CP4SBI: Local Conformal Calibration of Credible Sets in Simulation-Based Inference

Luben M. C. Cabezas[1, 2], Vagner S. Santos[1], Thiago R. Ramos[1], Pedro L. C. Rodrigues[3], and Rafael Izbicki[1]

[1]Department of Statistics, Federal University of São Carlos
[2]Institute of Mathematics and Computer Science, University of São Paulo
[3]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK

Abstract

Current experimental scientists have been increasingly relying on simulation-based inference (SBI) to invert complex non-linear models with intractable likelihoods. However, posterior approximations obtained with SBI are often miscalibrated, causing credible regions to undercover true parameters. We develop `CP4SBI`, a model-agnostic conformal calibration framework that constructs credible sets with local Bayesian coverage. Our two proposed variants, namely local calibration via regression trees and CDF-based calibration, enable finite-sample local coverage guarantees for any scoring function, including HPD, symmetric, and quantile-based regions. Experiments on widely used SBI benchmarks demonstrate that our approach improves the quality of uncertainty quantification for neural posterior estimators using both normalizing flows and score-diffusion modeling.

Keywords: Simulation-based inference, Credible regions, Conformal prediction, Local coverage, Conditional coverage

1. Introduction

In recent years, the machine learning research community has shown growing interest in tackling challenging open problems across a range of experimental sciences, including biology [Chicco, 2017, Min et al., 2017], astrophysics [Freeman et al., 2017, Rodríguez et al., 2022], neuroscience [Lueckmann et al., 2017], and fluid dynamics [Usman et al., 2021, Vinuesa and Brunton, 2022]. A fundamental challenge common to all these domains—and the focus of this work—is the task of inferring the parameter values of statistical models that best explain observed data [Tarantola, 2005, Casella and Berger, 2024]. These statistical models are typically formalized as stochastic simulators, which generate synthetic data $\mathbf{x} \in \mathcal{X}$ given input parameters $\theta \in \Theta$. Although it is easy to generate data from such simulators, these models generally lack a tractable likelihood

function, making classical Bayesian inference techniques for computing posterior distributions, such as Markov Chain Monte Carlo (MCMC), infeasible.

Simulation-based inference (SBI) [Cranmer et al., 2020] is a powerful framework for likelihood-free Bayesian inference, allowing posterior estimation from synthetic data generated by simulators. Recent advances in deep generative models enable SBI methods to approximate posterior distributions using expressive neural density estimators. Despite notable empirical successes, SBI is still maturing and faces key challenges that hinder wider adoption. These include sensitivity to model misspecification [Wehenkel et al., 2025], difficulties with hierarchical or weakly identified models [Rodrigues et al., 2021], and limited theoretical understanding of posterior consistency with finite data [Linhart et al., 2023]. A meta-analysis by Hermans et al. 2021 shows that current SBI methods often produce overconfident and miscalibrated posteriors. That is, credible regions $C(\mathbf{x}_{\text{obs}})$ built for parameters for θ using the approximate posterior $\widehat{p}(\theta \mid \mathbf{x}_{\text{obs}})$, where $\mathbf{x}_{\text{obs}}$ is the observed data, typically fail to achieve nominal coverage. This compromises the reliability of downstream tasks that rely on these posteriors [Murphy, 2022].

To mitigate the issues of miscalibrated posterior distributions, recent work draws on the success of Conformal prediction (CP) for prediction methods [Shafer and Vovk, 2008, Angelopoulos et al., 2023] to construct credible regions with finite-sample coverage guarantees in an SBI setting. However, vanilla CP yields only marginally calibrated regions [Patel et al., 2023, Baragatti et al., 2024], meaning that

$$\mathbb{P}(\theta \in C(\mathbf{X})) = 1 - \alpha, \tag{1}$$

which is often refered to as *marginal coverage*.

This is less informative than *conditional coverage* for a given observation $\mathbf{x}_{\text{obs}}$, defined as

$$\mathbb{P}(\theta \in C(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}_{\text{obs}}) = 1 - \alpha, \tag{2}$$

a property that is central to Bayesian inference. Moreover, current CP methods require access to a closed-form approximation to the posterior $p(\theta \mid \mathbf{x})$, which is not necessarily possible for all classes of generative models currently used for SBI, such as diffusion models [Linhart et al., 2024] and flow matching [Wildberger et al., 2023].

We address this gap by incorporating *local* CP techniques [Izbicki et al., 2022, Cabezas et al., 2025a, Dheur et al., 2025] into the SBI framework. Our method, `CP4SBI`, supports a broad range of posterior approximators, including both density- and sample-based approaches. As illustrated in Figure 1, `CP4SBI` yields more reliable, observation-specific credible regions, improving calibration across different $\mathbf{x}_{\text{obs}}$ and refining standard globally calibrated methods. It only requires a calibration dataset $\mathcal{D}_{\text{cal}} = \{(\theta_i, \mathbf{x}_i)\}_{i=1}^{B}$, where each $(\theta, \mathbf{X})$ is either simulated from the prior and the model, or obtained by withholding a small portion (e.g., 20%) of the training data used for the posterior approximator.

Paper Outline. Section 1.1 reviews related work. Section 1.2 presents our contributions. Background is shown in Section 2, followed by methodology in Section 3. Section 4 presents theoretical results, and Section 5 shows experiments on SBI benchmarks. Section 6 concludes the paper. Appendices A to C contain algorithms, proofs, and additional results and details.

1.1. Relation to Other Work

Approximations of posterior distributions. Various SBI strategies for posterior estimation have been developed over recent years. These include initial approaches like Approximate Bayesian Computation (ABC) [Marin et al., 2012, Sisson et al., 2018], which approximates the posterior for a specific observation $\mathbf{x}_{\text{obs}}$ via rejection sampling using a specified distance function and rejection sampling or other MCMC methods. Following this, amortized methods started being used to

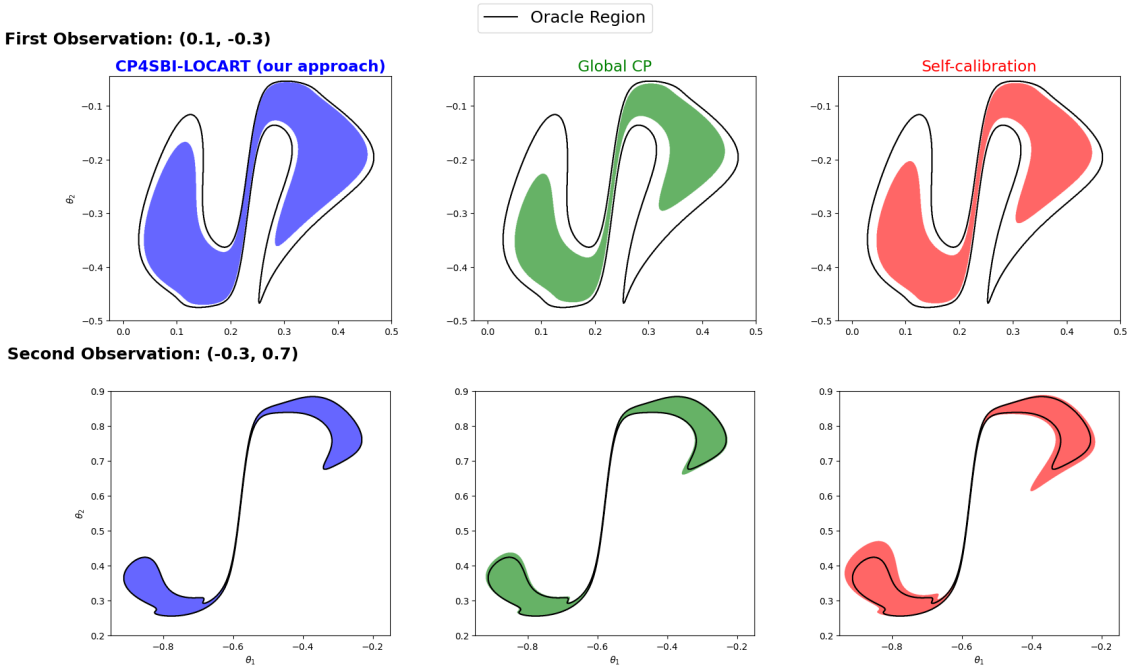

Figure 1: Comparison of credible regions on the Two Moons benchmark for two distinct observations $\mathbf{x}_{\text{obs}}$. The proposed `LoCart CP4SBI` method (blue) produces regions that more closely match the oracle regions (black) with correct coverage, compared to the global conformal (green) and standard self-calibration (red). This highlights the benefits of local adaptivity in improving the quality of credible region estimates.

estimate the posterior distribution and related quantities more efficiently [Cranmer et al., 2015, Gutmann and Corander, 2016, Izbicki et al., 2019]. These methods include neural posterior estimation (NPE) [Papamakarios and Murray, 2016, Lueckmann et al., 2017, Greenberg et al., 2019, Deistler et al., 2022], which directly approximates the posterior density $p(\theta \mid \mathbf{x})$; neural likelihood estimation (NLE) [Papamakarios et al., 2019, Frazier et al., 2023], which estimates the intractable likelihood function $p(\mathbf{x} \mid \theta)$; and density ratio estimators [Izbicki et al., 2014, Hermans et al., 2020, Durkan et al., 2020, Dalmasso et al., 2020, 2024], which estimate likelihood ratios such as $p(\mathbf{x} \mid \theta)/p(\mathbf{x})$. Additionally, a subsequent development in the field has been the adaptation of more implicit generative models, such as flow-matching and diffusion models, to SBI [Wildberger et al., 2023, Geffner et al., 2023, Linhart et al., 2024]. These advanced techniques aim to faithfully generate samples directly from the posterior $p(\theta \mid \mathbf{x})$ without requiring the estimation of a closed-form posterior density. Our approach uses such methods, but provides means to better calibrate credible sets derived from them.

Calibration of approximators and recalibration techniques. Despite the breadth of SBI's recent research and its impact on various scientific fields, Hermans et al. 2021 has revealed a prevalent challenge to all of its variants: they can suffer from statistical miscalibration and produce overconfident estimates in practical settings. Consequently, a significant body of recent work [Hermans et al., 2021, Dey et al., 2022, Bon et al., 2022, Delaunoy et al., 2022, Chung et al., 2024] has focused on improving posterior density calibration, with two main approaches: (1) Post-hoc procedures that recalibrate existing estimates, e.g., using local Probability Integral Transform (PIT) diagnostics [Zhao et al., 2021, Dey et al., 2022], or adjusting samples and probability estimates with Highest Density Region (HDR) corrections [Chung et al., 2024]; and (2) Pre-fitting procedures that integrate calibration during the estimation process, e.g., ensembling posterior estimations [Hermans et al., 2021] or regularizing estimator formulations [Delaunoy et al., 2022, 2023].

The main limitation of methods from both categories is that they do not offer explicit finite-sample guarantees for the validity of credible regions.

Conformal Prediction applied to approximate posteriors. Standard Conformal Prediction (CP) techniques, when applied to SBI, primarily provide post-hoc marginal coverage guarantees. Existing work in this direction includes Baragatti et al. 2024, which adapts CP for Approximate Bayesian Computation methods, and Patel et al. 2023, which explores its use in variational inference —a setting, like SBI, where one also works with estimates of the posterior distribution. Although these methods are effective in improving marginal coverage, they do not provide locally or approximately conditionally valid credible regions in general settings. Moreover, Patel et al. 2023 requires a closed-form approximation of the posterior, which is typically unavailable in the generative models used for SBI, where only sample-based estimates are provided.

Frequentist confidence sets for SBI. An emerging line of research aims to construct confidence sets with frequentist coverage in SBI settings, i.e., ensuring that $\mathbb{P}(\theta \in C(\mathbf{X}) \mid \theta) = 1 - \alpha$ for all $\theta \in \Theta$ as opposed to Equation (2). This includes methods based on Monte Carlo sampling [Cranmer et al., 2015], quantile regression [Dalmasso et al., 2020, 2024, Masserano et al., 2023, Carzon et al., 2025], and conformal prediction [Cabezas et al., 2024]. Our work does not target frequentist coverage but instead focuses on improving Bayesian posterior calibration.

