-specific membership weights. We encourage this by maximizing the empirical mutual information between samples and a latent component variable Z , denoted by $\hat{I}^{\text{norm}}(X; Z)$, where the membership weights define $q_\phi(Z = k | X_i) = w_{ik}$.

Combining these terms, we optimize

$$\mathcal{L}_{\text{SGCP}} = \mathcal{L}_{\text{rank}} + \lambda \mathcal{L}_{\text{null}} - \beta \hat{I}^{\text{norm}}(X; Z), \quad (9)$$

where λ controls the strength of local density fitting and β controls the strength of stochastic grouping information maximization. Detailed parameterizations and explicit loss definitions are deferred to Appendix B. Algorithm 1 summarizes the overall procedure.

The following result, proved in Appendix A.2, shows that SGCP retains the standard finite-sample marginal validity guarantee of split conformal prediction, which contains the true label with probability at least $1 - \alpha$. This guarantee does not depend on how accurately the local score law is learned, and the prediction set $\hat{C}_\alpha(X_{n+1})$ remains valid.

Proposition 2. *Under the standard split-conformal exchangeability assumption, the prediction set returned by SGCP satisfies $\mathbb{P}\{Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\} \geq 1 - \alpha$.*

Marginal validity alone does not explain subgroup reliability. A further question is how the learned transform affects subgroup coverage. The following (proved in Appendix A.3) shows that subgroup coverage is controlled by the mismatch between subgroup and overall transformed-score distributions. In the oracle property of Proposition 1, this mismatch vanishes for every subgroup. In practice, SGCP seeks to reduce it by learning a local score law transform through stochastic grouping. Appendix A.4 shows that uniformly approximating the oracle score law directly controls this mismatch.

Proposition 3. *For any measurable subgroup G with $\mathbb{P}(X \in G) > 0$, let Δ_G denote the Kolmogorov distance between the distribution of $V(X, Y)$ conditional on $X \in G$ and its marginal distribution. Under the assumptions of Proposition 2 and a continuous marginal distribution of $V(X, Y)$, subgroup coverage is lower bounded by $\mathbb{P}\{Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) | X_{n+1} \in G\} \geq 1 - \alpha - \Delta_G$.*

4 Numerical Experiments

Benchmarks and preprocessing. We evaluate SGCP on four benchmarks. For each dataset, let (X, Y, A) denote the feature vector, class label, and sensitive attributes, respectively. Below, we briefly describe how each dataset is processed and constructed for our evaluation.

(i) We generate synthetic (SYN) multiclass cardiopulmonary diagnosis benchmark with six disease categories, viz. Hypertension, Coronary Artery Disease, Heart Failure, Arrhythmia, Valvular Heart Disease, and Peripheral Artery Disease. Each sample contains four grouping attributes—gender, risk phenotype, age, and region—and six non-sensitive physiological variables. To mimic intersectional heterogeneity, individuals satisfying the XNOR relation between biological sex and risk phenotype, e.g., Female with Type I or Male with Type II, are assigned higher label uncertainty. This subgroup is therefore harder to classify and serves as the disadvantaged population. (ii) Nursery [19] is an admission-decision dataset for ranking nursery school applications, where family, socioeconomic, social, and health-related factors make it well suited for evaluating subgroup reliability. We use four priority ranks as targets and five categorical grouping attributes. Following [29, 27], we simulate controllable algorithmic bias by downsampling the subgroup defined by parents to 10%, adding label noise, and rounding labels to the nearest valid class. (iii) MIMIC-IV [13] is a real-world clinical tabular benchmark. We employ three binary grouping attributes, i.e., race, gender, and insurance status, for subgroup analysis, along with non-sensitive demographic and administrative covariates. We study two tasks: in-hospital mortality prediction and ICU-delay classification, where delays are grouped into short (≤ 4 hours), moderate (4-12 hours), and prolonged (> 12 hours) or no ICU admission. (iv) BACH [2] is a four-class histopathology benchmark for breast cancer diagnosis, viz. Normal, Benign, In Situ Carcinoma, and Invasive Carcinoma. Since demographic attributes are unavailable, we use the class label as a proxy grouping variable, setting $A := Y$. Images are encoded by a visual backbone as $X_i = \Phi(I_i) \in \mathbb{R}^d$. Subgroup heterogeneity arises from visual variability across tissue and potential acquisition bias [15], so the class-conditional score laws $F_c(s) := \mathbb{P}(S \leq s \mid Y = c)$ may differ substantially across classes.

Across all datasets, we randomly split data into training, calibration, and test sets; for SGCP, the training set is further divided equally into $\mathcal{D}_0$ and $\mathcal{D}_1$. Sensitive attributes are used only for fairness evaluation and never as model inputs. Preprocessing details are given in Appendix D.1.

We benchmark SGCP against AFCP [29], Clustered CP [8], FaReG [27], RLCP [12], marginal CP, Partial Equalized [20], and Exhaustive Equalized. Partial Equalized protects one sensitive attribute at a time and unions the resulting sets, while Exhaustive Equalized protects all candidate attributes jointly as an explicit subgroup-equalization baseline. All methods use the same backbone and nonconformity score. Implementation details and hyperparameters are provided in Appendix D.2. See the Appendix E for additional experiments.

Evaluation criteria. We evaluate all methods in terms of overall validity, subgroup reliability, and efficiency. Let $\{(X'_i, Y'_i)\}_{i=1}^{N'}$ be the test set, and let $\mathcal{C}(X'_i)$ denote the prediction set at miscoverage level α . We report empirical *overall coverage* $\text{Cov.} = \frac{1}{N'} \sum_{i=1}^{N'} \mathbf{1}\{Y'_i \in \mathcal{C}(X'_i)\}$, which should be close to the target level $1 - \alpha$. Efficiency is measured by the *average set size* $\text{AvgSize} = \frac{1}{N'} \sum_{i=1}^{N'} |\mathcal{C}(X'_i)|$. Among methods with comparable coverage, smaller AvgSize is preferable.

Given a grouping attribute A and its induced groups, we are interested in coverage conditional on each group. For group $g \in \mathcal{G}(A)$, let $\mathcal{I}_g := \{i : A(X'_i) = g\}$. We define the estimated conditional coverage, hereafter *group coverage*, as $\hat{c}_g = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} \mathbf{1}\{Y'_i \in \mathcal{C}(X'_i)\}$. We quantify subgroup disparity by the *coverage gap* (CovGap) $\text{CovGap}(A) = \frac{1}{|\mathcal{G}(A)|} \sum_{g \in \mathcal{G}(A)} |\hat{c}_g - (1 - \alpha)|$. Group coverage closer to $1 - \alpha$ and smaller CovGap indicate more reliable subgroup performance. All metrics other than AvgSize are reported as percentages.

4.1 Results and Discussion

Main results. Table 1 compares all methods across the four benchmarks (other experiments cf. Appendix E). AFCP is omitted on MIMIC-IV because its adaptive group selection requires an expensive repeated calibration procedure, making 20-run evaluation computationally prohibitive. We highlight AvgSize values exceeding half of the label space in red, as prediction sets of this size can reduce clinical utility, increase the cognitive burden on decision makers.

Table 1: Each method is run 20 times, and we report the mean and standard deviation. All results are reported for the largest sample size. Results with nominal coverage are shown in **blue**, optimal values are **boldfaced**, and AvgSize exceeding half of the total number of categories are highlighted in **red**.

Method	SYN			Nursery			MIMIC-IV			BACH		
	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$
Marginal	89.52 $\pm$ 1.52	2.309 $\pm$ 0.078	14.5 $\pm$ 1.6	84.15 $\pm$ 5.20	2.170 $\pm$ 0.260	20.4 $\pm$ 5.2	90.79 $\pm$ 10.62	1.226 $\pm$ 0.087	27.1 $\pm$ 7.1	89.20 $\pm$ 9.64	2.894 $\pm$ 0.267	15.7 $\pm$ 1.7
Partial	94.50 $\pm$ 1.12	2.419 $\pm$ 0.048	6.2 $\pm$ 0.7	88.97 $\pm$ 3.80	1.850 $\pm$ 0.200	12.1 $\pm$ 2.5	94.50 $\pm$ 4.54	1.258 $\pm$ 0.140	8.0 $\pm$ 3.0	96.57 $\pm$ 8.08	3.385 $\pm$ 0.299	7.6 $\pm$ 1.6
Exhaustive	95.85 $\pm$ 0.91	3.342 $\pm$ 0.053	6.0 $\pm$ 0.6	96.80 $\pm$ 3.10	2.850 $\pm$ 0.220	15.5 $\pm$ 3.2	96.45 $\pm$ 4.22	2.800 $\pm$ 0.142	19.7 $\pm$ 3.2	97.41 $\pm$ 6.89	3.867 $\pm$ 0.118	11.6 $\pm$ 1.5
RLCP	93.10 $\pm$ 2.45	2.540 $\pm$ 0.305	4.5 $\pm$ 1.8	88.67 $\pm$ 6.20	1.220 $\pm$ 0.410	13.5 $\pm$ 3.8	92.50 $\pm$ 2.80	1.350 $\pm$ 0.100	1.5 $\pm$ 0.8	91.88 $\pm$ 10.56	2.311 $\pm$ 0.319	8.1 $\pm$ 2.5
CluCP	89.80 $\pm$ 1.78	2.426 $\pm$ 0.050	4.9 $\pm$ 0.8	85.08 $\pm$ 1.60	1.121 $\pm$ 0.064	12.8 $\pm$ 1.3	92.22 $\pm$ 1.28	1.302 $\pm$ 0.059	3.5 $\pm$ 0.8	95.91 $\pm$ 3.59	3.257 $\pm$ 0.220	10.7 $\pm$ 2.6
AFCP	91.77 $\pm$ 1.23	2.432 $\pm$ 0.043	4.6 $\pm$ 0.5	92.85 $\pm$ 0.84	1.085 $\pm$ 0.024	14.1 $\pm$ 0.6	-	-	-	95.00 $\pm$ 2.60	2.830 $\pm$ 0.237	6.3 $\pm$ 1.7
FaReG	92.61 $\pm$ 0.98	2.579 $\pm$ 0.029	4.7 $\pm$ 0.5	93.21 $\pm$ 0.62	1.230 $\pm$ 0.012	13.3 $\pm$ 0.6	90.23 $\pm$ 0.09	1.174 $\pm$ 0.001	0.5 $\pm$ 0.0	93.60 $\pm$ 3.40	2.959 $\pm$ 0.239	7.4 $\pm$ 1.5
SGCP	90.13 $\pm$ 0.39	2.225 $\pm$ 0.035	4.2 $\pm$ 0.7	91.65 $\pm$ 0.70	1.089 $\pm$ 0.035	11.8 $\pm$ 0.7	90.13 $\pm$ 0.07	1.107 $\pm$ 0.001	0.8 $\pm$ 0.0	91.90 $\pm$ 2.70	2.224 $\pm$ 0.215	6.2 $\pm$ 1.1

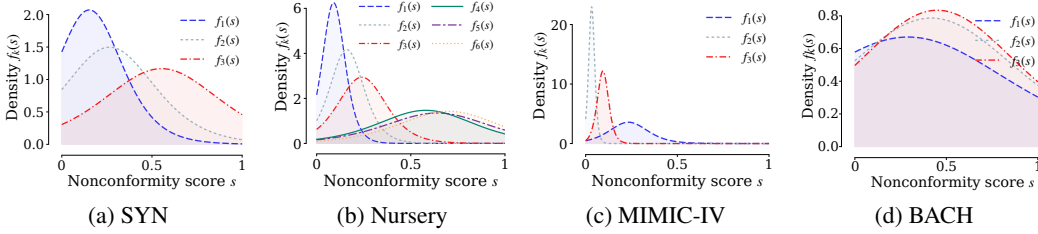


Figure 3: Learned latent component densities on four datasets. The components reveal distinct calibration regimes in the nonconformity-score space.

We highlight four main observations. **(i)** SGCP is the only method that consistently stays close to nominal coverage across all benchmarks while achieving strong subgroup reliability. Its empirical coverage remains near 0.9, and its CovGap is the smallest on SYN, Nursery, and BACH, with values 3.7, 11.8, and 6.2, respectively. On MIMIC-IV, it is close to the best result, with CovGap 0.8 compared with FaReG’s 0.5. **(ii)** The subgroup reliability gains of SGCP are not obtained through indiscriminate conservatism. SGCP achieves the smallest or nearly smallest AvgSize across all datasets. Taken together, these results show that SGCP improves subgroup reliability without enlarging prediction sets. **(iii)** Existing baselines exhibit less favorable or less consistent tradeoffs. Marginal CP often leaves large subgroup disparities, while explicit subgroup-protection methods reduce disparity at the cost of inflated coverage and larger prediction sets. RLCP, CluCP, AFCP, and FaReG are competitive in some settings, but their gains are less uniform across benchmarks.

Figure 3 visualizes the latent component densities learned by SGCP. Across datasets, the components exhibit diverse shapes in the nonconformity score space. Some components concentrate near small scores, while others are broader or place more mass on larger scores. This suggests that SGCP learns nontrivial score law structure, rather than merely reproducing a single global score distribution. Importantly, these components should not be viewed as demographic groups or semantic subpopulations. They are latent score law building blocks. For each sample, SGCP mixes these shared components to form a sample dependent local score law. In this way, samples with similar calibration behavior can share evidence even when they do not belong to the same observed subgroup. This provides a mechanism for improving subgroup reliability without requiring calibration to be tied to a single observed group label.

