Simulator-Based Inference with WALDO: Confidence Regions by Leveraging Prediction Algorithms and Posterior Estimators for Inverse Problems

Luca Masserano[1,2]   **Tommaso Dorigo**[3]   **Rafael Izbicki**[4]   **Mikael Kuusela**[1,2]   **Ann B. Lee**[1,2]

[1]Department of Statistics & Data Science, Carnegie Mellon University
[2]NSF AI Planning Institute for Data-Driven Discovery in Physics, Carnegie Mellon University
[3]INFN, Sezione di Padova,   [4]Department of Statistics, Federal University of São Carlos

Abstract

Prediction algorithms, such as deep neural networks (DNNs), are used in many domain sciences to directly estimate internal parameters of interest in simulator-based models, especially in settings where the observations include images or complex high-dimensional data. In parallel, modern neural density estimators, such as normalizing flows, are becoming increasingly popular for uncertainty quantification, especially when both parameters and observations are high-dimensional. However, parameter inference is an inverse problem and not a prediction task; thus, an open challenge is to construct *conditionally valid* and *precise* confidence regions, with a guaranteed probability of covering the true parameters of the data-generating process, no matter what the (unknown) parameter values are, and without relying on large-sample theory. Many simulator-based inference (SBI) methods are indeed known to produce biased or overly confident parameter regions, yielding misleading uncertainty estimates. This paper presents WALDO, a novel method to construct confidence regions with finite-sample conditional validity by leveraging prediction algorithms or posterior estimators that are currently widely adopted in SBI. WALDO reframes the well-known Wald test statistic, and uses a computationally efficient regression-based machinery for classical Neyman inversion of hypothesis tests. We apply our method to a recent high-energy physics problem, where prediction with DNNs has previously led to estimates with prediction bias. We also illustrate how our approach can correct overly confident posterior regions computed with normalizing flows.

1 INTRODUCTION

The vast majority of modern machine learning targets prediction problems, with algorithms such as Deep Neural Networks (DNNs) being particularly successful with point predictions of a target variable $Y \in \mathbb{R}$ when the input vectors $\mathbf{x} \in \mathcal{X}$ represent complex high-dimensional data. In many science applications, however, one is often interested in the "inverse" problem of estimating the internal parameters of a data-generating process with reliable measures of uncertainty. The parameters of interest, which we denote by $\boldsymbol{\theta}$, are then not directly observed but are the "causes" of the observed data $\mathbf{x}$.

In order to make inference on internal parameters, one needs a statistical model that relates the (unknown) parameters to the observed data. In science and engineering, simulations are often used to model the behavior of complex systems in lieu of an analytical likelihood, when the latter is too complicated to be evaluated explicitly. Let $\mathcal{D} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ denote observable data, where the "sample size" n refers to the number of observations at a fixed configuration of the parameters $\boldsymbol{\theta}$. *Likelihood-free inference* (LFI), which is a form of simulator-based inference (SBI; Cranmer et al. (2020)), refers to parameter estimation in a setting where the likelihood function $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) := p(\mathcal{D}|\boldsymbol{\theta})$ itself is intractable, but the scientist, in lieu of an explicit likelihood, has access to a simulator that can generate $\mathcal{D}$ given any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

LFI has undergone a revolution in terms of the complexity of problems that can be tackled, both because of faster and more realistic simulators that can generate a large number of examples $\mathcal{T} = \{(\boldsymbol{\theta}^{(j)}, \mathcal{D}^{(j)})\}_{j=1}^B$, and because of more powerful AI techniques that can learn various quantities of interest from these simulations. DNNs – such as convolutional neural networks (CNNs) (LeCun et al., 1995) – are now used in many domain sciences to directly *predict* internal parameters of interest in statistical models, especially in settings where $\mathbf{x}$ represents images or other high-dimensional data. Recent examples include estimating the energy (θ) of muons that radiate photons when traversing a finely segmented calorimeter ($\mathbf{x}$) (Kieseler et al., 2022); es-

timating the mass of a galaxy cluster (θ) from velocities and projected radial distances ($\mathbf{x}$) for a particular line-of-sight of the observer relative to the galaxy cluster (Ho et al., 2019); and estimating the range and noise-to-signal covariance parameters ($\boldsymbol{\theta}$) of spatial Gaussian processes from spatial fields or variograms ($\mathbf{x}$) (Gerber and Nychka, 2021). In parallel, modern neural density estimators, such as normalizing flows, are becoming increasingly popular for uncertainty quantification, especially when both parameters $\boldsymbol{\theta}$ and observations $\mathbf{x}$ are high-dimensional. Recent examples include Boyda et al. (2021); Mishra-Sharma and Cranmer (2022); Lueckmann et al. (2021).