1.2. Novelty

We introduce `CP4SBI`, a framework that integrates conformal prediction techniques into the simulation-based inference pipeline to produce calibrated credible regions. Our main contributions are:

- **Guaranteed Marginal Coverage:** Unlike standard credible sets from approximate posteriors which often suffer from miscalibration [Hermans et al., 2021], `CP4SBI` provides non-asymptotic guaranteed marginal coverage (Equation 1) by reinterpreting Bayesian scores as nonconformity scores.
- **Enhanced Local and Conditional Adaptivity:** We move beyond simple marginal guarantees, achieving stronger coverage criteria:
  - **`CDF CP4SBI`:** Achieves asymptotic conditional coverage as the estimate $\widehat{p}(\theta \mid \mathbf{x})$ gets closer to the true posterior distribution. This implies that $\mathbb{P}(\theta \in C(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}_{\text{obs}})$ typically gets closer to $1 - \alpha$ as the number of simulated training samples increases. This is achieved by efficiently estimating the conditional CDF of scores, reusing the already-trained posterior approximation.
  - **`LoCart CP4SBI`:** Provides local coverage by partitioning the data space with a regression tree, adapting credible region sizes based on inference difficulty. That is, **`LoCart CP4SBI`** achieves $\mathbb{P}(\theta \in C(\mathbf{X}) \mid \mathbf{X} \in A) = 1 - \alpha$, where A is a subset of $\mathcal{X}$ of datasets close to $\mathbf{x}_{\text{obs}}$ according to a data-driven metric. Moreover, it offers asymptotic conditional coverage in the sense that $\mathbb{P}(\theta \in C(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}_{\text{obs}})$ gets closer to $1 - \alpha$ as the number of *calibration* samples B increases.
- **Generality and Flexibility:** `CP4SBI` is a model-agnostic framework that can be applied on top of any posterior approximator, including density-based methods (e.g., NPE) and sample-based generative models (e.g., diffusion models, flow matching). Furthermore, it is compatible with various scoring functions, allowing for the construction of calibrated Highest Posterior Density regions, symmetric regions, or other custom credible sets. It also accommodates nuisance parameters and parameter transformations. This flexibility makes it a broadly applicable tool for enhancing the reliability of SBI methods.

2. Background

2.1. Conformal Prediction

Conformal methods have recently emerged as a powerful framework for constructing prediction regions under minimal assumptions [Vovk et al., 2005, Shafer and Vovk, 2008, Angelopoulos et al., 2023, Izbicki, 2025]. Given exchangeable data $\{(Y_i, \mathbf{X}_i)\}_{i=1}^{m+1}$, these methods construct a prediction set $C(\cdot)$ using the first m pairs such that $\mathbb{P}\big(Y_{m+1} \in C(\mathbf{X}_{m+1})\big) \geq 1-\alpha$, where the probability is taken with respect to the joint distribution of all $m+1$ samples.

A standard split-conformal method begins by defining a nonconformity score $s : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$, which quantifies how well a candidate output $y \in \mathcal{Y}$ conforms to the input $\mathbf{x} \in \mathcal{X}$ given a fitted regression model. In regression problems, a common choice of score is the absolute residual $s(y; \mathbf{x}) = \big|y - \widehat{\mathbb{E}}[Y \mid \mathbf{x}]\big|$, where $\widehat{\mathbb{E}}[Y \mid \mathbf{x}]$ is an estimate of the regression function $\mathbb{E}[Y \mid \mathbf{x}]$ fitted on a subset of the labeled data reserved for training. The prediction region then takes the form $C(\mathbf{x}) = \{y \in \mathcal{Y} : s(y; \mathbf{x}) \leq t_{1-\alpha}\}$, where $t_{1-\alpha}$ is the $(1+1/n)(1-\alpha)$-quantile of the conformity scores $s(Y_i; \mathbf{X}_i)$ evaluated on a calibration set of size n disjoint from the training data used to construct the nonconformity score s.

Conformal methods guarantee marginal coverage [Papadopoulos et al., 2008, Vovk, 2012, Lei and Wasserman, 2014, Valle et al., 2023], but generally do not ensure conditional coverage. Without strong assumptions on the data-generating distribution, achieving exact conditional coverage either requires trivial (often unbounded) prediction sets or results in coverage falling below the target level for some covariate values [Lei and Wasserman, 2014, Foygel Barber et al., 2021]. To address this, several recent methods aim to achieve asymptotic conditional coverage as $m \to \infty$. One approach is to construct conformity scores whose conditional distribution given $\mathbf{X}$ is approximately independent of $\mathbf{X}$. Examples include conformalized quantile regression [Romano et al., 2019], distributional conformal prediction [Chernozhukov et al., 2021], Dist-split [Izbicki et al., 2020], HPD-split [Izbicki et al., 2022], EPICSCORE [Cabezas et al., 2025b], and Plassier et al. 2025. In this work, we are particularly interested in the CDF-conformal score introduced by Dheur et al. [2025, Eq. 14]. Given a nonconformity score $s(y; \mathbf{x})$ and an estimate $\widehat{F}(\cdot \mid \mathbf{x})$ of its conditional cumulative distribution function, this approach defines the transformed score.

$$s'(y; \mathbf{x}) = \widehat{F}(s(y; \mathbf{x}) \mid \mathbf{x}).$$

This transformation improves conditional coverage by ensuring that the modified score s' is (close to) uniformly distributed, leading to asymptotically valid conditional prediction sets, while keeping marginal validity.

An alternative strategy to achieve asymptotic conditional coverage is to define a finite partition $\mathcal{A} = \{A_1, \ldots, A_K\}$ of the feature space $\mathcal{X}$, and construct prediction regions locally within each partition element. Concretely, let $T : \mathcal{X} \to \mathcal{A}$ be the function that maps each feature vector $\mathbf{x}$ to its corresponding region in $\mathcal{A}$. Then, divide the calibration set into subsets $I_j = \{i : T(\mathbf{X}_i) = A_j\}$, $j = 1, \ldots, K$. Finally, within each region A_j, compute the conformal quantile $t_{j,1-\alpha}$ as the $(1+1/n_j)(1-\alpha)$ empirical quantile of scores s_i for $i \in I_j$, where $n_j = |I_j|$. The local prediction region is then defined as

$$C_{\text{local}}(\mathbf{x}) = \{y \in \mathcal{Y} : s(y; \mathbf{x}) \leq t_{j,1-\alpha}\}, \quad \text{for } \mathbf{x} \in A_j.$$

This procedure guarantees $\mathbb{P}\left(Y_{n+1} \in C_{\text{local}}(\mathbf{X}_{n+1}) \mid \mathbf{X}_{n+1} \in A_j\right) \geq 1-\alpha$. As the partition becomes finer (i.e., as A_j shrinks), the method approaches conditional validity. Several strategies for defining such partitions have been proposed [Vovk, 2012, Lei and Wasserman, 2014, Boström and Johansson, 2020, Boström et al., 2021, Foygel Barber et al., 2021]. We use `LoCart` [Cabezas et al., 2025a], which constructs a regression tree to partition the feature space by predicting the conformity score

$s(y; \mathbf{x})$ from $\mathbf{x}$. The resulting partition approximates the coarsest one where the conditional distribution of $s \mid \mathbf{x}$ is constant [Meinshausen and Ridgeway, 2006, Cabezas et al., 2025a], enabling local conformalization to closely match conditional coverage with a minimal number of data-efficient regions.

2.2. Bayesian Credible Sets

Bayesian credible sets typically take the form

$$C(\mathbf{x}) = \{\theta : s(\theta; \mathbf{x}) \leq t_{1-\alpha}(\mathbf{x})\}, \tag{3}$$

where $s(\theta; \mathbf{x})$ is a scoring function computed using the posterior distribution, and the threshold $t_{1-\alpha}(\mathbf{x})$ is chosen to satisfy the conditional coverage condition:

$$\mathbb{P}\left(\theta \in C(\mathbf{x}) \mid \mathbf{x}\right) = \int \mathbb{I}\left(s(\theta; \mathbf{x}) \leq t_{1-\alpha}(\mathbf{x})\right) p(\theta \mid \mathbf{x})\, d\theta = 1 - \alpha.$$

In other words, the posterior probability that the credible set contains the parameter value must be $1 - \alpha$, where the miscoverage level α is defined beforehand.

Different choices of the scoring function s lead to different types of credible sets. For example (see Figure 2 for an illustration):

- **(HPD Regions)** If $s(\theta; \mathbf{x}) = -p(\theta \mid \mathbf{x})$, the resulting set corresponds to the highest posterior density (HPD) region;
- **(Symmetric Regions)** In the case $\theta \in \mathbb{R}$ and if $s(\theta; \mathbf{x}) = \dfrac{|\theta - \mathbb{E}[\theta \mid \mathbf{x}]|}{\sqrt{\mathrm{Var}[\theta \mid \mathbf{x}]}}$, the resulting set is a central region based on the posterior mean and variance (see Masserano et al. 2023 for multivariate extensions);
- **(Quantile-based Regions)** In the case $\theta \in \mathbb{R}$ and if $s(\theta; \mathbf{x}) = \max\{q_{\alpha_1}(\mathbf{x}) - \theta,\, \theta - q_{\alpha_2}(\mathbf{x})\}$, where $q_\alpha(\mathbf{x})$ denotes the α-quantile of the distribution $\theta \mid \mathbf{x}$ and $\alpha_2 - \alpha_1 = 1 - \alpha$, the resulting set is the quantile-based interval $(q_{\alpha_1}(\mathbf{x}), q_{\alpha_2}(\mathbf{x}))$. In this case, the threshold satisfies $t_{1-\alpha}(\mathbf{x}) = 0$ by construction.

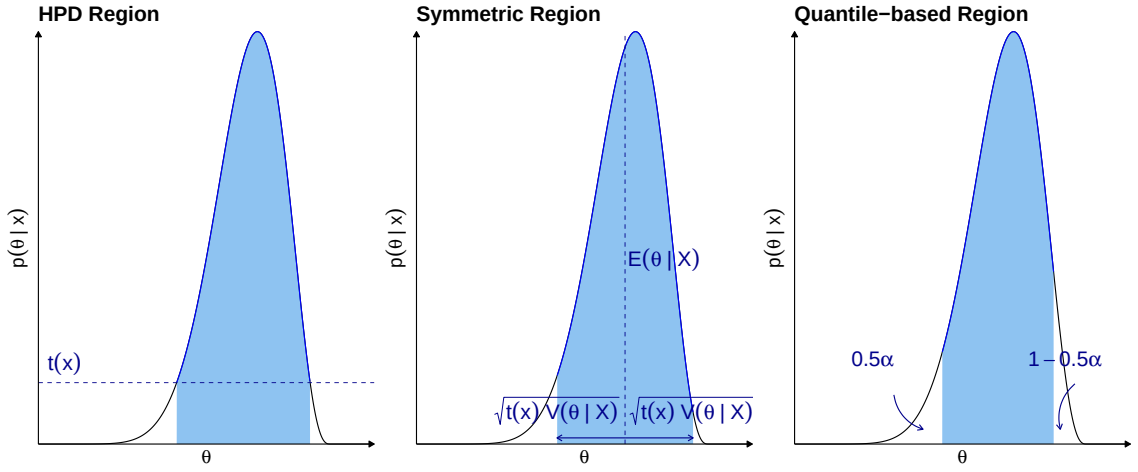

Figure 2: Credible regions for each distinct scoring function s.

When the true posterior $p(\theta \mid \mathbf{x})$ is available—either in closed form or through sampling—it is conceptually straightforward, albeit potentially computationally demanding, to compute a threshold $t_{1-\alpha}(\mathbf{x})$ that guarantees valid coverage as defined by the equation above.

Now suppose that an approximation $\widehat{p}(\theta \mid \mathbf{x})$ to the true posterior is used, for instance via neural posterior estimation, variational inference, or another method. The scoring function s can then be

derived from the approximate version of p. For instance, the score corresponding to HPD can be $-\widehat{p}(\theta \mid \mathbf{x})$. However, in this setting the threshold $t_{1-\alpha}(\mathbf{x})$ also needs to be estimated. A common approach is to choose its estimate $\widehat{t}_{1-\alpha}(\mathbf{x})$ so that

$$\widehat{\mathbb{P}}\left(\theta \in \widehat{C}(\mathbf{x}) \mid \mathbf{x}\right) := \int \mathbb{I}\left(s(\theta;\mathbf{x}) \leq \widehat{t}_{1-\alpha}(\mathbf{x})\right) \widehat{p}(\theta \mid \mathbf{x})\, d\theta = 1-\alpha. \tag{4}$$

That is, $\widehat{t}_{1-\alpha}(\mathbf{x})$ is set so that the resulting credible set contains θ with posterior probability $1-\alpha$ under the approximate $\widehat{p}$. However, if $\widehat{p}$ poorly approximates the true posterior p, the resulting coverage can deviate substantially from the nominal level. `CP4SBI` addresses this issue.

3. Our approach: `CP4SBI`

The key insight of our approach is to reinterpret $s(\theta;\mathbf{x})$ — defined in Equation (3) — as a conformity score in the conformal inference framework. This allows us to construct credible sets with improved coverage properties, even when the posterior $p(\theta \mid \mathbf{x})$ is poorly approximated.