Group coverage. To better understand where the improvement comes from, we visualize subgroup coverage against average prediction-set size, together with how set size varies with subgroup sample size. Figure 4 corresponds to BACH, where the subgroup variable is the class label. Two patterns are clear. First, SGCP remains close to the nominal target coverage while lying on the more efficient side of the tradeoff: in the top row, its points stay near the horizontal reference line at 0.9 and are typically located to the left of competing methods. Second, in the bottom row, its prediction sets shrink steadily as subgroup sample size increases, indicating that it can use additional subgroup evidence efficiently while still adapting locally. Overall, these subgroup-level plots show that the gains of SGCP are realized across individual groups rather than only in aggregate.

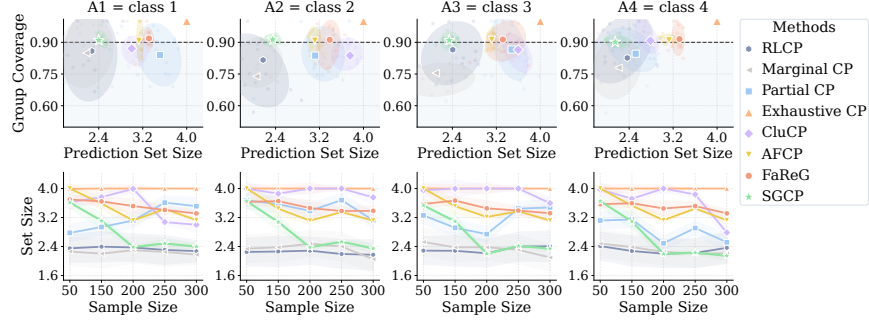


Figure 4: Group coverage on BACH dataset. SGCP remains close to the target coverage across classes while using smaller prediction sets than more conservative baselines.

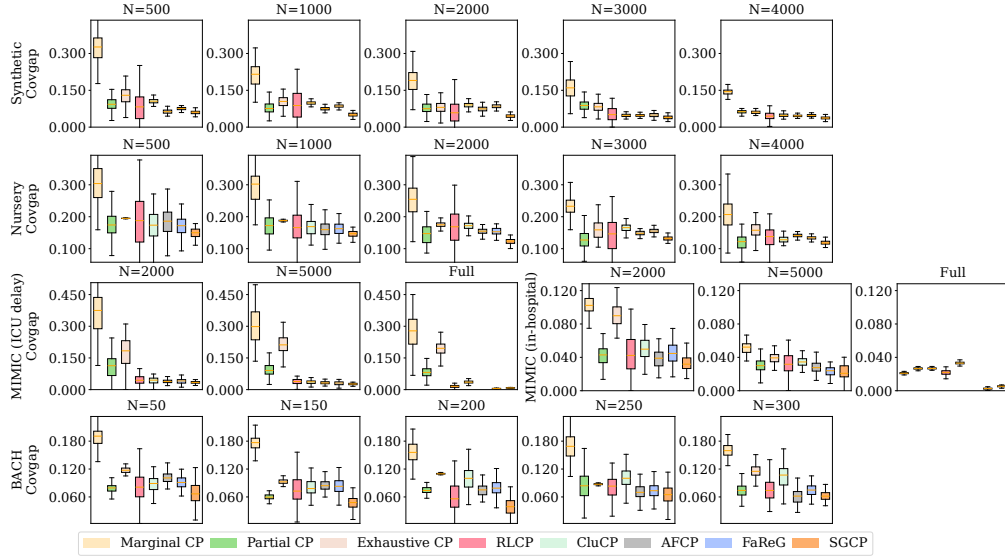


Figure 5: Distribution of CovGap over repeated runs under varying sample sparsity.

Distribution of CovGap. Figure 5 shows the distribution of CovGap over repeated runs under varying levels of sample sparsity. Beyond low average disparity, SGCP also exhibits stability: across datasets and sample sizes, it is consistently among the methods with the lowest median CovGap and relatively small variation across runs. This advantage is most pronounced in sparse-data settings, where subgroup calibration is most difficult. As sample size grows, the gap narrows for most methods, but SGCP remains competitive and typically retains one of the lowest CovGap values.

5 Conclusion

In this work, we discussed the limitation of standard split-conformal prediction in fairness-sensitive settings, where marginal validity can mask substantial subgroup miscoverage. We argued that subgroup reliability is closely tied to the local distribution of nonconformity scores, and that modeling this local score law provides a principled way to improve calibration across heterogeneous subpopulations. To this end, we proposed SGCP, a fairness-aware conformal prediction method that improves coverage for clinically meaningful subgroups without relying on predefined sensitive groups. Our theoretical analysis showed that SGCP preserves the standard split-conformal marginal coverage guarantee and accurate local score-law learning reduces subgroup coverage gaps by making calibration behavior more comparable across subpopulations. Extensive experiments across synthetic, and real-world benchmarks demonstrated that SGCP consistently improves subgroup reliability while maintaining efficient prediction sets. In future work, we plan to extend SGCP to personalized medicine, where reliable uncertainty quantification across diverse patient populations is crucial.

References

- [1] A. N. Angelopoulos and S. Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.
- [2] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q. D. Vu, M. N. N. To, E. Kim, J. T. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun, K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia, and P. Aguiar. Bach: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019.
- [3] K. Bairaktari, J. Wu, and S. Wu. Kandinsky conformal prediction: Beyond class- and covariate-conditional coverage. In *International Conference on Learning Representations (ICLR)*, 2025.
- [4] R. J. Chen, J. J. Wang, D. F. K. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, and F. Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6):719–742, 2023.
- [5] M. Y. Cheung, A. Veeraraghavan, and G. Balakrishnan. COMPASS: Robust feature conformal prediction for medical segmentation metrics. In *International Conference on Learning Representations (ICLR)*, 2026.
- [6] S. Cortes-Gomez, C. M. Patiño, Y. Byun, S. Wu, E. Horvitz, and B. Wilder. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. In *International Conference on Learning Representations (ICLR)*, 2025.
- [7] J. C. Cresswell, B. Kumar, Y. Sui, and M. Belbahri. Conformal prediction sets can cause disparate impact. In *International Conference on Learning Representations (ICLR)*, 2025.
- [8] T. Ding, A. N. Angelopoulos, S. Bates, M. I. Jordan, and R. J. Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 64555–64576, 2023.
- [9] S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, J. Zittrain, I. S. Kohane, and S. Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.
- [10] R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.
- [11] C. Gao, P. B. Gilbert, and L. Han. Bridging fairness and efficiency in conformal inference: a surrogate-assisted group-clustered approach. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.
- [12] R. Hore and R. F. Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):549–578, 2024.
- [13] A. Johnson, L. Bulgarelli, T. Pollard, B. Gow, B. Moody, S. Horng, L. A. Celi, and R. Mark. MIMIC-IV. *PhysioNet*, 2024. Version 3.1.
- [14] H. Kwon and D.-J. Kim. Conformal selective prediction with cost aware deferral for safe clinical triage under distribution shift. *Scientific Reports*, 16(1):10016, 2026.
- [15] J. Lee and R. M. Nishikawa. Identifying women with mammographically- occult breast cancer leveraging gan-simulated mammograms. *IEEE Transactions on Medical Imaging*, 41(1):225–236, 2022.
- [16] C. Lu, A. N. Angelopoulos, and S. Pomerantz. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, page 545–554, 2022.

- [17] C. Lu, A. Lemay, K. Chang, K. Höbel, and J. Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 12008–12016, 2022.
- [18] N. Martinez, D. C. Patel, C. Reddy, G. Ganapavarapu, R. Vaculin, and J. Kalagnanam. Identifying homogeneous and interpretable groups for conformal prediction. In *The Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.
- [19] V. Rajkovic. Nursery. UCI Machine Learning Repository, 1989. DOI: <https://doi.org/10.24432/C5P88W>.
- [20] Y. Romano, R. F. Barber, C. Sabatti, and E. Candès. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2(2), 2020.
- [21] Y. Romano, M. Sesia, and E. Candes. Classification with valid and adaptive coverage. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 3581–3591, 2020.
- [22] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.
- [23] A. T. Vadlamani, A. Srinivasan, P. Maneriker, A. Payani, and S. Parthasarathy. A generic framework for conformal fairness. In *International Conference on Learning Representations (ICLR)*, 2025.
- [24] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*, pages 1–476. Springer Cham, 2005.
- [25] V. Vovk, D. Lindsay, I. Nourtdinov, and A. Gammerman. Mondrian confidence machine. *Technical Report*, 2003.
- [26] Z. Wang and H. Wang. Variational imbalanced regression: Fair uncertainty quantification via probabilistic smoothing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [27] S. Xu, Y. Zhou, Y. Tan, Z. Li, Y. Yao, T. Chen, F. Xu, and X. Ma. Fair conformal classification via learning representation-based groups. In *International Conference on Learning Representations (ICLR)*, 2026.
- [28] D. Zhang, A. Chatzimparmpas, N. Kamali, and J. Hullman. Evaluating the utility of conformal prediction sets for ai-advised image labeling. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2024.
- [29] Y. Zhou and M. Sesia. Conformal classification with equalized coverage for adaptively selected groups. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

A Mathematical Proofs

A.1 Proof of Proposition 1

Proof. Let $S := s(x, y)$ and $V := \widehat{F}(S \mid X)$. Let λ denote the Lebesgue measure on $[0, 1]$. By the oracle assumption, for P_X -a.e. x ,

$$\widehat{F}(\cdot \mid x) = F_x^*(\cdot), \quad F_x^*(t) := \mathbb{P}(S \leq t \mid X = x).$$

Moreover, by assumption, the conditional distribution of $S \mid X = x$ is continuous for P_X -a.e. x . Hence, by the probability integral transform, for P_X -a.e. x ,

$$F_x^*(S) \mid X = x \sim \text{Unif}(0, 1).$$

Since $V = \widehat{F}(S \mid X) = F_X^*(S)$ almost surely, it follows that, for every Borel set $A \subseteq [0, 1]$,

$$\mathbb{E}[\mathbf{1}_{\{V \in A\}} \mid X] = \lambda(A) \quad \text{a.s.}$$

Now let $G \subseteq \mathcal{X}$ be any measurable subgroup with $\mathbb{P}(X \in G) > 0$. Using the tower property,

$$\begin{aligned} \mathbb{P}(V \in A \mid X \in G) &= \frac{\mathbb{E}[\mathbf{1}_{\{V \in A\}} \mathbf{1}_{\{X \in G\}}]}{\mathbb{P}(X \in G)} \\ &= \frac{\mathbb{E}[\mathbf{1}_{\{X \in G\}} \mathbb{E}[\mathbf{1}_{\{V \in A\}} \mid X]]}{\mathbb{P}(X \in G)} \\ &= \frac{\mathbb{E}[\mathbf{1}_{\{X \in G\}} \lambda(A)]}{\mathbb{P}(X \in G)} \\ &= \lambda(A). \end{aligned}$$

Therefore, for every set $A \subseteq [0, 1]$, $\mathbb{P}(V \in A \mid X \in G) = \lambda(A)$, which is precisely the statement that

$$V \mid \{X \in G\} \sim \text{Unif}(0, 1).$$

□

Remark 1. When this oracle local score law is recovered, the local percentile transform places scores from different inputs and subgroups on a common uniform scale, thereby eliminating subgroup reliability inconsistency. The objective of SGCP is to approximate this. Put differently, this proposition clarifies why we learn $\widehat{F}(s \mid x)$ instead of operating directly on the raw score s .

A.2 Proof of Proposition 2

Throughout the proof, we condition on the training split $\mathcal{D}_{\text{tra}}$. Thus the fitted transformation $(x, y) \mapsto \widehat{F}(s(x, y) \mid x)$ is fixed. For simplicity, we suppress this conditioning from the notation.

Let $m := |\mathcal{D}_{\text{cal}}|$ and write $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^m$. Define the transformed calibration scores

$$V_i := \widehat{F}(s(X_i, Y_i) \mid X_i), \quad i = 1, \dots, m,$$

and the transformed score of the test point under its true label

$$V_{\text{test}} := \widehat{F}(s(X_{n+1}, Y_{n+1}) \mid X_{n+1}).$$

Let $V_{(1)} \leq \dots \leq V_{(m)}$ denote the order statistics of the calibration scores, and set

$$k := \lceil (m+1)(1-\alpha) \rceil.$$

Define the split-conformal cutoff by $\widehat{q}_{1-\alpha} = V_{(k)}$, with the convention $\widehat{q}_{1-\alpha} = 1$ whenever $k > m$.

Proof. By assumption, conditional on $\mathcal{D}_{\text{tra}}$, the calibration examples in $\mathcal{D}_{\text{cal}}$ and the test point (X_{n+1}, Y_{n+1}) are exchangeable. Since the same deterministic measurable map is applied to each point, it follows that $V_1, \dots, V_m, V_{\text{test}}$ are exchangeable conditional on $\mathcal{D}_{\text{tra}}$.

Let $V_{(1)} \leq \dots \leq V_{(m)}$ denote the ordered transformed calibration scores, and define the split conformal cutoff by $\hat{q}_{1-\alpha} = V_{(\lceil (m+1)(1-\alpha) \rceil)}$, with the convention $\hat{q}_{1-\alpha} = 1$ whenever $\lceil (m+1)(1-\alpha) \rceil > m$. By exchangeability and the standard split conformal order-statistic argument, we have $\mathbb{P}(V_{\text{test}} \leq \hat{q}_{1-\alpha} \mid \mathcal{D}_{\text{tra}}) \geq 1 - \alpha$.