Purely predictive approaches are known to suffer from prediction bias in inverse problems, as the point prediction – e.g., $\mathbb{E}[\boldsymbol{\theta}|\mathbf{x}]$ under squared error loss – is generally different from the true (unknown) parameter $\boldsymbol{\theta}$. Concrete examples include Dorigo et al. (2022); Ho et al. (2019); Kiel et al. (2019), where attempts are made to correct for the observed bias post-hoc. At the same time, many posterior estimation methods are known to be overly confident, meaning that they yield confidence sets with empirical coverage lower than the desired nominal level (Hermans et al., 2021), hence leading to potentially misleading results.

At the heart of the matter is the fact that both predictive and posterior approaches in SBI rely heavily on how the values of $\boldsymbol{\theta}$ in the training set $\mathcal{T}$ are sampled. For reliable inference, however, the coverage guarantees of the confidence sets should be independent of the choice of prior $\pi_{\boldsymbol{\theta}}$, thereby allowing the user to design priors that can lead to tighter, *but guaranteed to be valid*, confidence sets. In this work, we present a solution without relying on large-sample theory or computationally intensive Monte Carlo sampling.

WALDO is a new LFI procedure that can leverage any prediction algorithm or neural posterior estimator to construct confidence regions for $\boldsymbol{\theta}$ with correct *conditional coverage*; that is, sets $\mathcal{R}(\mathcal{D})$ satisfying

$$\mathbb{P}(\boldsymbol{\theta} \in \mathcal{R}(\mathcal{D})|\boldsymbol{\theta}) = 1 - \alpha, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}, \qquad (1)$$

regardless of the size n of the observed sample, where $(1 - \alpha) \in (0, 1)$ is a prespecified confidence level. Correct conditional coverage implies correct *marginal coverage*, $\mathbb{P}(\boldsymbol{\theta} \in \mathcal{R}(\mathcal{D})) = 1 - \alpha$, but the former is a stronger requirement that checks that the confidence set is calibrated no matter what the true parameter is, whereas marginal coverage only requires the set to be calibrated on average over the parameter space $\boldsymbol{\Theta}$. WALDO reframes the Wald test (Wald, 1943) and leverages existing prediction or posterior algorithms to first compute a test statistic (Equation 4) based on estimates of the conditional mean $\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}]$ and conditional variance $\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]$. It then uses a recent approach (Dalmasso et al., 2021) to the Neyman construction (Neyman, 1937), which estimates critical values via quantile regression and converts hypothesis tests into a confidence region with finite-n conditional coverage. WALDO also

includes an independent diagnostics module to check that the constructed confidence sets achieve the correct nominal level of empirical coverage across the parameter space. Section 3.2 describes our methodology in detail, and Figure 1 summarizes its different components.

WALDO embraces the best sides of both the Bayesian and frequentist perspectives to statistical inference by providing confidence sets that *(i)* can effectively exploit available domain-specific knowledge, further constraining parameters when the prior is consistent with the data, and *(ii)* are guaranteed to have the nominal conditional coverage even in finite samples as long as the quantile regressor is well estimated, regardless of the correctness of the prior. WALDO is also amortized, meaning that once the procedure has been trained, it can be evaluated on any number of observations. We lay out the statistical and computational properties of WALDO, providing synthetic examples with analytical solutions to verify and support our claims (see Section 3.3 and Section 3.4). We then show its effectiveness on two complex applications, which confirm the results we obtained on the synthetic examples: the first one (Section 4.1) uses an established benchmark in SBI and leverages posterior distributions to construct valid confidence sets regardless of the prior distribution. The second application (Section 4.2) deals with a current problem in high-energy physics: inferring the energy of muons from a particle detector exploiting predictions from a custom CNN and an innovative source of information, i.e., the pattern of energy deposits left by muons in a finely segmented calorimeter. The results we obtain for this problem are of scientific interest by themselves, as a rigorous estimate of the uncertainty around estimated muon energies is essential in the search of new physics. A ready-to-use and flexible implementation of WALDO is available at `https://github.com/lee-group-cmu/lf2i`.