A first naive conformal approach defines the prediction set as $C(\mathbf{x}) = \{\theta \in \Theta : s(\theta;\mathbf{x}) \leq t_{1-\alpha}\}$, where $t_{1-\alpha}$ is the $(1+1/B)(1-\alpha)$-quantile of the scores $\{s(\theta_b;\mathbf{x}_b)\}_{b=1}^{B}$. This ensures marginal coverage: $\mathbb{P}\left(\theta \in C(\mathbf{X})\right) \geq 1-\alpha$, but offers no conditional coverage guarantees, which are central in Bayesian settings. Even as $B \to \infty$, t converges to the $(1-\alpha)$-quantile of the marginal distribution of $s(\theta;\mathbf{X})$, not the conditional distribution of $s(\theta;\mathbf{x})$ for fixed $\mathbf{x}$. To address this, we introduce conformal methods designed for asymptotic conditional validity. We present two implementations of `CP4SBI`, though other strategies from Section 2.1 also apply:

`LoCart CP4SBI`: We partition the covariate space $\mathcal{X}$ using a regression tree fitted to predict $s(\theta;\mathbf{x})$ from $\mathbf{x}$, following the `LoCart` method from Cabezas et al. 2025a. Conformal calibration is then applied within the tree leaf containing $\mathbf{x}$, using only calibration scores from that leaf. If the tree partitions are sufficiently representative of the structure of the conformity score, we expect the distribution of $s(\theta;\mathbf{x})$ to be approximately homogeneous within each region. As a result, the conditional probability $\mathbb{P}(\theta \in C(\mathbf{x}) \mid \mathbf{x} \in A)$ closely approximates $\mathbb{P}(\theta \in C(\mathbf{x}) \mid \mathbf{X} = \mathbf{x})$, since the partition captures local behavior in the score distribution. In our implementation of `LoCart`, we adopt an augmented version that enriches the feature space by including an estimate of the conditional variance of the conformity score, $\mathbb{V}[s(\theta;\mathbf{X}) \mid \mathbf{X}]$, as an additional feature. This augmentation allows for more informative partitions and improves the efficiency of the local procedure. The method is summarized in Algorithm 1 in Appendix A and illustrated in Figure 3. While marginal coverage guarantees theoretically require splitting the data into training and calibration sets, we omit this step in practice, as it has minimal empirical impact on coverage and may reduce practical performance.

`CDF CP4SBI`: This variant improves conditional coverage by transforming the score $s(\theta;\mathbf{x})$ via an estimate $\widehat{F}(\cdot \mid \mathbf{x})$ of its conditional cumulative distribution function, as proposed by Dheur et al. 2025: $s'(\theta;\mathbf{x}) = \widehat{F}(s(\theta;\mathbf{x}) \mid \mathbf{x})$. In its original prediction setting, CDF-conformal needs the conditional distribution of $s(Y;\mathbf{X}) \mid \mathbf{x}$ to be approximated from scratch, which requires estimating a conditional density (via e.g. normalizing flows or a kernel density estimator) on top of the original conformal score. This results in high computational costs and a separate holdout set dedicated to learning such a distribution. Note, however, that in our simulator-based inference setting, an estimate of the posterior $\widehat{p}(\theta \mid \mathbf{x})$ is already available and we can use it to obtain $\widehat{F}(s \mid \mathbf{x})$ with minimal additional computational cost. In practice, we approximate $\widehat{F}(s(\theta;\mathbf{x}) \mid \mathbf{x})$ using a Monte Carlo estimate based on posterior draws $\{\theta_j\}_{j=1}^{M} \sim \widehat{p}(\theta \mid \mathbf{x})$, i.e., samples generated from the

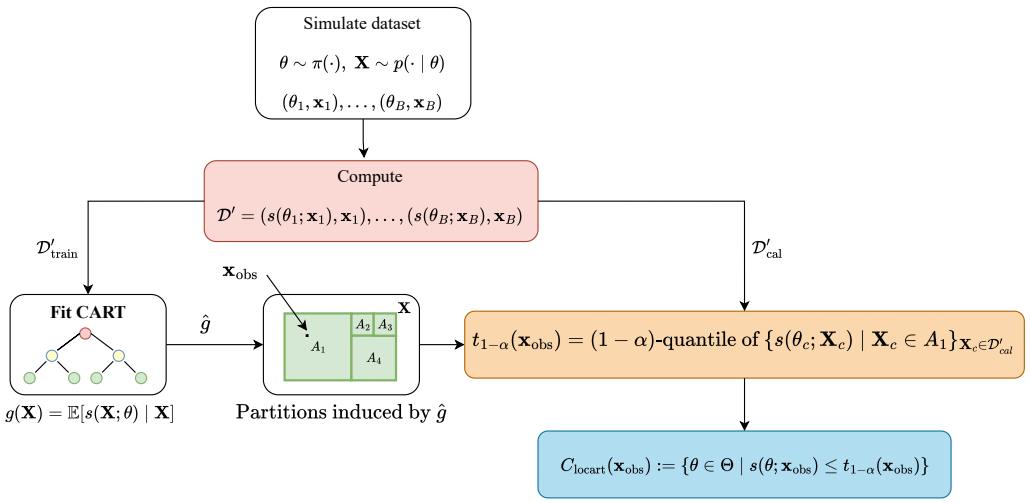

Figure 3: `LoCart CP4SBI`. Starting with a simulated calibration dataset (top-center), we compute conformity scores and build a new dataset $\mathcal{D}'$ (red). This is split into training and calibration sets. A regression tree is trained to predict s from $\mathbf{X}$ (CART panel), partitioning the feature space (partition panel). We locate the region containing $\mathbf{x}_{\text{obs}}$ and compute a local cutoff as the $(1-\alpha)$ quantile of scores from calibration points in that region (orange). This cutoff defines the final credible region (blue).

estimated posterior at $\mathbf{x}$. The approximation is given by the empirical CDF:

$$\widehat{F}_M(s(\theta; \mathbf{x}) \mid \mathbf{x}) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}\big(s(\theta_j; \mathbf{x}) \leq s(\theta; \mathbf{x})\big). \tag{5}$$

This corresponds to using the ECDF method from [Dheur et al., 2025]. `CDF CP4SBI` is summarized in Algorithm 2 in Appendix A and illustrated in Figure 4.

Nuisance parameters and transformations of the parameter space. In many applications, one may only need credible sets for a subset of θ or for a transformation $\phi = g(\theta)$. Our method naturally accommodates these cases. Specifically, we approximate the posterior $p(\phi \mid \mathbf{x})$ using the same techniques as for $p(\theta \mid \mathbf{x})$, then compute scores for ϕ as described in Section 2.2. To calibrate the cutoff $t_{1-\alpha}(\mathbf{x})$, we use the transformed calibration set $(\phi_1, \mathbf{X}_1), \ldots, (\phi_B, \mathbf{X}_B)$ with $\phi_i = g(\theta_i)$. Figure 5 shows our method in this setting, reducing a 10-dimensional problem (Gaussian Linear Uniform [Lueckmann et al., 2021]) to the first two dimensions (details on the illustration in Appendix C.2). Our approach yields regions close to the oracle with accurate coverage, while other methods tend to be overconfident.

Credible sets for implicit generative models. Many recent posterior estimation methods do not yield a closed-form expression for $\widehat{p}(\theta \mid \mathbf{x})$, or evaluating it is too costly. Instead, they provide independent samples $\theta_1, \ldots, \theta_L$ from $\widehat{p}(\theta \mid \mathbf{x})$ for each fixed $\mathbf{x}$—e.g., score-diffusion [Linhart et al., 2024] and flow-matching models [Wildberger et al., 2023]. `CP4SBI` remains applicable in this setting. Given a scoring function $s(\theta; \mathbf{x})$, both `LoCart CP4SBI` and `CDF CP4SBI` can be used directly, as they only require a posterior sampler (see Algorithms 1 and 2 in Appendix A). However, the standard scores from Section 2.2 assume access to an explicit density and thus cannot be used as-is. The approximations below allow their use with posterior samples:

- The HPD score can be approximated as $s(\theta; \mathbf{x}) \propto -\sum_{l=1}^{L} K(\theta, \theta_l)$, where K is a smoothing kernel, which corresponds to applying a kernel density estimator to the posterior samples.
- The symmetric-region score can be approximated by $s(\theta; \mathbf{x}) = \widehat{\mathbb{V}}^{-1/2}[\theta \mid \mathbf{x}] \cdot |\theta - \widehat{\mathbb{E}}[\theta \mid \mathbf{x}]|$, where $\widehat{\mathbb{E}}$ and $\widehat{\mathbb{V}}$ denote the empirical mean and variance of the samples $\{\theta_1, \ldots, \theta_L\}$.

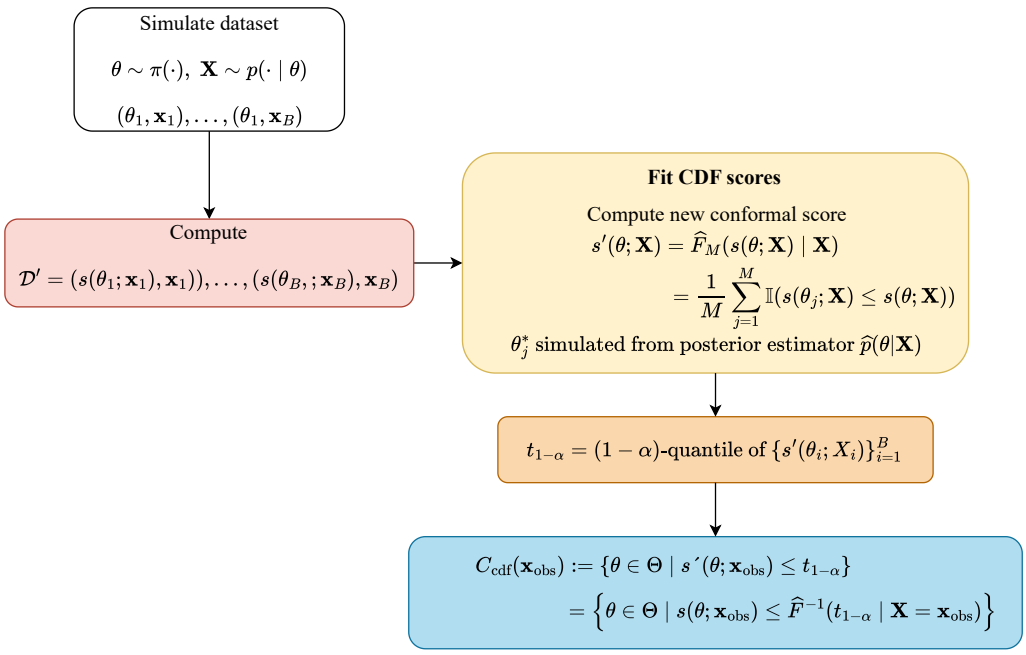

Figure 4: `CDF CP4SBI`. Starting with a simulated calibration dataset (top-left), we compute nonconformity scores and build the dataset $\mathcal{D}'$ (red). Using a posterior estimator, we estimate the conditional CDF of the scores and transform each into its CDF value $s'(\mathbf{X}; \theta)$ (yellow). A global cutoff is then computed from the transformed scores (orange). This threshold defines the credible region for $\mathbf{x}_{\text{obs}}$, either via the transformed scores or through a data-dependent cutoff on the original scores.

- The quantile-based score can be approximated via $s(\theta; \mathbf{x}) = \max\left\{\widehat{q}_{\alpha_1}(\mathbf{x}) - \theta,\ \theta - \widehat{q}_{\alpha_2}(\mathbf{x})\right\}$, where $\widehat{q}_{\alpha_1}(\mathbf{x})$ and $\widehat{q}_{\alpha_2}(\mathbf{x})$ are the empirical quantiles of the posterior samples.

Other scores from the multivariate conformal prediction literature can also be used in this framework, such as C-PCP [Wang et al., 2023, Dheur et al., 2025] and CP^2-PCP [Plassier et al., 2024]. In this work, we employ the Kernel Approximation to derive the HPD regions for these generative models. Algorithm 3 in Appendix A details the procedure for deriving these scores, and Figure 6 illustrates the regions obtained by using our method in this setting on the Gaussian Mixture task [Lueckmann et al., 2021] (details on the illustration can be found on Appendix C.2). `CP4SBI` better approximates the oracle region and achieves coverage closer to the nominal rate.

4. Theoretical guarantees

We now present theoretical guarantees for the methods in Section 3. From this point on, we assume a disjoint data split: a training set $\{(\theta_i, \mathbf{X}_i)\}_{i=1}^{K}$ to estimate the posterior $\widehat{p}(\theta \mid \mathbf{x})$, and a calibration set $\{(\theta_i, \mathbf{X}_i)\}_{i=1}^{B}$ to calibrate the conformal methods. Proofs are given in Appendix B.