It remains to connect this event to the prediction set. By definition,

$$\hat{C}_\alpha(X_{n+1}) = \left\{ y \in \mathcal{Y} : \hat{F}(s(X_{n+1}, y) \mid X_{n+1}) \leq \hat{q}_{1-\alpha} \right\}.$$

Substituting the true label Y_{n+1} gives

$$\left\{ Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \right\} = \left\{ \hat{F}(s(X_{n+1}, Y_{n+1}) \mid X_{n+1}) \leq \hat{q}_{1-\alpha} \right\} = \{V_{\text{test}} \leq \hat{q}_{1-\alpha}\}.$$

Hence,

$$\mathbb{P}(Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \mid \mathcal{D}_{\text{tra}}) \geq 1 - \alpha.$$

Taking expectations over $\mathcal{D}_{\text{tra}}$ gives

$$\mathbb{P}\left\{ Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \right\} = \mathbb{E}\left[\mathbb{P}\left\{ Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \mid \mathcal{D}_{\text{tra}} \right\}\right] \geq 1 - \alpha.$$

□

Remark 2. In the current implementation, the stochastic grouping map $w(x)$ may be approximated by Monte Carlo averaging over latent posterior samples. The validity proof above continues to hold as long as the transformed score is computed using a rule that is fixed conditional on $\mathcal{D}_{\text{tra}}$, or, more generally, uses auxiliary randomness that is independent across calibration and test samples.

A.3 Proof of Proposition 3

Proof. By definition of the prediction set,

$$\hat{C}_\alpha(X_{n+1}) = \left\{ y \in \mathcal{Y} : \hat{F}(s(X_{n+1}, y) \mid X_{n+1}) \leq \hat{q}_{1-\alpha} \right\}.$$

Therefore,

$$\left\{ Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \right\} = \{V_{\text{test}} \leq \hat{q}_{1-\alpha}\}.$$

Now fix a measurable subgroup $G \subseteq \mathcal{X}$ with $\mathbb{P}(X \in G) > 0$. Since the calibration sample is independent of the test point, $\hat{q}_{1-\alpha}$ is independent of (X_{n+1}, Y_{n+1}) . Hence, conditional on $\hat{q}_{1-\alpha}$,

$$\begin{aligned} \mathbb{P}(Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \mid X_{n+1} \in G, \hat{q}_{1-\alpha}) \\ = \mathbb{P}(V_{n+1} \leq \hat{q}_{1-\alpha} \mid X_{n+1} \in G, \hat{q}_{1-\alpha}) \\ = H_G(\hat{q}_{1-\alpha}), \end{aligned}$$

where

$$H_G(t) := \mathbb{P}(V(X, Y) \leq t \mid X \in G).$$

Taking expectations over $\hat{q}_{1-\alpha}$ gives

$$\mathbb{P}(Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \mid X_{n+1} \in G) = \mathbb{E}[H_G(\hat{q}_{1-\alpha})].$$

Next, let

$$H(t) := \mathbb{P}(V(X, Y) \leq t)$$

be the global CDF of the transformed score, and let $V^\dagger$ be an independent draw from this distribution. Since $V_1, \dots, V_m, V^\dagger$ are exchangeable and H is continuous, the standard split-conformal rank argument gives

$$\mathbb{P}(V^\dagger \leq \hat{q}_{1-\alpha}) = \frac{k}{m+1}.$$

On the other hand, conditioning on $\hat{q}_{1-\alpha}$,

$$\mathbb{P}(V^\dagger \leq \hat{q}_{1-\alpha} \mid \hat{q}_{1-\alpha}) = H(\hat{q}_{1-\alpha}).$$

Therefore,

$$\mathbb{E}[H(\hat{q}_{1-\alpha})] = \frac{k}{m+1}.$$

Combining the two identities,

$$\begin{aligned} & \left| \mathbb{P}(Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \mid X_{n+1} \in G) - \frac{k}{m+1} \right| \\ &= |\mathbb{E}[H_G(\hat{q}_{1-\alpha})] - \mathbb{E}[H(\hat{q}_{1-\alpha})]| \\ &= |\mathbb{E}[H_G(\hat{q}_{1-\alpha}) - H(\hat{q}_{1-\alpha})]| \\ &\leq \mathbb{E}[|H_G(\hat{q}_{1-\alpha}) - H(\hat{q}_{1-\alpha})|] \\ &\leq \sup_{t \in [0,1]} |H_G(t) - H(t)| = \Delta_G. \end{aligned}$$

This proves the claimed subgroup coverage gap bound.

Finally, since

$$\frac{k}{m+1} = \frac{\lceil (m+1)(1-\alpha) \rceil}{m+1} \geq 1-\alpha,$$

we obtain

$$\mathbb{P}(Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \mid X_{n+1} \in G) \geq 1-\alpha - \Delta_G.$$

□

Since $k/(m+1) \geq 1-\alpha$, this result implies that whenever a subgroup's transformed-score distribution is close to the overall transformed-score distribution, the global conformal cutoff keeps its coverage close to the nominal level $1-\alpha$.

Remark 3. Proposition 3 bounds subgroup miscoverage in terms of the residual transformed-score mismatch Δ_G . This mismatch can be further related to local score-law estimation error. If the learned transform satisfies

$$\sup_{x,s} |\hat{F}(s \mid x) - F_x^*(s)| \leq \varepsilon,$$

then, by the probability integral transform,

$$\Delta_G \leq 2\varepsilon$$

for every measurable subgroup G . Consequently,

$$\left| \mathbb{P}\{Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \mid X_{n+1} \in G\} - (1-\alpha) \right| \leq 2\varepsilon + \frac{1}{m+1}.$$

Thus, accurate local score-law learning provides a sufficient condition for small subgroup coverage gaps.

A.4 Empirical Subgroup Coverage

We now relate the empirical subgroup coverage measured on an independent test sample to the population subgroup coverage above. Under the assumptions of Proposition 3, let $\mathcal{D}_G^{\text{test}} = \{(X_j^{(G)}, Y_j^{(G)})\}_{j=1}^{n_G}$ be an i.i.d. test sample drawn from the conditional distribution of (X, Y) given $X \in G$, independent of $\mathcal{D}_{\text{tra}}$ and $\mathcal{D}_{\text{cal}}$.

Corollary 1. Let $\widehat{\text{Cov}}_G$ be the coverage measured on an independent subgroup test sample of size n_G , denoted as $\widehat{\text{Cov}}_G := \frac{1}{n_G} \sum_{j=1}^{n_G} \mathbf{1}\{Y_j^{(G)} \in \hat{C}_\alpha(X_j^{(G)})\}$. Then, for any $\delta \in (0, 1)$, with probability at least $1-\delta$,

$$\left| \widehat{\text{Cov}}_G - (1-\alpha) \right| \leq \Delta_G + \sqrt{\frac{\log(2/\delta)}{2n_G}} + \frac{1}{m+1}.$$

In particular,

$$\widehat{\text{Cov}}_G \geq 1-\alpha - \Delta_G - \sqrt{\frac{\log(2/\delta)}{2n_G}}.$$

Corollary 1 connects the theoretical subgroup coverage guarantee to the quantities reported in our experiments. When the learned transform reduces Δ_G and the subgroup test sample size n_G is sufficiently large, the empirical subgroup coverage concentrates near the nominal level $1 - \alpha$ up to the conformal correction $1/(m+1)$.

Proof. Conditional on $\mathcal{D}_{\text{tra}}$ and $\mathcal{D}_{\text{cal}}$, the prediction set $\widehat{C}_\alpha$ is fixed. Since the test points

$$(X_j^{(G)}, Y_j^{(G)}), \quad j = 1, \dots, n_G,$$

are i.i.d. draws from the conditional distribution of (X, Y) given $X \in G$, the indicators

$$Z_j := \mathbf{1} \left\{ Y_j^{(G)} \in \widehat{C}_\alpha(X_j^{(G)}) \right\}, \quad j = 1, \dots, n_G,$$

are i.i.d. Bernoulli random variables with mean $\text{Cov}_G := \mathbb{P} \left\{ Y \in \widehat{C}_\alpha(X) \mid X \in G \right\}$, where the probability is understood conditional on $\mathcal{D}_{\text{tra}}$ and $\mathcal{D}_{\text{cal}}$.

Therefore, by Hoeffding's inequality, for any $\delta \in (0, 1)$,

$$\mathbb{P} \left(\left| \widehat{\text{Cov}}_G - \text{Cov}_G \right| \geq \sqrt{\frac{\log(2/\delta)}{2n_G}} \mid \mathcal{D}_{\text{tra}}, \mathcal{D}_{\text{cal}} \right) \leq \delta.$$

Equivalently, with probability at least $1 - \delta$ over $\mathcal{D}_G^{\text{test}}$,

$$\left| \widehat{\text{Cov}}_G - \text{Cov}_G \right| \leq \sqrt{\frac{\log(2/\delta)}{2n_G}}. \quad (\text{A10})$$

By Proposition 3, $\left| \text{Cov}_G - \frac{k}{m+1} \right| \leq \Delta_G$. Combining this with Eq. (A10) and the triangle inequality yields

$$\left| \widehat{\text{Cov}}_G - \frac{k}{m+1} \right| \leq \left| \widehat{\text{Cov}}_G - \text{Cov}_G \right| + \left| \text{Cov}_G - \frac{k}{m+1} \right| \leq \sqrt{\frac{\log(2/\delta)}{2n_G}} + \Delta_G.$$

Next, since $k = \lceil (m+1)(1-\alpha) \rceil$, we have $0 \leq \frac{k}{m+1} - (1-\alpha) \leq \frac{1}{m+1}$. Hence

$$\left| \widehat{\text{Cov}}_G - (1-\alpha) \right| \leq \left| \widehat{\text{Cov}}_G - \frac{k}{m+1} \right| + \left| \frac{k}{m+1} - (1-\alpha) \right| \leq \Delta_G + \sqrt{\frac{\log(2/\delta)}{2n_G}} + \frac{1}{m+1}.$$

Finally, Proposition 3 implies $\text{Cov}_G \geq \frac{k}{m+1} - \Delta_G \geq 1 - \alpha - \Delta_G$. On the same high-probability event as in Eq. (A10),

$$\widehat{\text{Cov}}_G \geq \text{Cov}_G - \sqrt{\frac{\log(2/\delta)}{2n_G}},$$

and therefore

$$\widehat{\text{Cov}}_G \geq 1 - \alpha - \Delta_G - \sqrt{\frac{\log(2/\delta)}{2n_G}}.$$

□

B Model Details

This appendix specifies the concrete instantiation of the aggregate local score law in Eq. (8) and the corresponding training objective. The algorithm 2 demonstrates a specific process.

Latent calibration state. Let $h_\theta(x) \in \mathbb{R}^d$ denote the backbone feature representation and let $f_\theta(x) \in \Delta^{C-1}$ denote the corresponding predictive probability vector. We parameterize a Gaussian latent posterior $q_\phi(z \mid x) = \mathcal{N}(\mu_\phi(h_\theta(x)), \text{diag}(\sigma_\phi^2(h_\theta(x))))$.

Algorithm 2 Detailed Training and Inference Procedure for SGCP

Require: Training data $\mathcal{D}_{\text{tra}}$, calibration data $\mathcal{D}_{\text{cal}}$, backbone classifier f_θ , nonconformity score $s(\cdot, \cdot)$, K latent calibration components, M posterior samples, hyperparameters λ, β , miscoverage level α

Ensure: Prediction set $\hat{C}_\alpha(X_{N+1})$

- 1: Split $\mathcal{D}_{\text{tra}}$ into $\mathcal{D}_0$ and $\mathcal{D}_1$
 - 2: Train the backbone predictor f_θ on $\mathcal{D}_0$
 - 3: Initialize the stochastic assignment module and component laws $\{F_k\}_{k=1}^K$
 - 4: **for** each optimization step on $\mathcal{D}_1$ **do**
 - 5: Sample a mini-batch $\mathcal{B} = \{(X_i, Y_i)\}_{i=1}^B$
 - 6: Compute backbone features, and nonconformity scores $s_i = s(X_i, Y_i)$
 - 7: Estimate posterior-averaged memberships from $q_\phi(z | X_i)$
 - 8: Compute component transforms $C_{ik} = F_k(S_i)$ and local densities $\hat{f}(S_i | X_i) = \sum_{k=1}^K w_{ik} f_k(S_i)$
 - 9: Compute $\mathcal{L}_{\text{rank}}, \mathcal{L}_{\text{nl}},$ and $\hat{I}_B^{\text{norm}}(X; Z)$
 - 10: Update assignment and component parameters by minimizing $\mathcal{L}_{\text{SGCP}}$
 - 11: **end for**
 - 12: Construct the learned local score transform $\hat{F}(s | x) = \sum_{k=1}^K w_k(x) F_k(s)$
 - 13: **for** each $(X_i, Y_i) \in \mathcal{D}_{\text{cal}}$ **do**
 - 14: Compute $V_i = \hat{F}(s(X_i, Y_i) | X_i)$
 - 15: **end for**
 - 16: Sort $\{V_i\}_{i=1}^m$ and compute the conformal cutoff $\hat{q}_{1-\alpha} = V_{(\lceil (m+1)(1-\alpha) \rceil)}$ with fallback $\hat{q}_{1-\alpha} = 1$ if $\lceil (m+1)(1-\alpha) \rceil > m$
 - 17: **for** each candidate label $y \in \mathcal{Y}$ **do**
 - 18: Compute $V_{\text{test}}^y = \hat{F}(S(X_{n+1}, y) | X_{n+1})$
 - 19: **end for**
 - 20: **return** $\hat{C}_\alpha(X_{n+1}) = \{y \in \mathcal{Y} : V_{n+1}^y \leq \hat{q}_{1-\alpha}\}$
-

Stochastic grouping map. Let $c_1, \dots, c_K \in \mathbb{R}^r$ denote K learnable prototypes in the latent space. For a latent draw $z \sim q_\phi(z | x)$, define the assignment vector

$$\tilde{w}_k(x, z) = \frac{\exp(-\|z - c_k\|_2^2/T)}{\sum_{j=1}^K \exp(-\|z - c_j\|_2^2/T)}, \quad k = 1, \dots, K,$$

where $T > 0$ is a temperature parameter. Given $z^{(1)}, \dots, z^{(M)} \stackrel{\text{i.i.d.}}{\sim} q_\phi(z | x)$, the posterior-averaged membership vector $w(x) = (w_1(x), \dots, w_K(x)) \in \Delta^{K-1}$ is estimated by

$$w_k(x) = \mathbb{E}_{z \sim q_\phi(\cdot | x)}[\tilde{w}_k(x, z)] \approx \frac{1}{M} \sum_{m=1}^M \tilde{w}_k(x, z^{(m)}), \quad k = 1, \dots, K.$$

Remark 4. Marginal conformal validity only requires global exchangeability, so it can hold even when score behavior varies substantially across subgroups. Such heterogeneity can make a single global calibration rule locally unreliable. SGCP addresses this by learning latent calibration components that shift calibration evidence among samples with similar score behavior, providing a practical way to approximate oracle local score laws without assuming exact local exchangeability.