Notation We refer to parameters of interest as $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ and to a sample of size n of observable input data as $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, with $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and possibly $p \neq d$. Note that n is distinct from B, B' and B'', i.e., the number of simulations required at different steps of our method. We distinguish between observable data and actual observations by denoting the latter as D. We refer to confidence regions as $\mathcal{R}(\mathcal{D})$. The terms "set", "region" and (when $p = 1$) "interval" are used interchangeably.

2 RELATION TO OTHER WORK

There exist many approaches for calibrating predictive distributions $p(y|\mathbf{x})$ to achieve marginal or conditional validity in "forward" $\mathbf{x} \to y$ problems; examples include conformal inference (Vovk et al., 2005; Lei et al., 2018; Chernozhukov et al., 2021) and the calibration procedures of Bordoloi et al. (2010); Dey et al. (2022). In the Bayesian inference domain, such calibration procedures correspond to ensuring that an estimate $\widehat{p}(\theta|\mathbf{x})$ of the posterior $p(\theta|\mathbf{x})$ indeed cor-
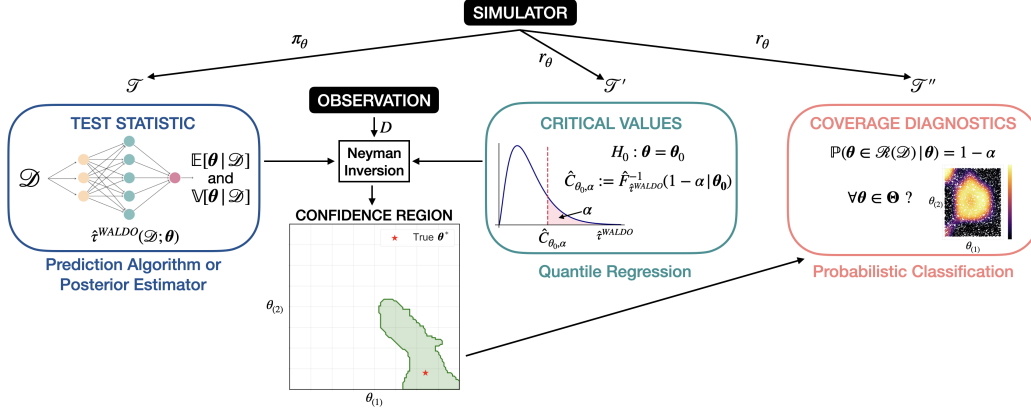
Luca Masserano, Tommaso Dorigo, Rafael Izbicki, Mikael Kuusela, Ann B. Lee

Figure 1: **Schematic diagram of WALDO.** *Left (blue)*: For a first simulated set $\mathcal{T}$, we estimate the conditional mean $\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}]$ and variance $\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]$ using a prediction algorithm (e.g., DNN) or posterior estimator (e.g., normalizing flows). This gives us the WALDO test statistic $\hat{\tau}^{\text{WALDO}}$ in Equation 4. *Center (green):* For a second simulated set $\mathcal{T}'$, we estimate critical values $\widehat{C}_{\boldsymbol{\theta}_0,\alpha}$ for all tests $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ across the parameter space $\boldsymbol{\Theta}$ via a quantile regression of $\hat{\tau}^{\text{WALDO}}$ on $\boldsymbol{\theta}$. *Bottom:* Given an observation D, Neyman inversion converts the tests (which compare test statistics with critical values) into a confidence region for θ. *Right (red):* For a third simulated set $\mathcal{T}''$, we provide an independent assessment of the conditional validity of constructed confidence regions by computing coverage diagnostics across the entire parameter space. See Section 3.2 and Algorithm 1 for details.

responds to the true posterior implied by the prior that was used. This work, on the other hand, deals with the question of constructing *confidence sets* with correct conditional coverage for internal unknown parameters θ in so-called "inverse problems" (recall Equation 1), which is not the same as achieving conditional coverage for prediction sets, or recalibrating posteriors.