4.1. `LoCart CP4SBI`

The coverage guarantee of `LoCart CP4SBI` follows from applying standard conformal prediction separately within each region A_j of the partition it defines. Once the partition is fixed, the calibration scores $s(\theta; \mathbf{x})$ remain exchangeable within each region, which allows us to apply conformal calibration independently in each subset. As a result, the quantile $t_{1-\alpha}(\mathbf{x})$, computed using calibration points in A_j, yields valid marginal coverage conditional on $\mathbf{x} \in A_j$.

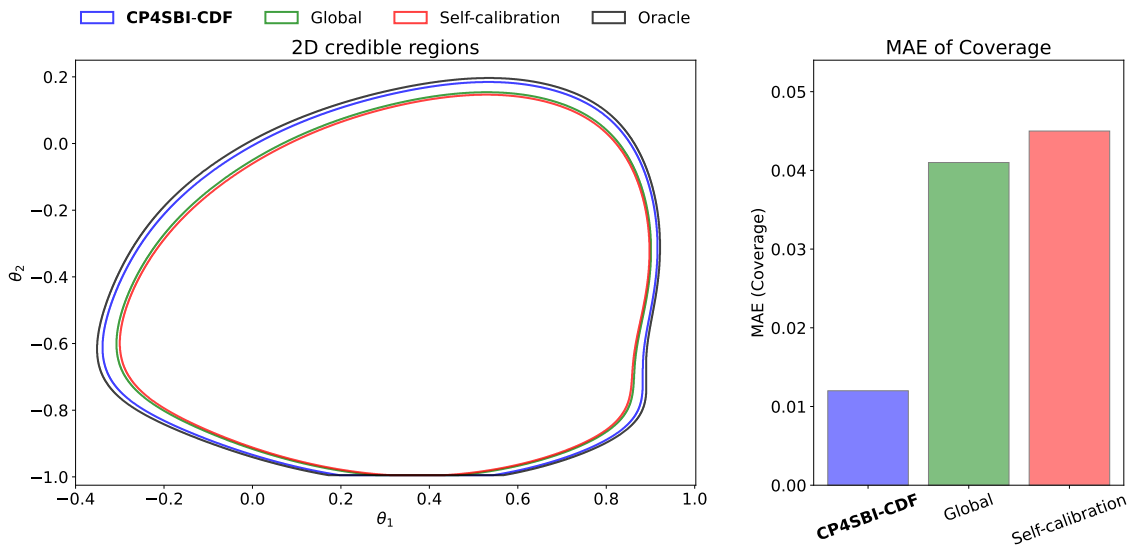

Figure 5: For a fixed observation $\mathbf{x}_{\text{obs}}$, we compare credible regions for the first two parameters from the 10-dimensional Gaussian Linear Uniform benchmark task. Our `CDF CP4SBI` (blue) creates a region that is closer to the oracle (black) with smaller deviation from nominal coverage.

Theorem 4.1 (`LoCart CP4SBI` local coverage)**.** *Suppose the calibration pairs* $\{(\theta_i, \mathbf{X}_i)\}_{i=1}^{B}$ *and the test pair* $(\theta, \mathbf{X})$ *are exchangeable. Let* $\{A_j\}_{j=1}^{J}$ *be the partition of* $\mathcal{X}$ *produced by the* `LoCart CP4SBI`*. Denote by* A_j *the cell of the partition containing* $\mathbf{X}$*, and let* $t_{1-\alpha}(\mathbf{X})$ *be the* $(1-\alpha)$ *quantile of the conformity scores* $s(\theta; \mathbf{x})$ *computed from the calibration pairs with* $\mathbf{X}_i \in A_j$*. Define the conformal set*

$$C_{locart}(\mathbf{x}) := \big\{\theta \in \Theta : s(\theta; \mathbf{x}) \le t_{1-\alpha}(\mathbf{x})\big\}.$$

Then $C_{locart}(\mathbf{x})$ *satisfies both local and marginal coverage:*

$$\mathbb{P}\big(\theta \in C_{locart}(\mathbf{X}) \,\big|\, \mathbf{X} \in A_j\big) \;\ge\; 1-\alpha, \qquad \mathbb{P}\big(\theta \in C_{locart}(\mathbf{X})\big) \;\ge\; 1-\alpha.$$

In addition, under the regularity conditions of Cabezas et al. [2025a, Theorem 5], `LoCart CP4SBI` achieves asymptotic conditional coverage.

Theorem 4.2 (`LoCart CP4SBI` asymptotic conditional coverage)**.** *Let* B *denote the size of the calibration set. Under the assumptions stated in Cabezas et al. [2025a, Theorem 5],* `LoCart CP4SBI` *satisfies*

$$\lim_{B\to\infty} \mathbb{P}\big(\theta \in C_{locart}(\mathbf{X}) \,\big|\, \mathbf{X} = \mathbf{x}\big) = 1-\alpha,$$

that is, it achieves conditional coverage in the limit as the calibration sample size grows.

4.2. `CDF CP4SBI`

When using the transformed scores $s'(\theta; \mathbf{x}) = \widehat{F}_M(s(\theta; \mathbf{x}) \mid \mathbf{x})$, the `CDF CP4SBI` procedure reduces to applying standard conformal prediction with a modified nonconformity score. Here, $\widehat{F}_M(\cdot \mid \mathbf{x})$ denotes the Monte Carlo estimate of the conditional CDF of the score $s(\theta; \mathbf{x})$ given $\mathbf{x}$ based on M samples drawn from the estimated posterior, defined by Equation 5. In the marginal setting, where calibration and test pairs $(\theta, \mathbf{X})$ are exchangeable, this transformation preserves validity, and the method retains the usual marginal coverage guarantee.

Theorem 4.3 (Marginal coverage)**.** *Assume the calibration pairs* $\{(\theta_i, \mathbf{X}_i)\}_{i=1}^{B}$ *and the test pair* $(\theta, \mathbf{X})$ *are exchangeable, and let* $s'_i = \widehat{F}_M(s(\theta_i; \mathbf{X}_i) \mid \mathbf{X}_i)$ *be the transformed conformity scores.*

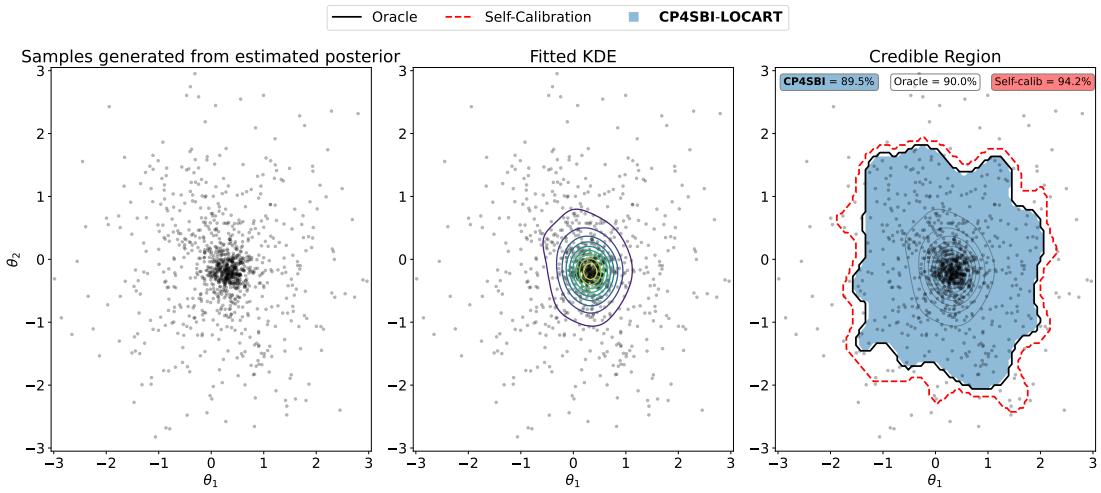

Figure 6: Approximation of the continuous-flow generative model's HPD region for the Gaussian Mixture benchmark using Kernel Approximation integrated with `CP4SBI`. Our method achieves closer alignment to the oracle region and better coverage than self-normalization.

Then the conformal region

$$C_{cdf}(\mathbf{x}) := \{\theta \in \Theta : \widehat{F}_M(s(\theta;\mathbf{x}) \mid \mathbf{x}) \leq t_{1-\alpha}\},$$

with $t_{1-\alpha}$ *the empirical* $(1-\alpha)$*-quantile of* $\{s'_i\}_{i=1}^B$*, satisfies*

$$\mathbb{P}(\theta \in C_{cdf}(\mathbf{X})) \geq 1-\alpha.$$

The conditional validity of `CDF CP4SBI` builds on the probability integral transform, following the approach of Dheur et al. [2025]. If the estimated posterior $\widehat{p}(\theta \mid \mathbf{x})$ converges to the true posterior $p(\theta \mid \mathbf{x})$ as the training size $K \to \infty$, and the number of posterior draws M used to compute $\widehat{F}_M$ grows to infinity, then the transformed score $s'(\theta;\mathbf{x}) = \widehat{F}_M(s(\theta;\mathbf{x}) \mid \mathbf{x})$ converges to the true conditional CDF $F(s(\theta;\mathbf{x}) \mid \mathbf{x})$ of the score. Since $\theta \sim p(\cdot \mid \mathbf{x})$, this implies that $s'(\theta;\mathbf{x})$ becomes approximately Uniform$(0,1)$ conditional on $\mathbf{x}$. In this case, the calibration scores and the test score are approximately conditionally i.i.d., and the empirical quantile converges to the target level $1-\alpha$, yielding asymptotic conditional coverage.

The asymptotic conditional validity of the procedure depends on the accuracy of the posterior approximation and of the CDF estimate: M must grow to ensure that $\widehat{F}_M$ converges to $\widehat{F}$ implied by $\widehat{p}$, B must grow so that the empirical quantile of the transformed scores converges to its population value, and K must grow to guarantee that $\widehat{p}$ converges to p.

Theorem 4.4 (`CDF CP4SBI` asymptotic conditional coverage)**.** *Let* B *be the calibration set size,* K *the training set size used to estimate the posterior distribution* $\widehat{p}(\theta \mid \mathbf{x})$*, and* M *the number of posterior draws used to compute* $\widehat{F}_M$*. Under Assumption B.1, if* $B \to \infty$*,* $K \to \infty$*, and* $M \to \infty$*, then*

$$\lim_{B,K,M \to \infty} \mathbb{P}\big(\theta \in C_{cdf}(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}\big) = 1-\alpha.$$

5. Experiments

We compare our approach to baselines in terms of conditional and marginal statistical validity using ten SBI benchmarks introduced by Lueckmann et al. 2021 (details in Table 1, in Appendix C.1). All experiments have a nominal level at $1-\alpha = 0.9$ with an overall simulation budget of $B_{\text{all}} = 10000$.

From this budget, 80% is allocated for training the posterior estimators, and the remaining 20% is reserved for the calibration set. We consider other simulation budgets in Appendix C.4. We focus on constructing HPD regions, defining the conformity score as $s(\theta; \mathbf{x}) = -\widehat{p}(\theta \mid \mathbf{x})$ or, for sample-based models, as $s(\theta; \mathbf{x}) \propto -\sum_{l=1}^{L} K(\theta, \theta_l)$, using Scott's rule to determine kernel bandwidth [Scott, 2015], where $\theta_l \sim \widehat{p}(\cdot \mid \mathbf{x})$. Posterior estimation is carried out using conditional normalizing flows (NPE; [Greenberg et al., 2019, Papamakarios et al., 2021]) and a conditional diffusion model (NPSE; [Geffner et al., 2023]). For both estimators, we use the implementations provided in the `sbi` package [Boelts et al., 2025], using their standard architectures.

5.1. Metrics for conditional and marginal validity

To assess conditional validity, we estimate the conditional coverage for each $\mathbf{x}$ as

$$\delta(\mathbf{x}, C) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}\left(\theta_k \in C\left(\mathbf{x}\right)\right) \ ,$$

where $\theta_k \sim p(\theta \mid \mathbf{x})$. We then compute the Mean Absolute Error (MAE) over B_{sim} fixed observations:

$$\text{MAE} = \frac{1}{B_{\text{sim}}} \sum_{i=1}^{B_{\text{sim}}} \left|\delta(\mathbf{x}_i, C) - (1-\alpha)\right| \ . \tag{6}$$

Lower MAE indicates conditional coverage closer to the nominal level, reflecting better calibration.

We assess marginal coverage by simulating an independent test set $\mathcal{D}_{\text{test}} = \{(\theta_j, \mathbf{x}_j)\}_{j=1}^{B_{\text{test}}}$ and compute the Average Marginal Coverage (AMC) as:

$$\text{AMC} = \frac{1}{B_{\text{test}}} \sum_{i=1}^{B_{\text{test}}} \mathbb{I}\left(\theta_i \in C(\mathbf{x}_i)\right) \ .$$

Values near the nominal level $1-\alpha$ indicate good average coverage. We fix $K = 1000$ and $B_{\text{test}} = 2000$. For MAE, we use $B_{\text{sim}} = 500$ or 10, depending on the simulator's posterior sample availability in `sbibm` [Lueckmann et al., 2021] (see Appendix C.1 for details).