Remark 5. In SGCP, posterior sampling approximates a stable stochastic grouping map rather than adding randomness to conformal calibration. For fixed x , draw $z^{(1)}, \dots, z^{(M)} \stackrel{\text{i.i.d.}}{\sim} q_\phi(z | x)$ and set

$$\hat{w}_k^{(M)}(x) = \frac{1}{M} \sum_{m=1}^M \tilde{w}_k(x, z^{(m)}), \quad w_k(x) = \mathbb{E}_{z \sim q_\phi(\cdot | x)}[\tilde{w}_k(x, z)].$$

The induced local score law is $\hat{F}^{(M)}(s | x) = \sum_{k=1}^K \hat{w}_k^{(M)}(x) F_k(s)$.

Samples share calibration evidence through the common components $\{F_k\}_{k=1}^K$, and posterior averaging reduces Monte Carlo variability in the sample-specific weights. As $M \rightarrow \infty$, the law of large

numbers gives $\hat{w}_k^{(M)}(x) \rightarrow w_k(x)$ almost surely, and therefore

$$\hat{F}^{(M)}(s | x) \rightarrow \hat{F}(s | x) = \sum_{k=1}^K w_k(x) F_k(s) \quad a.s.$$

B.1 Training Objective

We train the stochastic assignment module and the component parameters on $\mathcal{D}_1$, while keeping the backbone fixed. For each $(X_i, Y_i) \in \mathcal{D}_1$, define

$$S_i := s(X_i, Y_i), \quad w_{ik} := w_k(X_i), \quad C_{ik} := F_k(S_i).$$

The latent components are intended to serve as reusable local score laws. Thus, scores routed to the same component should be comparable after being mapped to their component-wise percentile scale, while the induced local mixture density should still explain the observed true-label nonconformity scores. These two requirements lead to the rank-uniformization and local density-fitting terms below.

Rank-uniformization. For a mini-batch $\mathcal{B}$ of size B and a fixed component k , let $\{C_{(1)k}, \dots, C_{(B)k}\}$ be the sorted values of $\{C_{ik} : i \in \mathcal{B}\}$, and let $w_{(1)k}, \dots, w_{(B)k}$ denote the correspondingly reordered weights. Normalize the weights within component k by

$$\pi_{(b)k} = \frac{w_{(b)k}}{\sum_{\ell=1}^B w_{(\ell)k}}.$$

The weighted rank midpoint is $r_{(b)k} = \sum_{\ell < b} \pi_{(\ell)k} + \frac{1}{2} \pi_{(b)k}$. We define the component-wise rank-uniformization loss as

$$\ell_{\text{rank}}^{(k)} = \sum_{b=1}^B \pi_{(b)k} (C_{(b)k} - r_{(b)k})^2.$$

Averaging over active components gives

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{K}_{\text{act}}|} \sum_{k \in \mathcal{K}_{\text{act}}} \ell_{\text{rank}}^{(k)}, \quad \mathcal{K}_{\text{act}} = \left\{ k : \sum_{i \in \mathcal{B}} w_{ik} > \tau \right\}.$$

Local density fitting. Let f_k denote the density associated with F_k . The local mixture density at (X_i, S_i) is $\hat{f}(S_i | X_i) = \sum_{k=1}^K w_{ik} f_k(S_i)$. We fit the observed true-label scores by the negative log-likelihood

$$\mathcal{L}_{\text{nll}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \hat{f}(S_i | X_i).$$

The rank term makes the component transforms behave like percentile maps, whereas the likelihood term ties the learned score laws to the empirical score distribution.

Information maximization for stochastic grouping. The preceding two terms specify how the component score laws should behave, but they do not by themselves ensure that the grouping map uses the components in an informative way. For example, a balance constraint on the average component mass can prevent all samples from collapsing onto a single component, but it still permits the opposite degenerate solution in which every sample is assigned nearly uniformly to all components. Such assignments are balanced in aggregate, yet they carry little sample-specific information and tend to make the learned component laws indistinguishable.

We therefore encourage the component assignment to retain information about the input. For a mini-batch $\mathcal{B}$, view the assignment as a conditional categorical distribution $q_\phi(Z = k | X_i) = w_{ik}$, where Z denotes the latent component index. The induced marginal component distribution on the mini-batch is

$$\bar{w}_k = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_{ik}, \quad \bar{w} = (\bar{w}_1, \dots, \bar{w}_K).$$

The empirical mutual information between inputs and component assignments is

$$\hat{I}_{\mathcal{B}}(X; Z) = H(\bar{w}) - \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} H(w_i),$$

where

$$H(\bar{w}) = - \sum_{k=1}^K \bar{w}_k \log \bar{w}_k, \quad H(w_i) = - \sum_{k=1}^K w_{ik} \log w_{ik}.$$

We use the normalized form

$$\hat{I}_{\mathcal{B}}^{\text{norm}}(X; Z) = \frac{1}{\log K} \left[H(\bar{w}) - \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} H(w_i) \right].$$

The marginal entropy term encourages all components to be used at the batch level, while the conditional entropy term encourages each sample to have a sharper, more informative assignment. Thus, maximizing $\hat{I}_{\mathcal{B}}^{\text{norm}}(X; Z)$ discourages both hard component collapse and soft uniform assignment.

Full objective. Taking the rank-uniformization loss as the reference scale, we optimize

$$\mathcal{L}_{\text{SGCP}} = \mathcal{L}_{\text{rank}} + \lambda \mathcal{L}_{\text{nl}} - \beta \hat{I}_{\mathcal{B}}^{\text{norm}}(X; Z),$$

where λ controls the strength of local density fitting and β controls the strength of stochastic-grouping information maximization.

C Complexity Analysis

Beyond backbone training, SGCP adds three sources of computation: stochastic grouping, local score-law fitting, and conformal calibration on transformed scores. Let N denote the number of training samples used for learning the grouping map and local score laws, N_{cal} the calibration size, and N_{test} the number of test samples. Treating the number of epochs, posterior samples, latent components, latent dimension, classes, and batch size as constants, the training overhead is linear in N , up to an optional $\mathcal{O}(N \log N)$ sorting step for constructing empirical component CDFs.

Calibration requires computing one transformed score per calibration sample and selecting a conformal quantile, costing $\mathcal{O}(N_{\text{cal}} \log N_{\text{cal}})$ with sorting or $\mathcal{O}(N_{\text{cal}})$ with linear-time selection. At test time, the grouping map is evaluated once per input and transformed scores are computed for all candidate labels, giving constant overhead per test sample and total cost $\mathcal{O}(N_{\text{test}})$.

Overall, under the standard mini-batch implementation, the additional complexity beyond backbone training is $\mathcal{O}(N + N_{\text{cal}} \log N_{\text{cal}} + N_{\text{test}})$, or $\mathcal{O}(N + N_{\text{cal}} + N_{\text{test}})$ with linear-time quantile selection.

D Detailed Experimental Setup

D.1 Additional Benchmark Details

Table A2: Summary of datasets and grouping attributes used for evaluation.

Dataset	Modality	#Classes	#Attrs	Sensitive / grouping attributes
SYN	tabular	6	3	$A_1 = \text{gender}, A_2 = \text{phenotype}, A_3 = \text{age}$
Nursery	tabular	4	5	$A_1 = \text{parents}, A_2 = \text{children}, A_3 = \text{finance},$ $A_4 = \text{social}, A_5 = \text{health}$
MIMIC-IV	tabular	2/3	3	$A_1 = \text{minority}, A_2 = \text{gender}, A_3 = \text{public insurance}$
BACH	image	4	-	class label (for class-wise reliability analysis)

In this section, we describe the dataset preprocessing in detail. A summary is provided in Table A2.

Synthetic. The synthetic benchmark is a controlled multiclass cardiopulmonary diagnosis task with six labels, $Y \in \{1, \dots, 6\}$, corresponding to Hypertension, Coronary Artery Disease, Heart Failure, Arrhythmia, Valvular Heart Disease, and Peripheral Artery Disease.

Each sample contains four grouping attributes, $A = (A_{\text{sex}}, A_{\text{pheno}}, A_{\text{age}}, A_{\text{region}})$, and six non-sensitive physiological variables $Z = (Z_1, \dots, Z_6)$, where $Z_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{U}(0, 1)$. We draw

$$A_{\text{sex}} \sim \text{Bernoulli}(1/2), \quad A_{\text{pheno}} \sim \text{Bernoulli}(1/2),$$

with $A_{\text{sex}} \in \{0, 1\}$ denoting Female/Male and $A_{\text{pheno}} \in \{0, 1\}$ denoting Type I/Type II. Age is assigned cyclically over $\{\text{Pediatric, Adult, Geriatric}\}$, and region is sampled uniformly from $\{\text{North, South, East, West}\}$.

To induce intersection-dependent difficulty, we define the disadvantaged subgroup through the XNOR relation between sex and phenotype,

$$G_{\text{bias}} = \{i : A_{\text{sex},i} = A_{\text{pheno},i}\},$$

which includes Female–Type I and Male–Type II individuals.

Labels are generated from a latent multiclass mechanism based on $X_i = (A_i, Z_i)$. We first construct a class-probability vector $\pi(X_i) \in \Delta^5$ over the six disease classes. Outside ambiguous regions of the feature space, labels follow the dominant class implied by $\pi(X_i)$. Inside ambiguous regions, we introduce subgroup-dependent uncertainty by mixing $\pi(X_i)$ with the uniform distribution $u = (1/6, \dots, 1/6)$:

$$Y_i \sim \begin{cases} \text{Cat}((1 - \delta_1)\pi(X_i) + \delta_1 u), & i \in G_{\text{bias}}, \\ \text{Cat}((1 - \delta_0)\pi(X_i) + \delta_0 u), & i \notin G_{\text{bias}}, \end{cases} \quad \delta_1 > \delta_0.$$

The parameters δ_1 and δ_0 control label ambiguity for the disadvantaged subgroup and the remaining population, respectively. Thus, G_{bias} exhibits systematically higher conditional uncertainty while the global task structure remains unchanged, yielding a controlled setting for studying subgroup-dependent calibration difficulty.

Nursery. Nursery is a public multiclass tabular benchmark derived from a hierarchical decision problem for ranking nursery-school applications. Following standard preprocessing, we remove the `recommend` class, which contains only two samples, and encode the remaining labels into four admission-priority levels, $Y \in \{0, 1, 2, 3\}$.

Each instance contains eight categorical covariates describing household background and application context, including parents’ occupation, number of children, housing, financial standing, social conditions, and health status. All categorical variables are label-encoded before training. We use five covariates as candidate sensitive attributes,

$$A = (A_{\text{parents}}, A_{\text{children}}, A_{\text{finance}}, A_{\text{social}}, A_{\text{health}}),$$

corresponding to parents’ occupation, number of children, financial standing, social conditions, and health status. These attributes are used only for subgroup construction and fairness evaluation.

To induce controlled subgroup disadvantage, we follow the intervention protocol of prior work [29, 27]. We define

$$G_{\text{nur}} = \{i : X_{\text{parents},i} = 1\},$$

corresponding to a particular preprocessed category of the `parents` attribute. We then modify only this subgroup by downsampling it to 10% of its original size and injecting label uncertainty. Specifically, labels are perturbed by independent noise $\varepsilon_i \sim \mathcal{U}(-4, 4)$, then rounded and clipped to $\{0, 1, 2, 3\}$. This creates a subgroup that is both underrepresented and noisier than the rest of the population, serving as the disadvantaged group in our fairness evaluation.

MIMIC-IV. We construct our clinical benchmark from MIMIC-IV at the hospital-admission level. Each admission is linked with ICU-stay information and diagnosis-code counts, yielding tabular covariates from demographic, administrative, and utilization records. Age is computed from `anchor_age`, adjusted by the difference between admission year and anchor year, and clipped to $[0, 89]$. We also derive ICU-related features, including ICU length of stay, time from hospital admission to first ICU entry, and the number of diagnosis codes.