Similarly, existing approaches for deep learning uncertainty quantification (see Gawlikowski et al. (2021) for a recent review), such as Monte Carlo drop out (Gal and Ghahramani, 2016) and conformal inference DNNs (Papadopoulos et al., 2007), construct prediction sets instead of confidence sets. Before WALDO, there has been no straightforward way to obtain confidence sets from point predictions or estimated posteriors obtained from deep neural networks and other predictive ML algorithms.

For example, various domain science applications have developed post-hoc corrections to predictive or posterior inferences to reduce observed biases and to improve the calibration of uncertainties. Such corrections are common in areas ranging from particle physics (Dorigo et al., 2022) to cosmology (Ho et al., 2019) and remote sensing (Kiel et al., 2019). Usually the goal of the corrections is to reduce the impact of the prior specification, but in contrast to WALDO, post-hoc correction approaches do not provide formal coverage guarantees. Similarly, in some settings, priors can be designed so that credible regions achieve correct conditional coverage (Bayarri and Berger, 2004; Berger, 2006; Kass and Wasserman, 1996; Scricciolo, 1999; Datta and Sweeting, 2005). However, this technique requires knowledge of the likelihood function (which is not available in LFI). Moreover, such prior distributions often do not encode actual prior information, a limitation that is not present in

WALDO.

Finally, posterior inferences do not control conditional coverage even for correctly specified priors (Patil et al., 2022). WALDO addresses this problem using Neyman inversion via an efficient regression-based approach proposed in Dalmasso et al. (2021). In the latter work, however, the authors construct likelihood-based test statistics (the Bayes factor or likelihood ratio) which require an extra numerical integration or optimization step that can lead to a loss of power of the resulting confidence sets. WALDO, on the other hand, has the ability of directly leveraging flexible prediction algorithms and posterior estimators to construct valid and potentially more precise finite-n confidence sets.

3 METHODOLOGY

WALDO leverages a regression-based approach to the Neyman construction, reframing the Wald test to use the output of common LFI prediction algorithms and posterior estimators. After outlining its statistical foundations, we describe our procedure and its properties using synthetic examples.

3.1 Foundational Tools from Classical Statistics

Neyman construction A key ingredient of WALDO is the equivalence between hypothesis tests and confidence sets, which was formalized by Neyman (1937). The basic idea is to invert a series of level-α hypothesis tests of the form

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \qquad (2)$$

for all $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$. After observing a sample D, one constructs a confidence region $\mathcal{R}(D)$ for θ by taking all $\boldsymbol{\theta}_0$ values that were not rejected by the series of tests above. By design, the

set $\mathcal{R}(\mathcal{D})$ satisfies Equation (1), i.e., it has the correct $1 - \alpha$ coverage across the *entire* parameter space Θ. Albeit simple, the Neyman construction requires one to control the type I error for every $\theta \in \Theta$. It is therefore hard to implement in practice within an LFI setting, without resorting to large-n approximations like Wilks' theorem (Wilks, 1938), or to Monte Carlo approaches, which become computationally prohibitive as the dimensionality of the parameter space increases (Cousins (2018); see also Section 3.4).

Wald test   Since any test that controls the type I error at level α can be used for the Neyman construction, we base WALDO on the classical Wald test (Wald, 1943), which is uniformly most powerful in many settings (Ghosh, 1991; Lehmann et al., 2005). The Wald test measures the agreement of the data with the null hypothesis for θ, and it has the following form for $p = 1$:

$$\tau^{\text{WALD}}(\mathcal{D}; \theta_0) := \frac{(\widehat{\theta}^{\text{MLE}} - \theta_0)^2}{\mathbb{V}(\widehat{\theta}^{\text{MLE}})}, \tag{3}$$

where $\widehat{\theta}^{\text{MLE}}$ is the maximum-likelihood estimator of θ and $\widehat{\mathbb{V}}(\widehat{\theta}^{\text{MLE}})$ can be any consistent estimator of its variance. However, in our setting, we do not have access to the likelihood and we cannot resort to assumptions on the distribution of $\tau^{\text{WALD}}(\mathcal{D}; \theta_0)$, nor to asymptotic regimes, which makes it difficult to directly compute the Wald test statistic.