5.2. Baselines

We compare our approach to three established methods for constructing credible regions:

- **Self-calibration**: This method constructs credible regions directly from the estimated posterior distribution, $\widehat{p}(\theta \mid \mathbf{x})$. For each fixed data point $\mathbf{x}$, the method draws a set of B_{self} (fixed at 1000) posterior samples, $\theta_i \sim \widehat{p}(\cdot \mid \mathbf{x})$, and evaluates their corresponding scores, $s(\theta_i; \mathbf{x})$. From these scores, an empirical threshold $\widehat{t}(\mathbf{x})$ is computed via Monte Carlo integration, following Equation (4), to ensure the desired coverage level of $1-\alpha$.
- **Global** [Patel et al., 2023]: This method is a direct application of the vanilla conformal approach for constructing credible regions. First, it uses the calibration set $\mathcal{D}_{\text{cal}}$ to compute nonconformity scores for each data point. Then, it determines a single threshold, t_α, from these scores, which is then applied uniformly to all new observations, $\mathbf{x}$.
- **HDR** [Chung et al., 2024]: This multivariate recalibration method corrects miscalibration by learning a monotonic mapping R from posterior density values using a calibration set $\mathcal{D}_{\text{cal}}$. For a new $\mathbf{x}$, it resamples from the posterior to align with R, producing calibrated samples that account for dependencies across dimensions. Cutoffs for HDRs are then computed using these recalibrated samples, similarly to self-calibration.

5.3. Results and discussion

Next, we evaluate the credible regions produced by different methods on the selected benchmarks. Focusing on the NPE base estimator, Figure 7 shows that our approach significantly outperforms existing methods in terms of conditional coverage. Specifically, both `LoCart CP4SBI` and `CDF CP4SBI` methods perform statistically better in 8 out of 10 benchmarks, all while maintaining marginal coverage close to the nominal level. Overall, our methods perform well in almost all benchmarks, except for the Lotka-Volterra simulator, where only the global method shows a better performance. The right panel further shows that our proposed methods consistently maintain near-nominal marginal coverage, whereas approaches such as self-calibration and HDR lack this property on some benchmarks. The ability of our methods to produce efficient and well-calibrated regions across various benchmarks, parameter spaces, and data distributions highlights their robust performance.

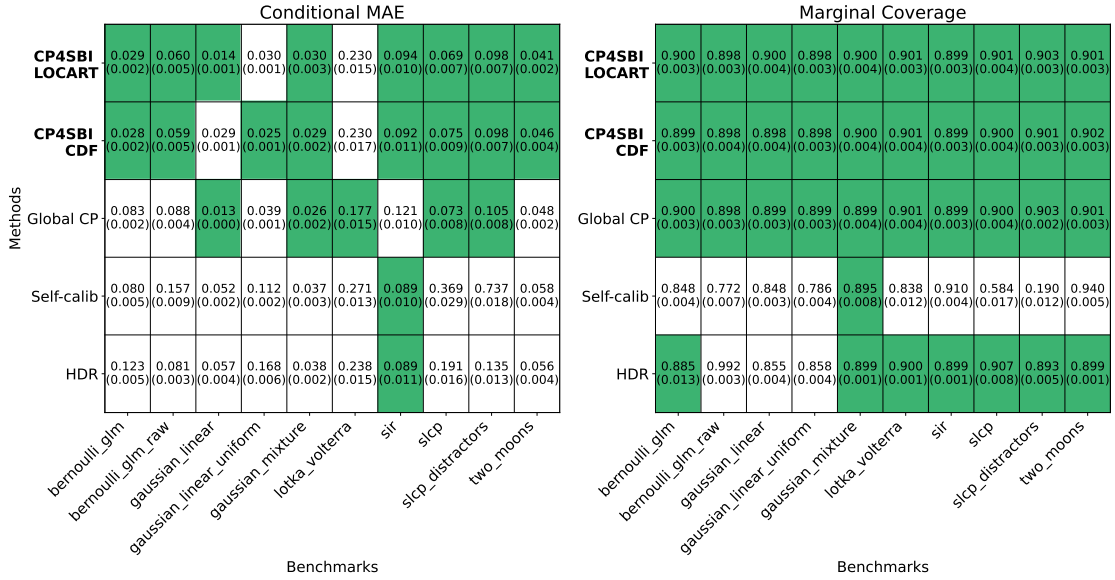

Figure 7: Conditional MAE (left) and marginal coverage (right) for NPE-based credible sets are shown across benchmark tasks, with means and 95% confidence intervals over 50 runs. Green cells indicate statistically significant improvements (lower MAE or coverage near the nominal level). Our methods, `LoCart CP4SBI` and `CDF CP4SBI`, outperform others in 8 of 10 tasks (left) and maintain close-to-nominal marginal coverage (right), demonstrating improved posterior calibration across varied tasks.

Figure 8 shows that our strategy for constructing calibrated HPD regions from an NPSE base estimator (i.e. a score diffusion model) is also successful. Both of our methods demonstrate better conditional coverage performance than other approaches, as well as solid marginal coverage. In particular, `LoCart CP4SBI` proved to be the most effective, producing credible regions that simultaneously improve conditional coverage while respecting marginal coverage. While the HDR method also shows acceptable conditional coverage in some settings, it does not achieve marginal coverage in any dataset. This highlights the efficiency of our approach in recalibrating credible regions across different posterior estimators. Figure 9 in Appendix C.4 shows that our approach consistently displays improved conditional coverage performance across two budgets ($B_{\text{all}} = 2000$ and $B_{\text{all}} = 20000$). A caveat exists, however: for the smaller budget, `LoCart CP4SBI` performs comparably to the global one, likely because its partitioning strategy struggles with sparse calibration data. For this case, `CDF CP4SBI` shows better results. This illustrates how `CP4SBI` adapts its performance based on the available data, highlighting its flexibility across different calibration

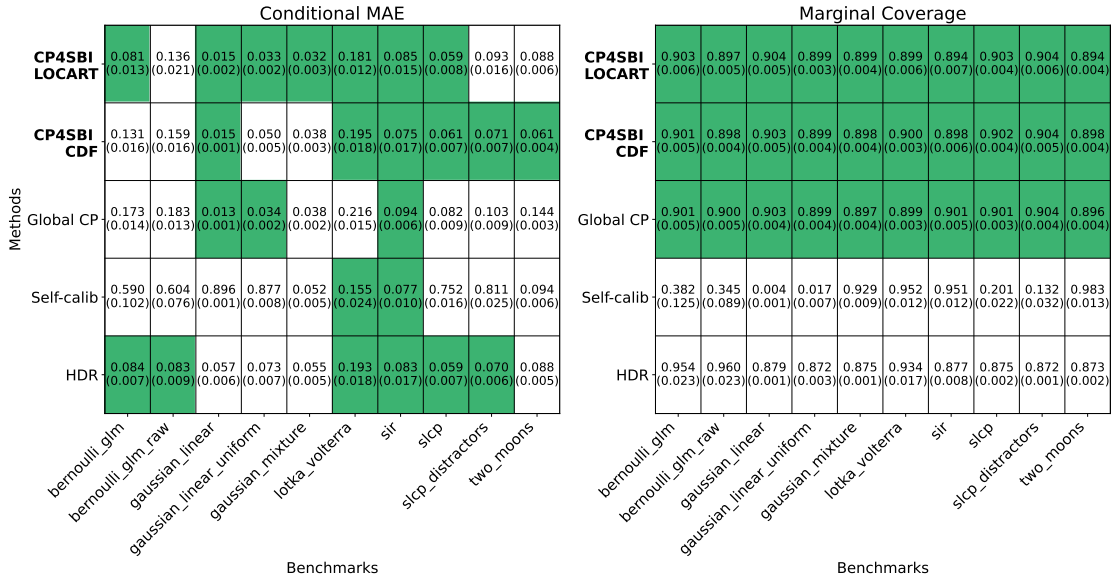

Figure 8: Conditional MAE (left) and marginal coverage (right) for NPSE-based credible sets across benchmark tasks, averaged over 15 runs with 95% confidence intervals. Green cells highlight statistically significant coverage near the nominal level. `LoCart CP4SBI` ranks among the best in 7 of 10 tasks. While HDR shows good conditional coverage, it fails to maintain marginal coverage. Our methods perform well on both metrics, ensuring calibrated inference.

budgets and posterior estimators of varying quality. The results obtained in this section reinforce the capacity of `CP4SBI` to enhance calibration in credible regions for fixed observations $\mathbf{x}$.

6. Final Remarks

In this paper, we tackled the fundamental problem of miscalibration in Simulation-Based Inference. We proposed `CP4SBI`, a post-hoc, model-agnostic framework using conformal prediction to build credible sets with finite-sample guarantees. It applies to any SBI method offering posterior samples or density estimates and supports any scoring function, enabling calibrated HPD regions, symmetric intervals, or custom sets.

We proposed two calibration strategies. The first, `CDF CP4SBI`, achieves asymptotic conditional coverage by recalibrating scores using an estimate of their conditional CDF. Its guarantees strengthen as the underlying posterior approximation improves—typically as the training sample size increases. The second, `LoCart CP4SBI`, provides finite-sample local coverage by partitioning the data space with a regression tree, adapting the calibration to regions of varying difficulty. This method also attains asymptotic conditional coverage as the size of the calibration set grows. Both methods guarantee correct marginal coverage. Our experiments on established SBI benchmarks confirmed that `CP4SBI` effectively improves uncertainty quantification for both neural posterior estimators and diffusion-based models.

Despite offering stronger coverage guarantees, `CP4SBI` has some limitations. The tightness of its credible regions depends on the quality of the initial posterior; poor approximations lead to wider, albeit calibrated, sets. Additionally, `LoCart CP4SBI` requires a sufficiently large calibration set to populate its partitions, which can be challenging in high-dimensional spaces. This can be partly mitigated by including posterior variance estimates or other summary statistics in the partitioning features.

This work opens several avenues for future research. Although we focused on two conformal methods, other techniques may yield better performance in specific settings (see, e.g., Plassier et al.

2025). We also plan to extend the `CP4SBI` framework to more challenging inference scenarios, such as those involving hierarchical models or discrete parameters. Finally, its applicability is not limited to SBI; it could be a valuable tool for calibrating posteriors from other methods like Variational Inference or Approximate Bayesian Computation.

In conclusion, `CP4SBI` offers a flexible and theoretically grounded approach to improving uncertainty quantification in computational science. By producing credible sets with stronger local and conditional coverage, it enhances the reliability of inferences from complex simulator models. This marks a meaningful advance for many scientific fields, providing a general-purpose tool to support trustworthy discovery.

Code to implement `CP4SBI` and reproduce the experiments and illustrations is available at https://github.com/Monoxido45/CP4SBI.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. L.M.C.C is grateful for the fellowship provided by São Paulo Research Foundation (FAPESP), grant 2022/08579-7. RI is grateful for the financial support of FAPESP (grant 2023/07068-1) and CNPq (grants 422705/2021-7 and 305065/2023-8)

References

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Meili Baragatti, Bertrand Cloez, David Métivier, and Isabelle Sanchez. Approximate bayesian computation with deep learning and conformal prediction. *arXiv preprint arXiv:2406.04874*, 2024.

Jan Boelts, Michael Deistler, Manuel Gloeckler, Álvaro Tejero-Cantero, Jan-Matthis Lueckmann, Guy Moss, Peter Steinbach, Thomas Moreau, Fabio Muratore, Julia Linhart, Conor Durkan, Julius Vetter, Benjamin Kurt Miller, Maternus Herold, Abolfazl Ziaeemehr, Matthijs Pals, Theo Gruner, Sebastian Bischoff, Nastya Krouglova, Richard Gao, Janne K. Lappalainen, Bálint Mucsányi, Felix Pei, Auguste Schulz, Zinovia Stefanidi, Pedro Rodrigues, Cornelius Schröder, Faried Abu Zaid, Jonas Beck, Jaivardhan Kapoor, David S. Greenberg, Pedro J. Gonçalves, and Jakob H. Macke. sbi reloaded: a toolkit for simulation-based inference workflows. *Journal of Open Source Software*, 10(108):7754, 2025. doi: 10.21105/joss.07754. URL https://doi.org/10.21105/joss.07754.

Joshua J Bon, David J Warne, David J Nott, and Christopher Drovandi. Bayesian score calibration for approximate models. *arXiv preprint arXiv:2211.05357*, 2022.