Table A3: Dataset specific hyperparameter settings for SGCP.

Hyperparameter	SYN	Nursery	M-ICU	M-Hosp	BACH	Description
<i>(i) Latent calibration components</i>						
K	3	6	3	3	6	number of latent calibration components
d_z	16	8	8	8	64	latent calibration dimension
hidden dim	64	64	32	32	64	hidden dimension of grouping network
<i>(ii) Score law learning</i>						
SGCP train epochs	100	100	100	100	100	number of SGCP training epochs
SGCP learning rate	10^{-3}	10^{-2}	10^{-3}	10^{-3}	10^{-3}	learning rate for SGCP training
λ	0.5	0.05	10^{-3}	10^{-3}	0.5	weight of negative log likelihood loss
β	0.1	10^{-3}	0.5	10^{-3}	5×10^{-5}	weight of balance regularization
M_{train}	3	6	3	3	8	posterior samples during training
M_{eval}	8	8	8	16	8	posterior samples during evaluation
<i>(iii) Backbone & optimization</i>						
train epochs	200	200	40	40	100	number of backbone training epochs
learning rate	5×10^{-4}	10^{-3}	5×10^{-4}	10^{-3}	10^{-3}	learning rate for backbone training
batch size	32	128	128	128	32	mini batch size
weight decay	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	optimizer weight decay

Note: M-ICU denotes MIMIC-IV (ICU delay) and M-Hosp denotes MIMIC-IV (in hospital).

Table A4: Ablation study of SGCP over 20 runs. Results with nominal coverage are shown in blue, optimal values are **boldfaced**, and AvgSize exceeding half of the total number of categories is highlighted in red. CovGap is reported in percent.

Method	SYN			Nursery			MIMIC			BACH		
	Cov.	AvgSize	CovGap	Cov.	AvgSize	CovGap	Cov.	AvgSize	CovGap	Cov.	AvgSize	CovGap
Default SGCP	90.13 ± 0.39	2.225 ± 0.035	4.2 ± 0.7	91.65 ± 0.70	1.089 ± 0.035	11.8 ± 0.7	90.13 ± 0.07	1.107 ± 0.001	0.8 ± 0.0	91.90 ± 2.70	2.224 ± 0.215	6.2 ± 1.1
(a) w/o Stochastic	89.82 ± 1.48	2.348 ± 0.081	7.35 ± 1.18	91.38 ± 1.15	1.123 ± 0.051	13.42 ± 1.76	89.94 ± 0.96	1.053 ± 0.013	2.36 ± 0.48	91.76 ± 4.42	2.447 ± 0.308	8.84 ± 2.35
(b) Feature space	90.96 ± 1.96	2.476 ± 0.127	7.18 ± 1.46	90.92 ± 1.43	1.248 ± 0.079	14.16 ± 2.05	90.48 ± 1.47	1.179 ± 0.044	4.08 ± 1.08	89.61 ± 5.74	3.094 ± 0.446	12.41 ± 3.16
(c) w/o rank	89.21 ± 4.42	1.898 ± 0.246	8.42 ± 4.10	88.63 ± 2.94	1.302 ± 0.118	15.31 ± 2.41	90.03 ± 5.41	1.334 ± 0.146	5.49 ± 3.71	91.18 ± 6.31	3.491 ± 0.541	9.26 ± 3.02
(d) w/o NLL	88.36 ± 1.17	2.074 ± 0.108	5.39 ± 1.08	92.39 ± 0.92	1.446 ± 0.094	12.43 ± 1.19	90.02 ± 2.96	1.351 ± 0.078	2.96 ± 0.77	93.11 ± 2.96	2.846 ± 0.317	7.44 ± 1.47
(e) w/o MI	92.28 ± 3.41	2.438 ± 0.241	6.53 ± 2.72	91.26 ± 2.43	1.149 ± 0.147	14.72 ± 1.42	89.84 ± 6.31	1.047 ± 0.116	4.95 ± 2.31	90.91 ± 7.82	2.347 ± 0.472	9.31 ± 2.11

For subgroup evaluation, we define three binary grouping attributes. The race indicator minority equals 1 if the recorded race string does not contain WHITE; gender_m equals 1 for male patients; and public_insurance equals 1 if the primary insurance is MEDICARE or MEDICAID. These attributes are used only for subgroup construction and fairness evaluation. Model inputs include age, one-hot indicators for insurance type, admission type, and marital status, ICU length of stay, and the number of diagnosis codes; ICU length of stay is excluded for ICU-delay prediction.

We consider two tasks. The first is in-hospital mortality prediction, with binary labels derived from hospital_expire_flag. Results for this task are reported as MIMIC-IV (in-hospital) in Figures 5, A9, and A10. The second is ICU-delay prediction, reported as MIMIC-IV (ICU delay), where admissions are categorized by the time from hospital admission to first ICU entry:

$$Y = \begin{cases} 0, & \text{if ICU delay} \leq 4 \text{ hours,} \\ 1, & \text{if ICU delay} \in (4, 12] \text{ hours,} \\ 2, & \text{if ICU delay} > 12 \text{ hours or no ICU admission occurs.} \end{cases}$$

Together, these tasks provide a real-world clinical benchmark for evaluating fairness across race, gender, and insurance-defined subgroups.

BACH. The Breast Cancer Histology (BACH) dataset [2] is a histopathology benchmark for evaluating subgroup reliability under visual heterogeneity. We use Part A, which contains 400 H&E-stained microscopy images evenly distributed across four classes: normal, benign, in situ carcinoma, and invasive carcinoma. Images were acquired at $20\times$ magnification with resolution 2048×1536 , annotated by two pathologists, and collected from 39 patients, though anonymization prevents complete patient-level tracing. The task is four-class classification with

$$Y \in \{\text{Normal, Benign, In Situ Carcinoma, Invasive Carcinoma}\}.$$

Since demographic sensitive attributes are unavailable, we use the class label as a proxy grouping variable and set $A := Y$. Subgroup reliability is therefore evaluated across disease categories,

Table A5: Performance comparison w.r.t. sample sparsity on the Synthetic Dataset (Target $1 - \alpha = 0.90$). **Takeaway:** SGCP remains consistently close to nominal coverage across all sample sizes while achieving the smallest subgroup coverage gap and comparatively small prediction sets. And the performance of each specific sensitive group is shown in the Fig. A6.

Method	N = 500			N = 1000			N = 2000			N = 4000		
	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$
Marginal	90.15 $\pm$ 6.50	2.150 $\pm$ 0.420	32.5 $\pm$ 6.5	90.20 $\pm$ 4.20	1.950 $\pm$ 0.400	21.5 $\pm$ 6.2	90.12 $\pm$ 4.10	1.820 $\pm$ 0.380	19.7 $\pm$ 6.0	89.52 $\pm$ 1.52	2.309 $\pm$ 0.078	14.5 $\pm$ 1.6
Partial	95.50 $\pm$ 3.20	4.520 $\pm$ 0.300	9.5 $\pm$ 2.8	94.20 $\pm$ 3.00	2.950 $\pm$ 0.280	7.9 $\pm$ 2.5	93.50 $\pm$ 2.80	2.820 $\pm$ 0.250	7.5 $\pm$ 2.2	94.50 $\pm$ 1.12	2.419 $\pm$ 0.048	6.2 $\pm$ 0.7
Exhaustive	100.00 $\pm$ 0.00	6.000 $\pm$ 0.000	12.5 $\pm$ 3.2	98.50 $\pm$ 2.20	4.850 $\pm$ 0.250	10.5 $\pm$ 3.0	98.20 $\pm$ 2.10	4.150 $\pm$ 0.220	7.9 $\pm$ 2.7	95.80 $\pm$ 0.91	3.342 $\pm$ 0.053	6.0 $\pm$ 0.6
RLCP	95.80 $\pm$ 3.80	2.910 $\pm$ 0.520	9.2 $\pm$ 8.2	95.10 $\pm$ 3.20	2.850 $\pm$ 0.480	8.5 $\pm$ 6.8	94.60 $\pm$ 2.90	2.850 $\pm$ 0.460	6.4 $\pm$ 5.5	93.15 $\pm$ 2.45	2.540 $\pm$ 0.305	4.5 $\pm$ 1.8
CluCP	90.22 $\pm$ 2.06	3.112 $\pm$ 0.332	10.7 $\pm$ 0.9	90.62 $\pm$ 0.98	2.722 $\pm$ 0.269	9.9 $\pm$ 0.7	91.11 $\pm$ 1.83	2.756 $\pm$ 0.250	8.9 $\pm$ 1.1	89.81 $\pm$ 1.78	2.426 $\pm$ 0.050	4.9 $\pm$ 0.8
AFCP	90.67 $\pm$ 1.44	3.162 $\pm$ 0.292	6.4 $\pm$ 0.8	91.30 $\pm$ 1.24	2.854 $\pm$ 0.328	7.6 $\pm$ 0.7	91.62 $\pm$ 1.71	2.848 $\pm$ 0.227	7.3 $\pm$ 0.9	91.77 $\pm$ 1.23	2.432 $\pm$ 0.043	4.6 $\pm$ 0.5
FaReG	92.18 $\pm$ 1.26	3.536 $\pm$ 0.388	7.5 $\pm$ 0.6	92.41 $\pm$ 1.13	3.178 $\pm$ 0.288	8.5 $\pm$ 0.6	92.49 $\pm$ 1.57	3.141 $\pm$ 0.238	8.5 $\pm$ 0.9	92.61 $\pm$ 0.98	2.579 $\pm$ 0.029	4.7 $\pm$ 0.5
SGCP	91.20 $\pm$ 1.20	2.050 $\pm$ 0.120	6.1 $\pm$ 0.9	90.80 $\pm$ 1.10	1.820 $\pm$ 0.110	5.1 $\pm$ 0.8	90.50 $\pm$ 1.00	1.710 $\pm$ 0.090	4.7 $\pm$ 0.8	90.13 $\pm$ 0.39	2.225 $\pm$ 0.035	4.2 $\pm$ 0.7

Table A6: Dataset access and preprocessing sources.

Dataset	Source	Access note
Synthetic	Procedurally generated; see Appendix D.1	Generated from the described simulation protocol
Nursery	https://archive.ics.uci.edu/dataset/76/nursery	Public UCI dataset
MIMIC-IV	https://physionet.org/content/mimiciv/3.1/	Credentialed access via PhysioNet
BACH	https://zenodo.org/records/3632035	Public Zenodo record

which naturally differ in visual complexity, intra-class variability, and class-boundary ambiguity. For preprocessing, images are resized to 224×224 , augmented during training with random horizontal and vertical flips, and normalized using ImageNet statistics. We train a visual backbone Φ and use the learned representations $X_i = \Phi(I_i) \in \mathbb{R}^d$ for downstream prediction and evaluation.

In this benchmark, heterogeneity is primarily visual rather than demographic. Differences in tissue morphology, staining, lesion structure, and acquisition conditions can induce distinct class-conditional score laws, $F_c(s) := \mathbb{P}(S \leq s \mid Y = c)$. Thus, BACH complements the tabular benchmarks by testing whether SGCP can maintain reliable conditional coverage when subgroup structure reflects semantic visual heterogeneity rather than explicit sensitive attributes.

D.2 Implementation Details

Computing resources. All experiments are conducted on a workstation equipped with an Intel i5-14600KF CPU, 32 GB of RAM, and a single NVIDIA GeForce RTX 4090 D GPU. All models are implemented in PyTorch, and each experimental setting is trained with multiple random seeds. The hyperparameters for each dataset are reported in Table A3. Data access and code availability are summarized in Tables A6 and A7, respectively.

Open access to data and code. All external datasets used in our experiments are publicly documented and accessible online, subject to their original access requirements. The synthetic dataset is generated procedurally following Appendix D.1. The code for SGCP is provided anonymously in the supplementary material, and baseline implementations are based on the publicly available repositories listed below.

D.2.1 Baseline implementation details

We compare SGCP with six baselines: Partial Equalized, Exhaustive Equalized, CluCP, AFCP, FaReG, and RLCP. Unless otherwise stated, all methods use the same backbone classifier and softmax-based nonconformity score as SGCP. Hyperparameters are adopted from the original papers or released implementations whenever available.

Partial and Exhaustive Equalized. We extend the equalized-coverage framework of [20] to conformal prediction by applying group-conditional calibration over sensitive groups. Partial Equalized calibrates separately for each value of one sensitive attribute, whereas Exhaustive Equalized calibrates over all intersectional groups induced by the full combination of sensitive attributes. In both cases, the classifier is trained on all samples, and only the calibration step is group-conditional.

Table A7: Code availability and baseline implementations.

Method	Code source
SGCP	https://anonymous.4open.science/r/sgcp/
Partial Equalized	https://github.com/yromano/cqr
RLCP	https://github.com/rohanhore/RLCP
CluCP	https://github.com/tiffanyding/class-conditional-conformal
AFCP	https://github.com/FionaZ3696/Adaptively-Fair-Conformal-Prediction
FaReG	https://github.com/Xusr1123/FaReG

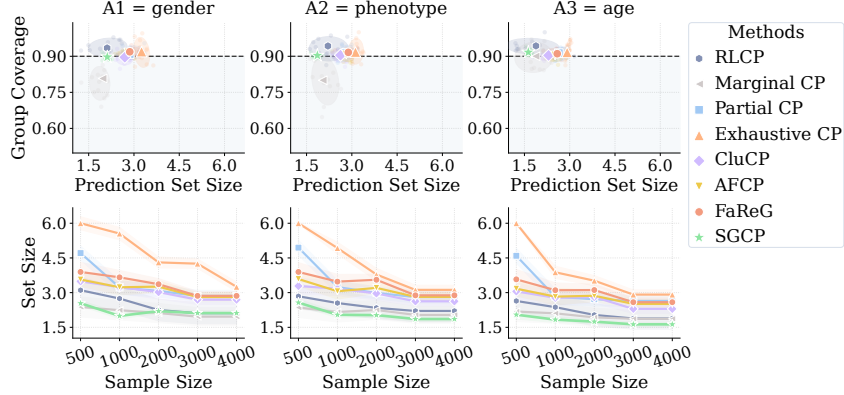


Figure A6: Group coverage on SYN dataset. SGCP remains close to the target coverage across classes while using smaller prediction sets than more conservative baselines.