3.2   Confidence Sets from Predictions and Posteriors

From Wald to WALDO   WALDO reframes the Wald test by replacing $\widehat{\theta}^{\text{MLE}}$ and its variance with quantities that are easily computable with prediction algorithms or posterior estimators commonly used in LFI. We define the WALDO test statistic for parameters of arbitrary dimensionality p as

$$\tau^{\text{WALDO}}(\mathcal{D}; \theta_0) = (\mathbb{E}[\theta | \mathcal{D}] - \theta_0)^T \mathbb{V}[\theta | \mathcal{D}]^{-1} (\mathbb{E}[\theta | \mathcal{D}] - \theta_0), \tag{4}$$

where $\mathbb{E}[\theta | \mathcal{D}]$ and $\mathbb{V}[\theta | \mathcal{D}]$ are, respectively, the conditional mean and covariance matrix of θ given $\mathcal{D}$. The connection to the Wald test follows from the asymptotic behavior of Bayes estimators (e.g., Chao (1970); Ghosh and Ramamoorthi (2003); Ghosh et al. (1982); Li et al. (2020)):

$$\mathbb{E}[\theta | \mathcal{D}] - \widehat{\theta}^{\text{MLE}} = \mathcal{O}_p(n^{-1/2}) \quad \text{and}$$
$$\mathbb{V}[\theta | \mathcal{D}] - \frac{1}{n} H^{-1}(\widehat{\theta}^{\text{MLE}}) = \mathcal{O}_p(n^{-1}),$$

where $H^{-1}(\widehat{\theta}^{\text{MLE}})$ is the negative inverse Fisher information matrix evaluated at $\widehat{\theta}^{\text{MLE}}$. The above result implies that WALDO would enjoy the same asymptotic properties typical of the Wald test, making it a pivotal test statistic. On the other hand, this does not mean that Wald and WALDO will give the same results for small n: indeed, in Section 3.3 and Appendix B.2, we demonstrate that WALDO can benefit

from a prior over θ that is consistent with the data to achieve smaller confidence sets, whereas the Wald test statistic only depends on the likelihood.

Likelihood-Free Frequentist Inference (LF2I)   WALDO expands on the LF2I framework formalized in Dalmasso et al. (2021), which proposed a fast construction of Neyman confidence sets using quantile regression to bypass large-sample approximations or expensive Monte-Carlo simulations. In its original formulation, the LF2I machinery includes three modular procedures which, respectively, **(i)** estimate a likelihood-based test statistic via odds ratios, **(ii)** estimate critical values $C_{\theta, \alpha}$ via quantile regression, and **(iii)** check that the constructed confidence sets achieve the desired coverage level for all $\theta \in \Theta$. Each module is based on a independent simulated sample from a high-fidelity simulator F_θ. WALDO replaces **(i)** and instead uses posteriors or predictions to compute τ^{WALDO} in (4). We break down the construction of a confidence set (including diagnostics) in the following steps, as outlined in Figure 1 and Algorithm 1:

(i) Estimate the test statistic via prediction algorithms or neural posterior estimators.   Use the simulated set $\mathcal{T} = \{(\theta^{(j)}, \mathcal{D}^{(j)})\}_{j=1}^{B}$, where θ can be drawn from any prior distribution π_θ, to estimate $\mathbb{E}[\theta | \mathcal{D}]$ and $\mathbb{V}[\theta | \mathcal{D}]$. This can be done by choosing between two methods: if using a prediction algorithm, we can leverage the fact that they approximate the conditional mean of the outcome variable given the inputs $\mathcal{D}$, when minimizing the squared error loss (lines 4-6 in Algorithm 1). Conversely, if using modern neural posterior estimators (such as normalizing flows (Papamakarios et al., 2021)), we can approximate $\mathbb{E}[\theta | \mathcal{D}]$ and $\mathbb{V}[\theta | \mathcal{D}]$ via Monte Carlo sampling from the estimated posterior distribution (lines 16-18 in Algorithm 1);