Henrik Boström and Ulf Johansson. Mondrian conformal regressors. In *Conformal and Probabilistic Prediction and Applications*, pages 114–133. PMLR, 2020.

Henrik Boström, Ulf Johansson, and Tuwe Löfström. Mondrian conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pages 24–38. PMLR, 2021.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255. URL https://books.google.com.br/books?id=koNqWRluhP0C.

Luben Cabezas, Guilherme P Soares, Thiago R Ramos, Rafael B Stern, and Rafael Izbicki. Distribution-free calibration of statistical confidence sets. *arXiv preprint arXiv:2411.19368*, 2024.

Luben MC Cabezas, Mateus P Otto, Rafael Izbicki, and Rafael B Stern. Regression trees for fast and adaptive prediction intervals. *Information Sciences*, 686:121369, 2025a.

Luben Miguel Cruz Cabezas, Vagner Silva Santos, Thiago Ramos, and Rafael Izbicki. Epistemic uncertainty in conformal scores: A unified approach. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025b.

James Carzon, Luca Masserano, Joshua D Ingram, Alex Shen, Antonio Carlos Herling Ribeiro Junior, Tommaso Dorigo, Michele Doro, Joshua S Speagle, Rafael Izbicki, and Ann B Lee. Trustworthy scientific inference for inverse problems with generative models. *arXiv preprint arXiv:2508.02602*, 2025.

George Casella and Roger Berger. *Statistical inference*. CRC press, 2024.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48), 2021.

Davide Chicco. Ten quick tips for machine learning in computational biology. *BioData mining*, 10 (1):35, 2017.

Youngseog Chung, Ian Char, and Jeff Schneider. Sampling-based multi-dimensional recalibration. In *Forty-first International Conference on Machine Learning*, 2024.

Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Niccolo Dalmasso, Rafael Izbicki, and Ann Lee. Confidence sets and hypothesis testing in a likelihood-free inference setting. In *International Conference on Machine Learning*, pages 2323–2334. PMLR, 2020.

Niccolò Dalmasso, Luca Masserano, David Zhao, Rafael Izbicki, and Ann B Lee. Likelihood-free frequentist inference: bridging classical statistics and machine learning for reliable simulator-based inference. *Electronic Journal of Statistics*, 18(2):5045–5090, 2024.

Michael Deistler, Pedro J Goncalves, and Jakob H Macke. Truncated proposals for scalable and hassle-free simulation-based inference. *Advances in neural information processing systems*, 35: 23135–23149, 2022.

Arnaud Delaunoy, Joeri Hermans, François Rozet, Antoine Wehenkel, and Gilles Louppe. Towards reliable simulation-based inference with balanced neural ratio estimation. *Advances in Neural Information Processing Systems*, 35:20025–20037, 2022.

Arnaud Delaunoy, Benjamin Kurt Miller, Patrick Forré, Christoph Weniger, and Gilles Louppe. Balancing simulation-based inference for conservative posteriors. *arXiv preprint arXiv:2304.10978*, 2023.

Biprateep Dey, David Zhao, Jeffrey A Newman, Brett H Andrews, Rafael Izbicki, and Ann B Lee. Conditionally calibrated predictive distributions by probability-probability map: application to galaxy redshift estimation and probabilistic forecasting. *arXiv preprint arXiv:2205.14568*, 2022.

Victor Dheur, Matteo Fontana, Yorick Estievenart, Naomi Desobry, and Souhaib Ben Taieb. Multi-output conformal regression: A unified comparative study with new conformity scores. *arXiv preprint arXiv:2501.10533*, 2025.

Conor Durkan, Iain Murray, and George Papamakarios. On contrastive learning for likelihood-free inference. In *International conference on machine learning*, pages 2771–2781. PMLR, 2020.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

David T Frazier, David J Nott, Christopher Drovandi, and Robert Kohn. Bayesian inference using synthetic likelihood: asymptotics and adjustments. *Journal of the American Statistical Association*, 118(544):2821–2832, 2023.

Peter E Freeman, Rafael Izbicki, and Ann B Lee. A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting. *Monthly Notices of the Royal Astronomical Society*, 468(4):4556–4565, 2017.

Tomas Geffner, George Papamakarios, and Andriy Mnih. Compositional score modeling for simulation-based inference. In *International Conference on Machine Learning*, pages 11098–11116. PMLR, 2023.

David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International conference on machine learning*, pages 2404–2414. PMLR, 2019.

Michael U Gutmann and Jukka Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125):1–47, 2016.

Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International conference on machine learning*, pages 4239–4248. PMLR, 2020.

Joeri Hermans, Nilanjan Banik, Christoph Weniger, Gianfranco Bertone, and Gilles Louppe. Towards constraining warm dark matter with stellar streams through neural simulation-based inference. *Monthly Notices of the Royal Astronomical Society*, 507(2):1999–2011, 2021.

Rafael Izbicki. *Machine Learning Beyond Point Predictions: Uncertainty Quantification*. 1st edition, 2025. ISBN 978-65-01-20272-3.

Rafael Izbicki, Ann Lee, and Chad Schafer. High-dimensional density ratio estimation with extensions to approximate likelihood computation. In *Artificial intelligence and statistics*, pages 420–429. PMLR, 2014.

Rafael Izbicki, Ann B Lee, and Taylor Pospisil. Abc–cde: Toward approximate bayesian computation with complex high-dimensional data and limited simulations. *Journal of Computational and Graphical Statistics*, 28(3):481–492, 2019.

Rafael Izbicki, Gilson T Shimizu, and Rafael B Stern. Distribution-free conditional predictive bands using density estimators. *Proceedings of Machine Learning Research (AISTATS Track)*, 2020.

Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. Cd-split and hpd-split: efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 2022.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014.

Julia Linhart, Alexandre Gramfort, and Pedro L. C. Rodrigues. L-c2ST: Local diagnostics for posterior approximations in simulation-based inference. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=k2UVKezeWn.

Julia Linhart, Gabriel Victorino Cardoso, Alexandre Gramfort, Sylvain Le Corff, and Pedro LC Rodrigues. Diffusion posterior sampling for simulation-based inference in tall data settings. *arXiv preprint arXiv:2404.07593*, 2024.

Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30, 2017.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International conference on artificial intelligence and statistics*, pages 343–351. PMLR, 2021.

Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate bayesian computational methods. *Statistics and computing*, 22(6):1167–1180, 2012.

Luca Masserano, Tommaso Dorigo, Rafael Izbicki, Mikael Kuusela, and Ann Lee. Simulator-based inference with waldo: Confidence regions by leveraging prediction algorithms and posterior estimators for inverse problems. In *International Conference on Artificial Intelligence and Statistics*, pages 2960–2974. PMLR, 2023.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.

Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

Harris Papadopoulos, Alex Gammerman, and Volodya Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pages 64–69, 2008.

George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29, 2016.

George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd international conference on artificial intelligence and statistics*, pages 837–848. PMLR, 2019.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Yash Patel, Declan McNamara, Jackson Loper, Jeffrey Regier, and Ambuj Tewari. Variational inference with coverage guarantees in simulation-based inference. *arXiv preprint arXiv:2305.14275*, 2023.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Vincent Plassier, Alexander Fishkov, Maxim Panov, and Eric Moulines. Conditionally valid probabilistic conformal prediction. *stat*, 1050:1, 2024.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. *arXiv preprint arXiv:2502.16336*, 2025.

Pedro L. C. Rodrigues, Thomas Moreau, Gilles Louppe, and Alexandre Gramfort. HNPE: Leveraging global parameters for neural posterior estimation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=E8BxwYR8op.

José-Víctor Rodríguez, Ignacio Rodríguez-Rodríguez, and Wai Lok Woo. On the application of machine learning in astronomy and astrophysics: A text-mining-based scientometric analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(5):e1476, 2022.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression, 2019.

David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian computation*. CRC press, 2018.

Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

Ali Usman, Muhammad Rafiq, Muhammad Saeed, Ali Nauman, Andreas Almqvist, and Marcus Liwicki. Machine learning computational fluid dynamics. In *2021 Swedish artificial intelligence society workshop (SAIS)*, pages 1–4. IEEE, 2021.

Denis Valle, Rafael Izbicki, and Rodrigo Vieira Leite. Quantifying uncertainty in land-use land-cover classification using conformal statistics. *Remote Sensing of Environment*, 295:113682, 2023.

Ricardo Vinuesa and Steven L Brunton. Enhancing computational fluid dynamics with machine learning. *Nature Computational Science*, 2(6):358–366, 2022.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490, 2012.

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

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David Blei. Probabilistic conformal prediction using conditional random samples. In *International Conference on Artificial Intelligence and Statistics*, pages 8814–8836. PMLR, 2023.

Antoine Wehenkel, Juan L. Gamella†, Ozan Sener, Jens Behrmann, Guillermo Sapiro, Jörn-Henrik Jacobsen, and Marco Cuturi. Addressing misspecification in simulation-based inference through data-driven calibration. In *ICML*, 2025. URL https://arxiv.org/abs/2405.08719.

Jonas Wildberger, Maximilian Dax, Simon Buchholz, Stephen Green, Jakob H Macke, and Bernhard Schölkopf. Flow matching for scalable simulation-based inference. *Advances in Neural Information Processing Systems*, 36:16837–16864, 2023.

David Zhao, Niccolò Dalmasso, Rafael Izbicki, and Ann B Lee. Diagnostics for conditional density models and bayesian inference algorithms. In *Uncertainty in Artificial Intelligence*, pages 1830–1840. PMLR, 2021.

Appendix

A. Algorithms

This appendix details all algorithms related to our approach. Algorithms 1 and 2 describe the procedures for `LoCart CP4SBI` and `CDF CP4SBI`, respectively. Algorithm 3 further details how we approximate Highest Posterior Density (HPD) scores for implicit generative methods by using Kernel Density Estimation.

Algorithm 1: CP4SBI - LoCart: Local Conformal Prediction for SBI

Input: Calibration set $\mathcal{D} = \{(\theta_i, \mathbf{x}_i)\}_{i=1}^{B}$, posterior approximator $\widehat{p}(\theta \mid \mathbf{x})$, conformity score $s(\theta; \mathbf{X})$, nominal level α, new observation $\mathbf{x}_{\text{obs}}$

Step I: Score computation

1: For each $(\theta_i, \mathbf{x}_i) \in \mathcal{D}$, compute $s_i = s(\theta_i; \mathbf{x}_i)$ using $\widehat{p}$.

2: Form the scored dataset $\mathcal{D}' = \{(s_i, \mathbf{x}_i)\}_{i=1}^{B}$, then randomly split it into training $\mathcal{D}'_{\text{train}}$ and calibration $\mathcal{D}'_{\text{calib}}$ subsets.

Step II: Partition learning

1: Fit a regression tree on $\mathcal{D}'_{\text{train}}$ to predict s from $\mathbf{x}$.

2: Use the tree to induce a partition $\mathcal{A} = \{A_1, \ldots, A_K\}$ of the feature space.

3: Define a region mapping $T : \mathcal{X} \to \mathcal{A}$ such that $T(\mathbf{x}) = A_j$ if $\mathbf{x} \in A_j$.

Step III: Local quantile estimation

1: For each region $A_j \in \mathcal{A}$, define the set of calibration indices $I_j = \{i \mid \mathbf{x}_i \in A_j,\ (s_i, \mathbf{x}_i) \in \mathcal{D}'_{\text{calib}}\}$.

2: Compute the local cutoff t_j as the empirical $(1 + 1/|I_j|)(1-\alpha)$-quantile of $\{s_i\}_{i \in I_j}$.

Step IV: Credible region construction

1: Assign $\mathbf{x}_{\text{obs}}$ to region $A_k = T(\mathbf{x}_{\text{obs}})$.

2: Return the credible region:

$$R_{\text{CP4SBI}}(\mathbf{x}_{\text{obs}}) = \{\theta \mid s(\theta; \mathbf{x}_{\text{obs}}) \leq t_k\}$$

Output: Credible region $R_{\text{CP4SBI}}(\mathbf{x}_{\text{obs}})$ with marginal and local $1-\alpha$ coverage

Algorithm 2: CP4SBI-CDF

Input: Calibration set $\mathcal{D} = \{(\theta_i, \mathbf{x}_i)\}_{i=1}^{B}$, estimated posterior $\widehat{p}(\theta \mid \mathbf{x})$, conformity score $s(\theta; \mathbf{x})$, nominal level α, new observation $\mathbf{x}_{\text{obs}}$

Step I: Score computation

1: For each $(\theta_i, \mathbf{x}_i) \in \mathcal{D}$, compute $s_i = s(\theta_i; \mathbf{x}_i)$ using $\widehat{p}$.