CluCP. We implement CluCP following its original clustered conformal pipeline. The clustering fraction γ and the number of clusters M are selected using the heuristic proposed in the paper. Class embeddings are clustered by weighted k -means with the default KMeans implementation. Although CluCP targets class-wise rather than subgroup-fair calibration, it provides a useful non-global calibration baseline.

AFCP. We implement AFCP using its original adaptive attribute-selection procedure, directly applying it to the sensitive attributes available in each dataset.

FaReG. We follow the original FaReG representation-learning pipeline: an encoder-decoder first learns latent representations, representation-based low-coverage groups are then identified, and prediction sets are adjusted accordingly. We use the original three-layer MLP encoder and decoder. For Synthetic and Nursery, we adopt the reported hyperparameters: hidden dimensions [64, 32], batch size 500, learning rates 10^{-3} and 10^{-2} , respectively, with $(\beta, \delta, T) = (2.0, 0.3, 20)$ for Synthetic and $(0.1, 0.1, 100)$ for Nursery. For the remaining datasets, we keep the architecture and optimization settings as close as possible to the released implementation.

RLCP. We implement RLCP following its weighted localization procedure. Calibration samples are weighted by their similarity to each test point, and prediction sets are constructed using the corresponding weighted conformal quantile. For all datasets, we use the Gaussian kernel and the bandwidth-selection rule from the original method, without subgroup-specific tuning.

E Additional Experiments

E.1 Ablation Study

Table A4 studies the contribution of each component in SGCP. The default model achieves nominal marginal coverage on all datasets and obtains the best or tied-best CovGap on three out of four datasets, while maintaining the smallest or near-smallest prediction sets. This suggests that the full objective provides the most favorable balance between validity, subgroup reliability, and efficiency.

Removing posterior stochasticity consistently degrades subgroup reliability. The deterministic variant increases CovGap on all datasets, indicating that posterior averaging is useful for stabilizing the local

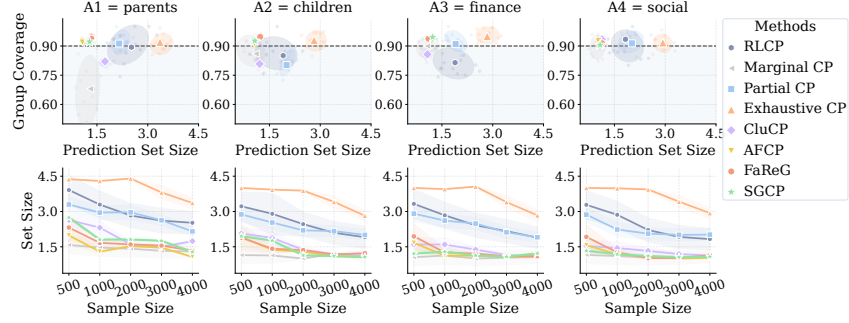


Figure A7: Group coverage on Nursery. Top: subgroup coverage vs. average prediction set size at a sample size of 4000, with the dashed line indicating nominal coverage of 0.9. Bottom: prediction set size varies with subgroup sample size. SGCP stays close to the nominal target while remaining among the most efficient methods across subgroup definitions.

Table A8: Performance comparison w.r.t. sample sparsity on the Nursery Dataset (Target $1 - \alpha = 0.90$). **Takeaway:** SGCP remains near nominal coverage even in the sparse regime and consistently achieves the smallest subgroup coverage gap, while competing methods improve subgroup reliability only with larger prediction sets or at the cost of under- or over-coverage. And the performance of each specific sensitive group is shown in the Fig. A7.

Method	N = 500			N = 1000			N = 2000			N = 4000		
	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$
Marginal	48.25 $\pm$ 8.50	1.120 $\pm$ 0.350	29.5 $\pm$ 6.5	57.59 $\pm$ 8.20	1.080 $\pm$ 0.330	29.1 $\pm$ 5.2	59.72 $\pm$ 7.80	1.050 $\pm$ 0.310	26.2 $\pm$ 5.8	84.15 $\pm$ 5.20	2.170 $\pm$ 0.260	20.4 $\pm$ 5.2
Partial	84.61 $\pm$ 4.80	2.850 $\pm$ 0.310	17.7 $\pm$ 3.8	86.07 $\pm$ 4.50	2.620 $\pm$ 0.280	16.8 $\pm$ 3.5	87.86 $\pm$ 4.20	2.450 $\pm$ 0.250	15.2 $\pm$ 3.2	88.97 $\pm$ 3.80	1.850 $\pm$ 0.200	12.1 $\pm$ 2.5
Exhaustive	100.00 $\pm$ 0.00	4.000 $\pm$ 0.000	19.5 $\pm$ 0.1	98.00 $\pm$ 0.10	3.980 $\pm$ 0.001	18.8 $\pm$ 0.2	98.00 $\pm$ 0.12	3.979 $\pm$ 0.001	17.5 $\pm$ 0.8	96.80 $\pm$ 3.10	2.850 $\pm$ 0.220	15.5 $\pm$ 3.2
RLCP	82.45 $\pm$ 14.20	2.250 $\pm$ 0.580	18.5 $\pm$ 9.5	83.36 $\pm$ 13.80	1.720 $\pm$ 0.550	17.8 $\pm$ 6.2	83.31 $\pm$ 11.50	1.480 $\pm$ 0.510	16.5 $\pm$ 5.8	88.67 $\pm$ 6.20	1.220 $\pm$ 0.410	13.5 $\pm$ 3.8
CluCP	86.26 $\pm$ 5.61	1.506 $\pm$ 0.265	17.3 $\pm$ 4.5	83.80 $\pm$ 3.59	1.516 $\pm$ 0.265	17.1 $\pm$ 2.9	84.42 $\pm$ 2.47	1.261 $\pm$ 0.096	16.9 $\pm$ 1.8	85.08 $\pm$ 1.60	1.121 $\pm$ 0.064	12.8 $\pm$ 1.3
AFCP	92.78 $\pm$ 4.81	1.551 $\pm$ 0.345	18.7 $\pm$ 3.9	94.85 $\pm$ 3.69	1.163 $\pm$ 0.074	16.4 $\pm$ 2.8	93.12 $\pm$ 1.33	1.092 $\pm$ 0.031	15.6 $\pm$ 1.1	92.85 $\pm$ 0.84	1.085 $\pm$ 0.024	14.1 $\pm$ 0.6
FaReG	93.77 $\pm$ 3.66	1.879 $\pm$ 0.515	17.4 $\pm$ 2.9	92.47 $\pm$ 3.04	1.346 $\pm$ 0.146	16.2 $\pm$ 2.3	94.37 $\pm$ 1.44	1.257 $\pm$ 0.027	15.4 $\pm$ 1.1	93.21 $\pm$ 0.62	1.230 $\pm$ 0.012	13.3 $\pm$ 0.6
SGCP	92.72 $\pm$ 1.50	1.320 $\pm$ 0.080	15.1 $\pm$ 1.5	92.10 $\pm$ 1.20	1.210 $\pm$ 0.060	14.6 $\pm$ 1.2	91.97 $\pm$ 1.00	1.120 $\pm$ 0.050	12.3 $\pm$ 1.0	91.65 $\pm$ 0.70	1.089 $\pm$ 0.035	11.8 $\pm$ 0.7

score law rather than committing each sample to a single calibration state. Replacing the learned calibration state with feature-space grouping also leads to larger CovGap and larger prediction sets, especially on MIMIC and BACH. This supports our motivation that similarity in raw feature space is not necessarily the right notion of locality for conformal calibration; what matters is similarity in score behavior.

The loss ablations further clarify the role of each training signal. Without the rank-uniformization loss, coverage becomes less stable and CovGap increases substantially, showing that aligning transformed scores to a common percentile scale is central to subgroup reliability. Removing the density-fitting term also worsens the tradeoff between coverage and efficiency, often producing larger prediction sets, which suggests that likelihood-based fitting helps the learned local law remain faithful to the observed score distribution. Finally, removing the mutual-information term leads to less reliable subgroup calibration and higher variance across runs, consistent with component underuse or less informative stochastic grouping. Overall, the ablation results support the design of SGCP: subgroup reliability improves when local score laws are learned through both distributional fitting and rank-based normalization, and when calibration evidence is shared through informative stochastic components.

E.2 Sensitivity to Hyperparameters and Subgroup Configurations

E.2.1 Effect of sample sparsity

Tables A5–A10 study how performance changes as the effective sample size varies. As the effective sample size decreases, methods based on explicit subgroup equalization or localization tend to become increasingly conservative, while global calibration remains efficient but suffers from large subgroup disparity. By contrast, SGCP degrades more gracefully in the sparse regime and benefits more consistently from additional data as sample size grows. This pattern is strongest on the synthetic

Table A9: Performance comparison w.r.t. sample sparsity on MIMIC-IV (target $1 - \alpha = 0.90$). Each method is run 20 times. **Takeaway:** On the MIMIC-IV dataset, SGCP consistently achieves near-nominal coverage while maintaining the lowest subpopulation disparity (CovGap) in sparse settings ($N = 2000, 5000$). Even with the full dataset, SGCP provides a highly efficient prediction set with excellent fairness metrics compared to conventional baselines.

Method	N = 2000			N = 5000			Full (All Samples)		
	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$
Marginal	90.06 $\pm$ 10.14	1.074 $\pm$ 0.211	34.4 $\pm$ 10.9	90.79 $\pm$ 10.62	1.026 $\pm$ 0.087	29.6 $\pm$ 9.3	90.79 $\pm$ 10.62	1.226 $\pm$ 0.087	27.1 $\pm$ 7.1
Partial	95.73 $\pm$ 3.62	1.564 $\pm$ 0.243	11.0 $\pm$ 5.6	94.50 $\pm$ 4.54	1.335 $\pm$ 0.207	9.2 $\pm$ 4.0	94.50 $\pm$ 4.54	1.258 $\pm$ 0.140	8.0 $\pm$ 3.0
Exhaustive	100.00 $\pm$ 0.00	3.000 $\pm$ 0.000	18.7 $\pm$ 8.0	98.86 $\pm$ 5.21	2.989 $\pm$ 0.004	21.4 $\pm$ 5.3	96.45 $\pm$ 4.22	2.800 $\pm$ 0.142	19.7 $\pm$ 3.2
RLCP	95.10 $\pm$ 2.40	1.650 $\pm$ 0.120	4.2 $\pm$ 2.6	93.20 $\pm$ 2.10	1.480 $\pm$ 0.150	3.5 $\pm$ 1.4	92.50 $\pm$ 2.80	1.350 $\pm$ 0.100	1.5 $\pm$ 0.8
CluCP	95.11 $\pm$ 1.92	1.654 $\pm$ 0.192	4.5 $\pm$ 1.5	93.19 $\pm$ 1.28	1.422 $\pm$ 0.059	3.5 $\pm$ 1.0	92.22 $\pm$ 1.28	1.302 $\pm$ 0.059	3.5 $\pm$ 0.8
AFCP	91.81 $\pm$ 2.51	1.400 $\pm$ 0.075	3.9 $\pm$ 1.0	91.40 $\pm$ 1.12	1.114 $\pm$ 0.045	3.3 $\pm$ 0.8	-	-	-
FaReG	95.78 $\pm$ 2.09	1.601 $\pm$ 0.078	4.1 $\pm$ 1.2	92.58 $\pm$ 1.05	1.357 $\pm$ 0.050	3.0 $\pm$ 1.1	90.23 $\pm$ 0.09	1.174 $\pm$ 0.001	0.5 $\pm$ 0.0
SGCP	90.12 $\pm$ 1.91	1.060 $\pm$ 0.068	3.3 $\pm$ 0.7	90.48 $\pm$ 1.36	1.116 $\pm$ 0.050	2.6 $\pm$ 0.5	90.17 $\pm$ 0.07	1.107 $\pm$ 0.001	0.8 $\pm$ 0.0

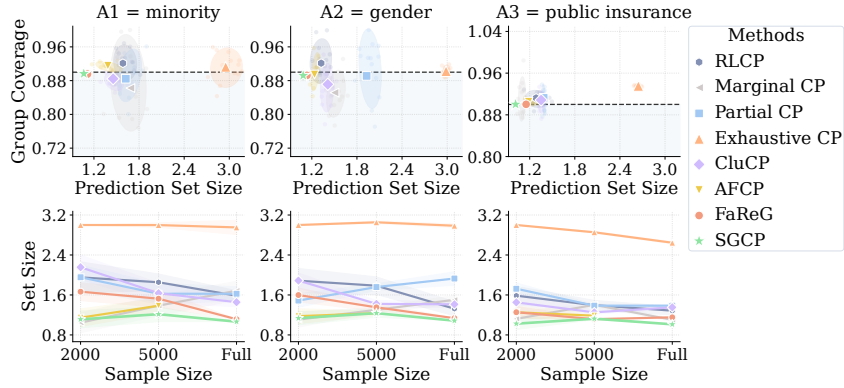


Figure A8: Group coverage on MIMIC-IV dataset. SGCP remains close to the target coverage across classes while using smaller prediction sets than more conservative baselines.

benchmark and on BACH, where subgroup heterogeneity is more structured, and remains visible on MIMIC-IV, where SGCP is particularly competitive in the sparse and moderate sample regimes.