2: Form the scored dataset $\mathcal{D}' = \{(s_i, \mathbf{x}_i)\}_{i=1}^{B}$

Step II: Compute CDF scores

1: Compute $s'_i = s'(\theta_i; \mathbf{x}_i) = \widehat{F}(s(\theta_i; \mathbf{X}) \mid \mathbf{X} = \mathbf{x}_i)$ on $\mathcal{D}'$, using $\widehat{p}$ to estimate each $\widehat{F}(\cdot \mid \mathbf{X} = \mathbf{x}_i)$.

Step III: Cutoff estimation on transformed scores

1: Compute the cutoff $t_{1-\alpha}$ as the empirical $(1 + 1/B)(1-\alpha)$-quantile of $\{s'_i\}_{i=1}^{B}$

Step IV: Compute credibility set

1: Compute the set $C_{\text{CP4SBI}}(\mathbf{x}_{n+1})$ as:

$$\begin{aligned} R_{\text{CP4SBI}}(\mathbf{x}_{\text{obs}}) &= \{\theta \mid s'(\theta; \mathbf{x}_{\text{obs}}) \leq t_{1-\alpha}\} \\ &= \{\theta \mid s(\theta; \mathbf{x}_{\text{obs}}) \leq \widehat{F}^{-1}(t_{1-\alpha} \mid \mathbf{X} = \mathbf{x}_{\text{obs}})\}\ . \end{aligned}$$

Output: Credible region $R_{\text{CP4SBI}}(\mathbf{x}_{\text{obs}})$ with marginal and asymptotic conditional $1-\alpha$ coverage

Algorithm 3: KDE approximation of HPD score

Input: Calibration set $\mathcal{D} = \{(\theta_i, \mathbf{x}_i)\}_{i=1}^{B}$, estimated posterior $\widehat{p}(\theta \mid \mathbf{x})$, KDE budget L

For each $(\theta, \mathbf{x}) \in \mathcal{D}$, do:

  Simulate L samples $\tilde{\theta} \sim \widehat{p}(\cdot \mid \mathbf{x})$

  Fit a KDE on data generated from posterior estimator $\{\tilde{\theta}_l\}_{l=1}^{L}$

  Define $s(\theta; \mathbf{x}) = \frac{1}{L} \sum_{l=1}^{L} K(\theta, \theta_l)$

Output: Conformal scores $\{s(\theta_i, \mathbf{x}_i)\}_{i=1}^{B}$

B. Proofs

This appendix contains formal proofs for the theoretical results stated in Section 4. We organize the material according to the corresponding methods: `LoCart CP4SBI` and `CDF CP4SBI`.

B.1. `LoCart CP4SBI`

We begin by justifying the finite-sample local and marginal coverage guarantees of `LoCart CP4SBI`. The result follows directly from standard conformal prediction arguments applied independently within each partition element.

Proof of Theorem 4.1. This is a straightforward application of standard conformal prediction (see, e.g., Angelopoulos et al. 2023). Since the conformal threshold $t_{1-\alpha}(\mathbf{x})$ is computed within each region A_j using calibration scores that are exchangeable within that region, the conformal guarantee holds conditional on $\mathbf{x} \in A_j$. Marginal validity follows by averaging over all regions via the law of total probability. A complete version of this argument is presented in Cabezas et al. [2025a, Theorem 2] □

We now turn to the proof of the asymptotic conditional coverage guarantee for `LoCart CP4SBI`.

Proof of Theorem 4.2. This result follows from Cabezas et al. [2025a, Theorem 5], which establishes asymptotic conditional validity for local conformal prediction procedures under general regularity conditions. These include mild assumptions on the partition structure and the convergence of estimated quantiles to the true conditional quantiles within each region. □

B.2. `CDF CP4SBI`

We begin by formalizing the transformation used in `CDF CP4SBI`. Let

$$F(s(\theta_*;\mathbf{x}) \mid \mathbf{x}) = \int \mathbb{I}\{s(\theta;\mathbf{x}) \le s(\theta_*;\mathbf{x})\}\, p(\theta \mid \mathbf{x})\, d\theta$$

denote the true conditional cumulative distribution function (CDF) of the score $s(\theta;\mathbf{x})$, evaluated at a fixed test point $(\theta_*, \mathbf{x})$. The estimated version is defined as

$$s'(\theta_*;\mathbf{x}) = \widehat{F}(s(\theta_*;x) \mid \mathbf{x}) = \int \mathbb{I}\{s(\theta;\mathbf{x}) \le s(\theta_*;\mathbf{x})\}\, \widehat{p}(\theta \mid \mathbf{x})\, d\theta,$$

where $\widehat{p}(\theta \mid \mathbf{x})$ is an estimate of the conditional posterior density $p(\theta \mid \mathbf{x})$.

In the marginal setting, the `CDF CP4SBI` procedure can be interpreted as standard conformal prediction applied to a transformed score function $s'(\theta;x) = \widehat{F}(s(\theta;x) \mid x)$. The calibration scores $s'_i = \widehat{F}(s(\theta_i;\mathbf{X}_i) \mid \mathbf{X}_i)$ are computed from i.i.d. samples $(\theta_i, \mathbf{X}_i)$, and the conformal region is constructed using their adjusted empirical $(1-\alpha)$-quantile. Since the procedure follows the usual conformal recipe—changing only the score function—the marginal coverage guarantee holds by the standard conformal prediction argument.

Proof of Theorem 4.3. The procedure applies conformal prediction to the transformed scores $s'_i = \widehat{F}(s(\theta_i;\mathbf{X}_i) \mid \mathbf{X}_i)$, which are computed from exchangeable calibration pairs. Since the conformal region is defined by comparing the test score $s'(\theta;\mathbf{x})$ to the empirical $(1-\alpha)$-quantile of the calibration scores, the standard conformal guarantee ensures that

$$\mathbb{P}(s'(\theta;\mathbf{X}) \le t_{1-\alpha}) \ge 1-\alpha,$$

where $(\theta, \mathbf{X})$ is drawn jointly from the same distribution as the calibration data. This implies marginal coverage of the conformal region $C_{\text{cdf}}(\mathbf{x})$. □

We now turn to the conditional coverage properties of `CDF CP4SBI`. Unlike the marginal case, conditional validity depends on three sources of approximation: the accuracy of the posterior estimator $\widehat{p}(\theta \mid \mathbf{x})$, the Monte Carlo approximation used to compute the transformed score $\widehat{F}_M(s(\theta;\mathbf{x}) \mid \mathbf{x})$, and the estimation of the conformal quantile from a finite calibration set. To build intuition, we first consider an idealized setting in which the posterior is exact ($\widehat{p} = p$) and the conditional CDF $F(s(\theta;x) \mid x)$ is known exactly, so that the only remaining source of error comes from estimating the conformal quantile using a finite calibration set. In this setting, conditional coverage follows from the probability integral transform as $B \to \infty$, as formalized in Theorem B.1.

Theorem B.1 (Asymptotic conditional coverage: idealized case)**.** *Suppose $\widehat{p} = p$ and the CDF transformation $F(s(\theta;\mathbf{x}) \mid \mathbf{x})$ is known exactly. Then, as the calibration size $B \to \infty$, `CDF CP4SBI` achieves asymptotic conditional coverage:*

$$\lim_{B\to\infty} \mathbb{P}\big(\theta \in \widehat{C}(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}\big) = 1 - \alpha.$$

Proof. If $\widehat{p} = p$, then the transformed score becomes

$$s'(\theta;\mathbf{x}) = F(s(\theta;\mathbf{x}) \mid \mathbf{x}),$$

where $F(\cdot \mid \mathbf{x})$ denotes the true conditional CDF of the score. Since $\theta \sim p(\cdot \mid \mathbf{x})$, it follows from the probability integral transform that

$$s'(\theta;\mathbf{x}) \,\big|\, \mathbf{x} \;\sim\; \text{Uniform}(0,1).$$

Therefore, for any fixed $\mathbf{x}$,

$$\mathbb{P}\big(s'(\theta;\mathbf{x}) \le t_{1-\alpha} \mid \mathbf{x}\big) = t_{1-\alpha}.$$

Moreover, if the threshold $t_{1-\alpha}$ is computed as the empirical $(1 + 1/B)(1-\alpha)$-quantile of i.i.d. Uniform$(0,1)$ calibration scores, then

$$t_{1-\alpha} \to 1 - \alpha \quad \text{as } B \to \infty,$$

which implies that the procedure achieves asymptotic conditional coverage at level $1-\alpha$ [Dheur et al., 2025]. □

Suppose now that the posterior estimate differs from the true posterior, but that the CDF transformation $\widehat{F}(s(\theta;\mathbf{x}) \mid \mathbf{x})$ is still evaluated exactly from its integral definition under the approximate posterior. To make the dependence on the training sample size explicit, we write $\widehat{p}_K(\theta \mid \mathbf{x})$ for the posterior estimate obtained from a training set of size K, and keep $p(\theta \mid \mathbf{x})$ for the true posterior. Accordingly, we define

$$\widehat{F}_K(s(\theta;\mathbf{x}) \mid \mathbf{x}) \;=\; \int \mathbb{I}\{s(\theta';\mathbf{x}) \le s(\theta;\mathbf{x})\}\, \widehat{p}_K(\theta' \mid \mathbf{x})\, d\theta'.$$

In this setting, the transformed score $s'(\theta;\mathbf{x}) = \widehat{F}_K(s(\theta;\mathbf{x}) \mid \mathbf{x})$ is no longer uniformly distributed under p, and conditional validity may be compromised. Our goal is to quantify the deviation from the nominal coverage level as a function of the discrepancy between $\widehat{p}_K$ and p.

To recover conditional coverage in the limit, we require that the divergence between $\widehat{p}_K$ and p vanishes as the posterior approximation improves.

Assumption B.1 (KL convergence of the approximate posterior)**.** *Let*

$$\delta_K(\mathbf{x}) := \mathrm{KL}\left(\widehat{p}_K(\cdot \mid \mathbf{x}) \,\|\, p(\cdot \mid \mathbf{x})\right).$$

We assume that $\widehat{p}_K(\cdot \mid \mathbf{x})$ *is absolutely continuous with respect to* $p(\cdot \mid \mathbf{x})$*, and that for almost every* $\mathbf{x}$ *with respect to the distribution of* $\mathbf{X}$*, we have* $\delta_K(\mathbf{x}) \to 0$ *as* $K \to \infty$*.*

This assumption is reasonable when $\widehat{p}_K$ is obtained through methods that aim to approximate the true posterior by minimizing a divergence such as the KL. In such cases, it is natural to expect $\delta_K(\mathbf{x})$ to decrease as K grows, provided the optimization is effective and the approximating family is sufficiently flexible.

Theorem B.2 (Asymptotic conditional coverage: approximate posterior case)**.** *Under Assumption B.1, and assuming that the calibration size* $B \to \infty$*, the* `CDF CP4SBI` *procedure achieves asymptotic conditional coverage:*

$$\lim_{B\to\infty} \lim_{K\to\infty} \mathbb{P}\big(\theta \in C_{cdf}(\mathbf{x}) \,\big|\, \mathbf{X} = \mathbf{x}\big) \;=\; 1-\alpha \quad \textit{for almost every } \mathbf{x}.$$

Proof of Theorem B.2. If $\widehat{p}_K \neq p$, then the distribution of

$$s'(\theta;\mathbf{x}) \;=\; \widehat{F}_K(s(\theta;\mathbf{x}) \mid \mathbf{x})$$

under $p(\theta \mid \mathbf{x})$ is not uniform. For any fixed $\mathbf{x}$, the conditional coverage can be written as

$$\begin{aligned}\mathbb{P}\big(s'(\theta;\mathbf{x}) \le t \,\big|\, \mathbf{x}\big) &= \mathbb{P}\left(s(\theta;\mathbf{x}) \le \widehat{F}_K^{-1}(t \mid \mathbf{x}) \,\middle|\, \mathbf{x}\right) \\ &= F\left(\widehat{F}_K^{-1}(t \mid \mathbf{x}) \,\middle|\, \mathbf{x}\right),\end{aligned}$$

where $F(\cdot \mid \mathbf{x})$ and $\widehat{F}_K(\cdot \mid \mathbf{x})$ denote the CDFs of $s(\theta;\mathbf{x})$ under p and $\widehat{p}_K$, respectively.

Since $\widehat{p}_K(\cdot \mid \mathbf{x})$ is absolutely continuous with respect to $p(\cdot \mid \mathbf{x})$, Pinsker's inequality [Boucheron et al., 2013, Theorem 4.19] gives

$$\begin{aligned}\left|F\left(\widehat{F}_K^{-1}(t \mid \mathbf{x}) \mid \mathbf{x}\right) - t\right| &= \left|F\left(\widehat{F}_K^{-1}(t \mid \mathbf{x}) \mid \mathbf{x}\right) - \widehat{F}_K\left(\widehat{F}_K^{-1}(t \mid \mathbf{x}) \mid \mathbf{x}\right)\right| \\ &\le \sqrt{\frac{1}{2}\,\mathrm{KL}\left(\widehat{p}_K(\cdot \mid \mathbf{x}) \,\|\, p(\cdot \mid \mathbf{x})\right)} \\ &= \sqrt{\delta_K(\mathbf{x})/2}.\end{aligned}$$

Therefore,

$$\mathbb{P}\big(s'(\theta;\mathbf{x}) \le t \mid \mathbf{x}\big) \;\ge\; t - \sqrt{\delta_K(\mathbf{x})/2}.$$

The correction term $\sqrt{\delta_K(\mathbf{x})/2}$ quantifies the effect of posterior misspecification. Under Assumption B.1, we have $\delta_K(\mathbf{x}) \to 0$ for almost every $\mathbf{x}$ as $K \to \infty$, and hence

$$\lim_{K\to\infty} \mathbb{P}\big(s'(\theta;\mathbf{x}) \le t \mid \mathbf{x}\big) = t.$$

The remainder of the argument is identical to the proof of Theorem B.1, yielding the stated asymptotic conditional coverage. □

We now address the final source of approximation in the `CDF CP4SBI` procedure: the Monte Carlo estimation of the transformed score. While the previous results accounted for the error introduced by approximating the posterior $\widehat{p}_K(\theta \mid \mathbf{x})$, in practice the CDF $\widehat{F}_K(s(\theta;\mathbf{x}) \mid \mathbf{x})$ must also be approximated using a finite set of posterior samples. We show that this additional source of error vanishes as the number of samples grows, completing the proof of asymptotic conditional coverage in Theorem 4.4.

Proof of Theorem 4.4. Fix $\mathbf{x}$. Let $\{\theta_j\}_{j=1}^{M} \overset{\text{i.i.d.}}{\sim} \widehat{p}_K(\cdot \mid \mathbf{x})$ and denote by $\widehat{F}_{K,M}(\cdot \mid \mathbf{x})$ the empirical CDF based on these samples, and by $\widehat{F}_K(\cdot \mid \mathbf{x})$ its population counterpart under $\widehat{p}_K$. For any

$t \in (0,1)$, consider

$$\Delta_{K,M}(t;\mathbf{x}) := \Big| F(\widehat{F}^{-1}_{K,M}(t \mid \mathbf{x}) \mid \mathbf{x}) - t \Big|.$$

We decompose

$$\Delta_{K,M}(t;\mathbf{x}) \le \Big| F(\widehat{F}^{-1}_{K,M}(t) \mid \mathbf{x}) - F(\widehat{F}^{-1}_{K}(t) \mid \mathbf{x}) \Big| + \Big| F(\widehat{F}^{-1}_{K}(t) \mid \mathbf{x}) - t \Big|.$$

For fixed K, the Glivenko–Cantelli theorem ensures $\widehat{F}_{K,M} \to \widehat{F}_K$ uniformly as $M \to \infty$, hence $\widehat{F}^{-1}_{K,M}(t) \to \widehat{F}^{-1}_{K}(t)$. By continuity of $F(\cdot \mid \mathbf{x})$, the first difference converges to zero, so that

$$\lim_{M\to\infty} \Delta_{K,M}(t;\mathbf{x}) = \Big| F(\widehat{F}^{-1}_{K}(t) \mid \mathbf{x}) - t \Big|.$$

Next, under Assumption B.1 the divergence $\mathrm{KL}(\widehat{p}_K(\cdot \mid \mathbf{x}) \,\|\, p(\cdot \mid \mathbf{x}))$ vanishes as $K \to \infty$, which by Pinsker's inequality implies that the distribution functions converge. In particular,

$$\Big| F(\widehat{F}^{-1}_{K}(t) \mid \mathbf{x}) - t \Big| \to 0 \quad \text{for a.e. } \mathbf{x}.$$

Thus,

$$\lim_{K\to\infty} \lim_{M\to\infty} \Delta_{K,M}(t;\mathbf{x}) = 0.$$

In other words, the transformed scores converge to a uniform distribution in the limit, so that we are precisely in the setting of Theorem B.1. Therefore,

$$\mathbb{P}\big(\theta \in C_{\mathrm{cdf}}(\mathbf{x}) \,\big|\, X = \mathbf{x}\big) \to 1 - \alpha \quad \text{for a.e. } \mathbf{x}.$$

□

C. Experiment details

C.1. Benchmark details

Table 1 details the specific characteristics of each benchmark task, including the number of dimensions, the prior used, and the availability of true posterior samples. For tasks such as *Bernoulli GLM* (and Raw), *SLCP* (and Distractors), *Lotka-Volterra*, and *SIR*, only 10 pre-set observations with true posterior samples are available in the `sbibm` package. Meanwhile, the remaining tasks allow for a higher budget of up to 500 observations with true posterior samples.

Table 1: Summary of benchmark tasks and their key characteristics.

Task	**Parameters**	**Prior**	B_{sim}	**Description**
Bernoulli GLM	10	Conjugate	10	Generalized Linear Model with Bernoulli observations and uses sufficient statistics (10-D)
Bernoulli GLM Raw	10	Conjugate	10	Raw data version (100-D) of Bernoulli GLM.
Gaussian Linear	10	Gaussian	500	Mean inference with fixed covariance and conjugate prior
Gaussian Linear Unif.	10	Uniform	500	Same as Gaussian Linear, but using a uniform prior.
Gaussian Mixture	2	Uniform	500	Bimodal mixture of Gaussians and ABC benchmark case
Lotka-Volterra	4	Lognormal	10	An ecological model describing the dynamics of two interacting species, such as a prey-predator relationship.
SIR	2	Lognormal	10	An influential three-state epidemiological model parameterized by contact rate and mean recovery.
SLCP	5	Uniform	10	A challenging inference task that starts from a simple likelihood and a complex posterior.
SCLP Distractors	5	Uniform	10	Same task as SLCP, but with additional 100 noisy features (distractors).
Two Moons	2	Uniform	10	Crescent-shaped posterior and tests multimodal performance

For the *Gaussian Mixture* task, we made specific modifications to the default hyperparameters of the simulator and prior to improve its behavior and enable more fruitful comparisons. We set the uniform prior's limits to -3 and 3 and changed the multiplicative factor of each mixture mean

parameter (which are the parameters of interest) from 1 to 0.8.

C.2. Illustration Details

For both illustrations, we used a credibility level of $1 - \alpha = 0.9$ and an overall simulation budget of $B_{\text{all}} = 20000$. Of this budget, 80% was used for training the posterior approximator, with the remaining 20% reserved for the calibration step.

Nuisance parameter illustration. We utilized the 10-dimensional Gaussian Linear Uniform task, as detailed in Table 1. The first two parameters were set as our parameters of interest, $\phi = (\theta_1, \theta_2)$. The posterior estimator $\widehat{p}(\phi \mid \mathbf{x})$was a Neural Posterior Estimator (NPE) based on conditional normalizing flows from the `sbi` package, using its standard architecture. We compare the credible regions derived from the global, self-calibrated, and `CDF CP4SBI` methods for the fixed observation:

$$\mathbf{x}_{\text{obs}} = (0.3416, -0.4812, -0.0749, 0.3471, -0.7253, 0.1747, -0.1242, -0.3328, 0.0409, -0.5498),$$

which was generated from the fixed full parameter vector:

$$\theta = (0.25, 0.1, 0, 0, 0, 0, 0, 0, 0, 0),$$

with the true parameter of interest being the first two entries, $\phi = (0.25, 0.1)$.

Continuous-flow generative model credible set illustration. This illustration uses the 2-dimensional Gaussian Mixture task, as detailed in Table 1. To derive the posterior estimator, we used a Neural Posterior Score Estimator (NPSE) based on a conditional diffusion model [Geffner et al., 2023], using the standard architecture from the `sbi` [Boelts et al., 2025] implementation. Since the NPSE is a sample-based model, we use the KDE approximation of the HPD score detailed in Algorithm 3 to build the set $\mathcal{D}' = \{(s(\theta_i; \mathbf{x}_i), \mathbf{x}_i)\}_{i=1}^{B}$ and to compute conformal scores $s(\theta; \mathbf{x})$ for any other fixed $\mathbf{x}$, with Scott's rule used to determine the kernel bandwidth [Scott, 2015]. We compare the self-calibrated and `LoCart CP4SBI` credible regions for the fixed observation $\mathbf{x}_{\text{obs}} = (0.2651, -0.1454)$, which was generated under the fixed parameter $\theta = (0.15, -0.1)$.

C.3. Computational Details

To derive the partitions for `LoCart CP4SBI` using regression trees, we fixed the minimum number of samples per leaf at 300 for large overall budgets ($B_{\text{all}} = 10000$ or $B_{\text{all}} = 20000$) and at 75 for a small budget ($B_{\text{all}} = 2000$) to ensure well-populated partitions. Post-pruning was also performed via cost-complexity methods (*ccp_alpha*) to balance partition complexity and predictive performance, with the remaining hyperparameters set to the *scikit-learn* defaults [Pedregosa et al., 2011]. For `CDF CP4SBI`, its sole hyperparameter, the sample size for estimating the empirical CDF, M, was fixed at 1000.

C.4. Additional results

Conditional coverage performance for NPE-based posterior estimators is shown across benchmark tasks for a smaller overall budget ($B_{\text{all}} = 2000$) and a larger one ($B_{\text{all}} = 20000$), in comparison to the original budget of 10000. For the smaller budget, the performance of our approaches and Global CP is similar, with `CDF CP4SBI` being a top performer and the best method for this case, excelling in 9 out of 10 benchmarks. For this limited sample size, the Global CP approach may be a better option than `LoCart CP4SBI`, as its local partitioning strategy is less effective with sparse calibration data. On the other hand, when considering a larger budget, both `LoCart CP4SBI` and `CDF CP4SBI` show a clear advantage over competing approaches, with each excelling in 7

out of 10 benchmarks. The best competing method, Global CP, only performs well in 5 tasks, particularly in Lotka-Volterra. Ultimately, these results demonstrate the robustness of our method across different posterior estimator qualities and calibration budgets.

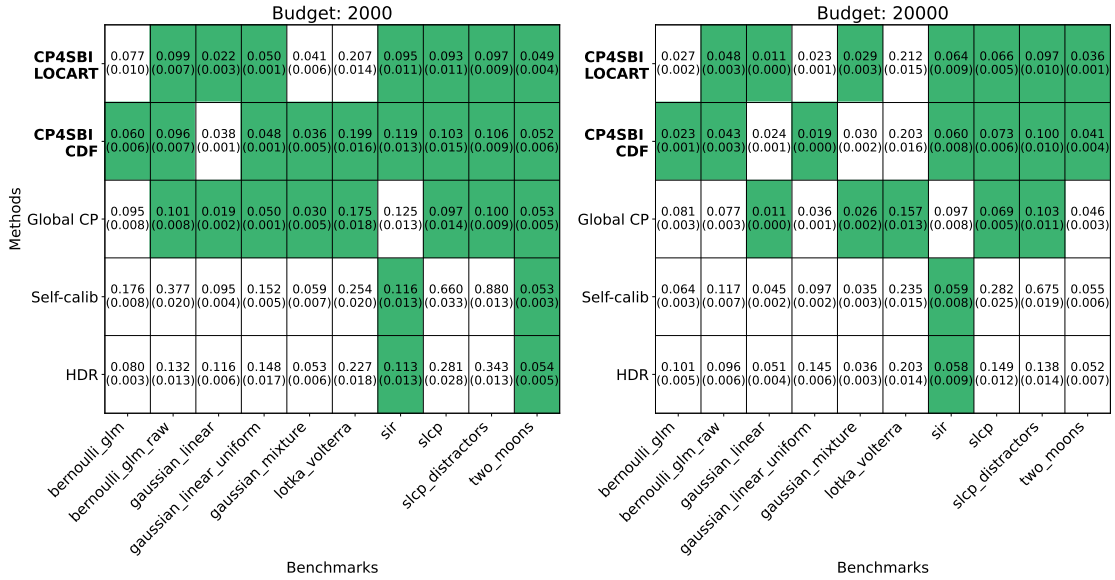

Figure 9: Conditional MAE for NPE-based posterior estimators across benchmark tasks for the overall budgets of $B_{\text{all}} = 2000$ (left) and $B_{\text{all}} = 20000$ (right). Mean metric values and 95% confidence intervals, calculated over 30 repetitions, are reported. Green cells indicate statistically significant superiority. For the smaller budget of $B_{\text{all}} = 2000$, `CDF CP4SBI` shows improved performance, standing out in 9 out of 10 benchmarks. For the larger budget of $B_{\text{all}} = 20000$, both `CDF CP4SBI` and `LoCart CP4SBI` perform well, each standing out in 7 out of 10 benchmarks. This showcases how our method adapts well across different budgets.