For a clearer comparison, we also provide separate box plots for each metric across methods and sample sizes. Figures A9, A10, and 5 show the distributions of overall coverage, average prediction set size, and subgroup coverage gap under varying sample sparsity. Each box summarizes 20 independent trials with different random seeds for a given dataset and sample-size setting. Across datasets, SGCP remains near the nominal target with competitive set sizes and consistently low subgroup disparity, especially in sparse regimes where competing methods become more conservative or more variable.

E.2.2 Effect of the number of latent components

Figure A11 visualizes the learned component densities on SYN under different choices of K . When K is small, the model learns a few distinct score law patterns that separate low score and high score calibration regimes. As K increases, additional components tend to overlap with existing ones or become nearly flat, suggesting that the dominant calibration structure has already been captured by a smaller number of components. This behavior indicates that K controls the granularity of stochastic sharing rather than simply increasing model capacity. In practice, overly large K may introduce redundant components, while a moderate K provides a compact set of reusable score law building blocks for constructing sample dependent local score laws.

Figure A12 shows a similar sensitivity pattern on MIMIC-IV, but with a different score structure. Most learned components concentrate near small nonconformity scores, reflecting the strong mass of the empirical score distribution in the low score region. Increasing K refines this region and introduces broader components that capture the remaining tail behavior. However, when K becomes large, several components overlap substantially or become weakly differentiated. This suggests that

Table A10: Performance comparison w.r.t. sample sparsity on BACH (target $1 - \alpha = 0.90$). Each method is run 20 times. **Takeaway:** Under extreme sparsity, most methods become conservative, but with the increase in sample size, SGCP recovers the fastest, with near-nominal coverage, a smaller prediction set, and lower subpopulation disparity, outperforming other baseline methods. The performance of each individual subgroup is further shown in Fig. 4.

Method	N = 50			N = 150			N = 200			N = 300		
	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$	Cov. $\uparrow$	AvgSize $\downarrow$	CovGap $\downarrow$
Marginal	89.34 $\pm$ 3.35	3.700 $\pm$ 0.474	19.0 $\pm$ 2.1	90.50 $\pm$ 3.86	3.604 $\pm$ 0.494	17.6 $\pm$ 2.0	90.69 $\pm$ 15.75	3.545 $\pm$ 0.445	15.9 $\pm$ 2.9	89.20 $\pm$ 9.64	2.894 $\pm$ 0.267	15.7 $\pm$ 1.7
Partial	98.33 $\pm$ 2.66	3.735 $\pm$ 0.316	7.8 $\pm$ 1.0	97.45 $\pm$ 3.25	3.711 $\pm$ 0.290	6.2 $\pm$ 0.6	97.17 $\pm$ 11.56	3.460 $\pm$ 0.258	7.5 $\pm$ 0.7	96.57 $\pm$ 8.08	3.385 $\pm$ 0.299	7.6 $\pm$ 1.6
Exhaustive	100.00 $\pm$ 0.00	4.000 $\pm$ 0.105	11.7 $\pm$ 0.7	100.00 $\pm$ 0.00	4.000 $\pm$ 0.115	9.2 $\pm$ 0.5	98.73 $\pm$ 0.63	3.837 $\pm$ 0.233	11.0 $\pm$ 0.1	97.41 $\pm$ 6.89	3.867 $\pm$ 0.118	11.6 $\pm$ 1.5
RLCP	91.74 $\pm$ 3.06	2.855 $\pm$ 0.440	8.2 $\pm$ 3.6	92.25 $\pm$ 3.80	2.763 $\pm$ 0.422	7.9 $\pm$ 3.8	91.99 $\pm$ 10.88	2.611 $\pm$ 0.411	6.4 $\pm$ 3.8	91.88 $\pm$ 10.56	2.311 $\pm$ 0.319	8.1 $\pm$ 2.5
CluCP	97.80 $\pm$ 3.14	3.936 $\pm$ 0.199	8.7 $\pm$ 1.5	96.60 $\pm$ 3.29	3.882 $\pm$ 0.180	8.1 $\pm$ 1.9	98.34 $\pm$ 0.50	3.985 $\pm$ 0.110	10.0 $\pm$ 2.6	95.91 $\pm$ 3.59	3.257 $\pm$ 0.220	10.7 $\pm$ 2.6
AFCP	100.00 $\pm$ 0.50	4.000 $\pm$ 0.120	10.0 $\pm$ 1.2	96.50 $\pm$ 2.64	3.817 $\pm$ 0.116	8.3 $\pm$ 1.3	95.80 $\pm$ 2.39	3.447 $\pm$ 0.297	7.5 $\pm$ 1.2	95.00 $\pm$ 2.60	2.830 $\pm$ 0.237	6.3 $\pm$ 1.7
FaReG	99.00 $\pm$ 2.00	3.946 $\pm$ 0.153	9.3 $\pm$ 1.3	98.40 $\pm$ 1.78	3.914 $\pm$ 0.212	8.4 $\pm$ 1.8	96.80 $\pm$ 1.93	3.769 $\pm$ 0.250	7.7 $\pm$ 1.5	93.60 $\pm$ 3.40	2.959 $\pm$ 0.239	7.4 $\pm$ 1.5
SGCP	96.44 $\pm$ 2.61	3.842 $\pm$ 0.136	6.8 $\pm$ 2.2	92.63 $\pm$ 4.04	3.344 $\pm$ 0.240	5.0 $\pm$ 1.58	92.22 $\pm$ 2.39	2.370 $\pm$ 0.197	4.6 $\pm$ 2.1	92.00 $\pm$ 2.70	2.224 $\pm$ 0.215	6.2 $\pm$ 1.1

additional components mainly refine existing calibration patterns rather than uncover entirely new score regimes. Together, the SYN and MIMIC-IV results show that a moderate number of latent components is sufficient to capture the dominant score law structure, while larger K may lead to redundant calibration sharing.

E.2.3 Hyperparameter sensitivity

Figures A13 and A14 examine the sensitivity of SGCP w.r.t. several key hyperparameters, including λ , β , the number of latent calibration components K , and the number of evaluation latent samples M_{eval} , under different D_1 split ratios. Across both datasets, the hyperparameter curves are generally flat, indicating that the performance of SGCP does not rely on delicate tuning.

On the SYN, this stability is particularly clear: coverage remains close to the nominal target across all tested settings, while the changes in average prediction set size and subgroup coverage gap are modest. The most visible effect comes from the D_1 split ratio, where allocating more data to score law learning consistently improves subgroup reliability and reduces prediction set size. On MIMIC-IV, it show a similar pattern. Coverage remains tightly concentrated around the nominal level across all tested settings, with modest changes in AvgSize and CovGap. Compared with hyperparameter choices, allocation training data to local score law learning has a more visible effect on the final tradeoff. Overall, these results suggest that the gains of SGCP stem primarily from its structured local score law design rather than from aggressive hyperparameter tuning.

E.3 Wall-Clock Time

We compare the wall-clock time of the full SGCP pipeline with all baselines under varying calibration sample sizes. Figure A15 reports the average runtime over repeated runs, with error bars indicating one standard deviation. Marginal CP and Partial CP are the fastest methods, as they require only global calibration. RLCP and SGCP incur additional cost from sample-dependent calibration, but remain moderate and scale smoothly with sample size. AFCP is the most expensive baseline by a large margin, reflecting its repeated adaptive subgroup selection and conformal recalibration, while CluCP and FaReG lie in between. Overall, SGCP provides a favorable computational tradeoff, remaining practical across all tested sample sizes while achieving stronger subgroup reliability than simpler global or rigid group based baselines.

E.4 Limitations

Our study has several limitations. All experiments are conducted on retrospective benchmarks, so prospective validation is needed to understand the practical impact of subgroup reliable prediction sets in real clinical workflows. In addition, our current framework is developed for multiclass classification and does not address regression or other prediction settings. Extending the method beyond classification and evaluating it under deployment level constraints are left to future work.

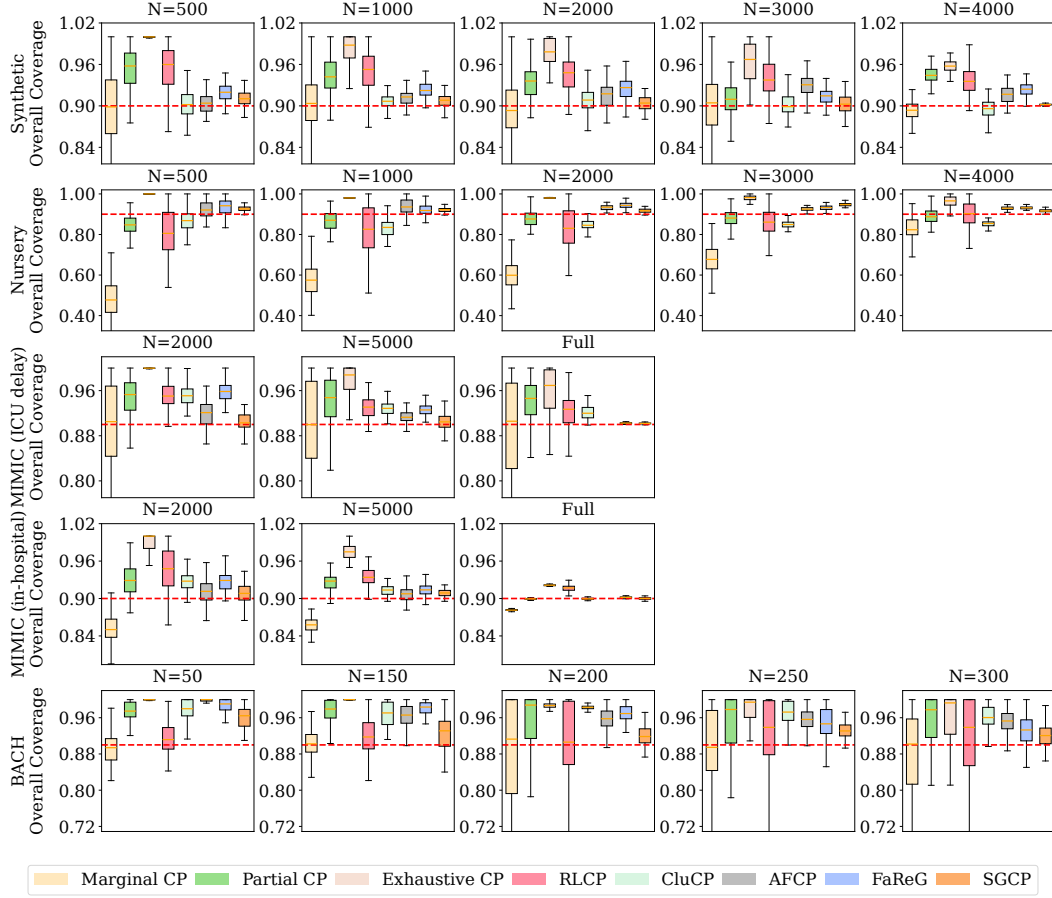


Figure A9: Distribution of overall coverage over repeated runs under varying sample sparsity. The dashed line marks the nominal target coverage of 0.9. **Takeaway:** SGCP remains consistently close to the target across datasets and sample sizes, whereas global calibration may under-cover in sparse regimes and stronger equalization baselines often become overly conservative.

F Supplement to Related Work

Group-conditional and fairness-aware conformal prediction. A substantial body of work studies fairness-aware conformal prediction through coverage parity across groups that are either specified in advance or adaptively selected from data. Vadlamani et.al [23] develop a unified post-hoc framework that extends classical fairness notions to prediction sets by adjusting conformity thresholds in a fairness-aware manner. AFCEP [29] study conformal classification with equalized coverage for adaptively selected groups, selecting features that reflect potential model limitations or biases in order to balance predictive efficiency and group reliability. FAREG [27] goes beyond predefined sensitive groups by learning representation-based subgroups and adaptively promoting equalized coverage for populations that would otherwise receive unfair treatment. SAGCCI [11] examines the tradeoff between fairness and efficiency by clustering protected groups with similar conformal score distributions and by using surrogate outcomes to improve threshold estimation. This is particularly close in spirit to our work in its use of score-distribution similarity to share calibration information. The key distinction lies in the granularity of the sharing mechanism. SAGCCI shares calibration information across clusters of observed protected groups, whereas SGCP learns a sample-dependent stochastic mixture over latent calibration components and thereby induces a local score law for each individual sample. Consequently, the calibration-sharing structure learned by SGCP need not align with protected-group labels and does not rely on surrogate outcomes.

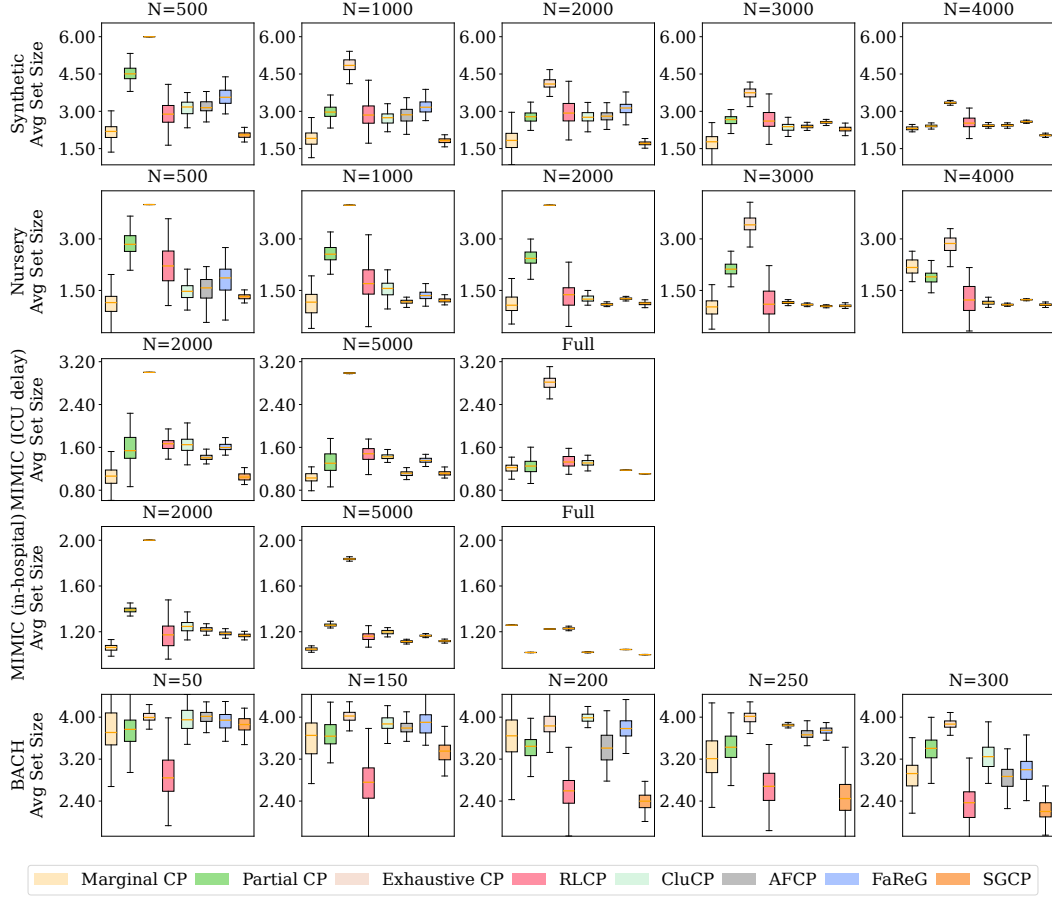


Figure A10: Distribution of average prediction set size over repeated runs under varying sample sparsity. Smaller values indicate better efficiency. **Takeaway:** SGCP remains among the most efficient methods across datasets and sample sizes, often matching or improving upon the smallest prediction sets while avoiding the severe conservatism of explicit equalization and localized baselines.

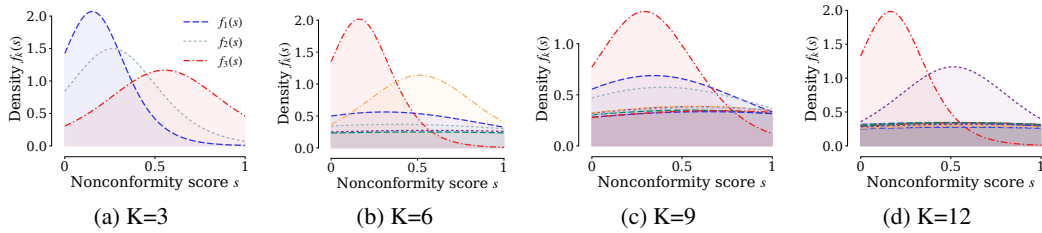


Figure A11: Effect of K on learned latent component densities for SYN. Increasing K refines the score law components, but large K introduces overlapping or redundant densities.

Structured sharing and learned calibration groups. A related line of work improves upon a single global calibration rule by sharing calibration evidence across structured partitions. Clustered conformal prediction [8] groups labels with similar score distributions in many-class settings and performs calibration at the cluster level, thereby interpolating between standard conformal prediction and fully class-conditional conformal prediction. Conceptually similar ideas also arise beyond conformal prediction. For instance, VIR [26] smooths probabilistic uncertainty estimates across neighboring label regions in imbalanced regression. Together, these works underscore the value of structured sharing: calibration and uncertainty estimation can often benefit from borrowing information across statistically related regions, rather than treating all samples globally or calibrating each group in

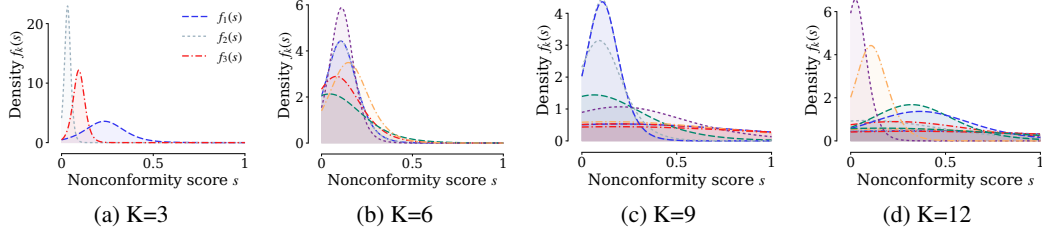


Figure A12: Effect of K on learned latent component densities for MIMIC-IV. Larger K refines the dominant low score region and tail behavior, but also introduces increasingly overlapping components.

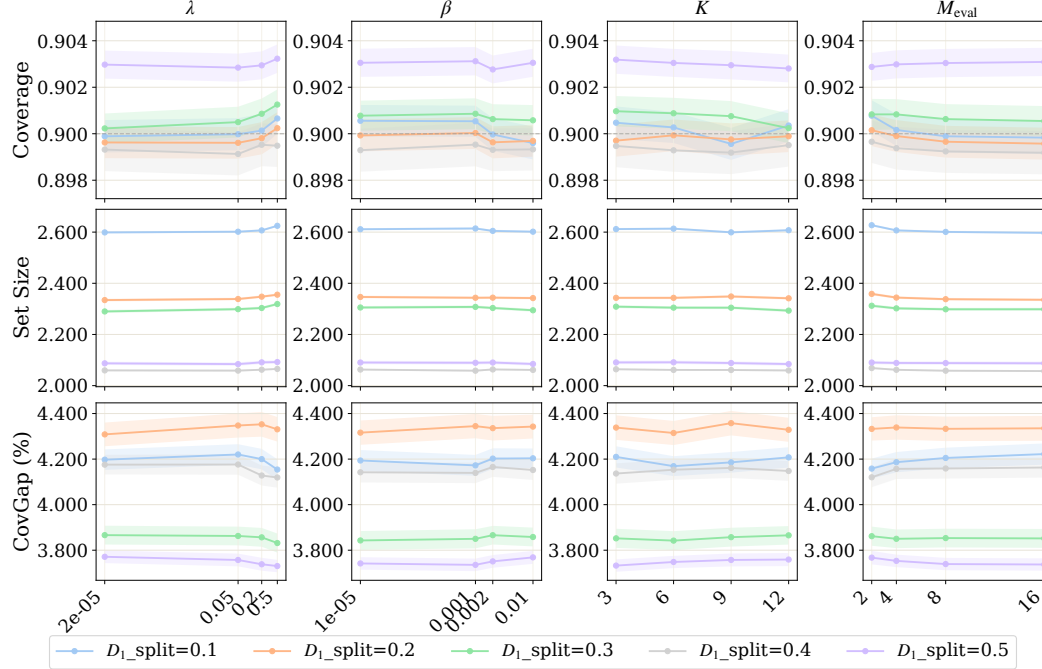


Figure A13: Hyperparameter sensitivity of SGCP on SYN. We vary λ_{nil} , λ_{bal} , K , and M_{eval} , and report coverage, average set size, and subgroup coverage gap under different D_1 split ratios. Shaded bands show one standard deviation. **Takeaway:** SGCP is largely insensitive to these hyperparameters, while larger samples for score-law learning reduce covgap.

isolation. Another line of work enlarges or learns the group structure used for conditional calibration. Kandinsky Conformal Prediction [3] generalizes group-conditional conformal inference to overlapping and fractional group functions defined on (X, Y) , thereby expanding the collection of groups over which weighted coverage deviations can be controlled. Recent work on identifying homogeneous and interpretable groups for conformal prediction likewise uses nonconformity-score behavior to learn explicit partitions of the input space before applying group-conditional calibration [18]. These methods are closely related to ours in their shared emphasis on calibration-relevant structure beyond predefined sensitive attributes.

Our work adopts a complementary perspective. Instead of specifying a family of group functions or assigning each sample to a hard discovered partition, SGCP learns a sample-dependent local score law through stochastic mixing over shared latent calibration components. This enables calibration information to be shared according to learned score behavior, while the final prediction set is still obtained through standard split conformal calibration applied to transformed scores.

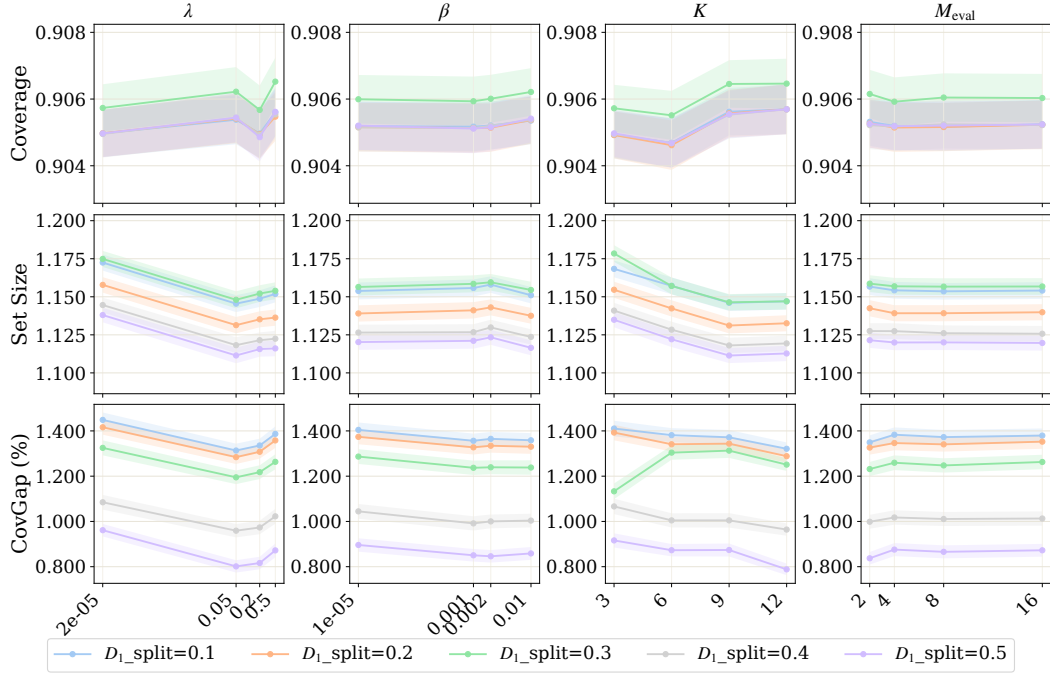


Figure A14: Hyperparameter sensitivity of SGCP on MIMIC-IV (ICU delay). **Takeaway:** SGCP is largely insensitive to these hyperparameters. Coverage remains close to the nominal level, and most of the variation in efficiency and subgroup reliability is driven by the D_1 split ratio rather.

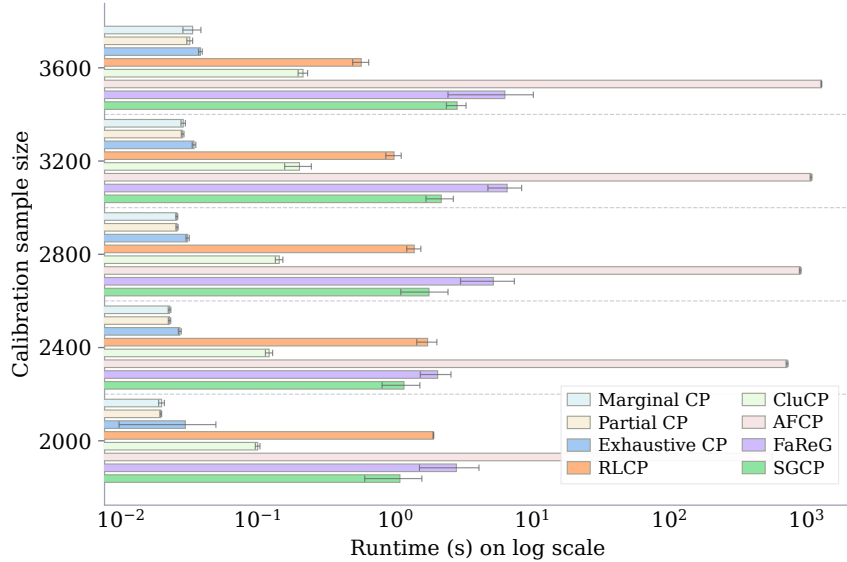


Figure A15: Wall clock time comparison w.r. calibration sample. We report the wall-clock time for each method, averaged over repeated runs; error bars indicate one standard deviation. The horizontal axis is shown on a log scale. **Takeaway:** SGCP has moderate computational cost and is faster than AFCP, while remaining competitive with other subgroup aware baselines across all sample sizes.