(ii) Estimate critical values via quantile regression.   Estimate $C_{\theta, \alpha} := F_{\widehat{\tau}^{\text{WALDO}}}^{-1}(1 - \alpha | \theta)$ by learning the conditional $(1 - \alpha)$-quantile of $\widehat{\tau}^{\text{WALDO}}(\mathcal{D}; \theta)$ using quantile regression over a simulated set $\mathcal{T}' = \{(\theta^{(j)}, \mathcal{D}^{(j)})\}_{j=1}^{B'}$, where θ is drawn uniformly (r_θ in Figure 1) over Θ to allow calibration $\forall \theta \in \Theta$;

(i) + (ii) Neyman inversion.   Once D is observed, evaluate $\widehat{\tau}^{\text{WALDO}}(D; \theta_0)$ and $\widehat{C}_{\theta_0; \alpha}$ over a fine grid of parameters $\theta_0 \in \Theta$, and retain all θ_0 for which the corresponding test does not reject the null:

$$\mathcal{R}(D) = \{\theta_0 \in \Theta : \tau^{\text{WALDO}}(D; \theta_0) \leq \widehat{C}_{\theta_0, \alpha}\}.$$

As we show in Appendix A, step *(ii)* leads to valid level-α hypothesis tests as long as the quantile regressor is well estimated, which then implies that $\mathcal{R}(D)$ satisfies conditional coverage (Eq. 1) at level $1 - \alpha$, regardless of the true value of θ and of the size n of the observed sample D.

(iii) Coverage diagnostics.   To check that the constructed confidence sets indeed achieve the desired level of conditional coverage, we leverage the diagnostics procedure

Algorithm 1 Confidence set for $\boldsymbol{\theta}$ via WALDO

1: **// Estimate building blocks of test statistic**
2: Simulate $\mathcal{T} = \{(\boldsymbol{\theta}^{(j)}, \mathcal{D}^{(j)})\}_{j=1}^{B}$
3: **if** prediction algorithm **then**
4:     Estimate $\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}]$ on $\mathcal{T}$ under squared error loss
5:     Compute $\{\mathbf{z}^{(j)} = (\boldsymbol{\theta}^{(j)} - \mathbb{E}[\boldsymbol{\theta}|\mathcal{D}^{(j)}])^2\}_{j=1}^{B}$
6:     Estimate $\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}] = \mathbb{E}[\mathbf{z}|\mathcal{D}]$ under squared error loss
7: **else if** posterior estimator **then**
8:     Estimate posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ on $\mathcal{T}$
9: **end if**

10: **// Estimate critical values**
11: Simulate $\mathcal{T}' = \{(\boldsymbol{\theta}^{(j)}, \mathcal{D}^{(j)})\}_{j=1}^{B'}$
12: **if** prediction algorithm **then**
13:     Predict $\{\widehat{\mathbb{E}}[\boldsymbol{\theta}|\mathcal{D}^{(j)}], \widehat{\mathbb{V}}[\boldsymbol{\theta}|\mathcal{D}^{(j)}]\}_{j=1}^{B'}$
14: **else if** posterior estimator **then**
15:     **for** each $\mathcal{D}$ that appears in $\mathcal{T}'$ **do**
16:         Draw N samples from $\widehat{p}(\boldsymbol{\theta}|\mathcal{D})$
17:         $\widehat{\mathbb{E}}[\boldsymbol{\theta}|\mathcal{D}] \approx \frac{\sum_i \boldsymbol{\theta}_i}{N}$
18:         $\widehat{\mathbb{V}}[\boldsymbol{\theta}|\mathcal{D}] \approx \frac{\sum_i (\boldsymbol{\theta}_i - \widehat{\mathbb{E}}[\boldsymbol{\theta}|\mathcal{D}])(\boldsymbol{\theta}_i - \widehat{\mathbb{E}}[\boldsymbol{\theta}|\mathcal{D}])^T}{N-1}$
19:     **end for**
20: **end if**
21: Compute $\{\widehat{\tau}^{\text{WALDO}}(\mathcal{D}^{(j)}; \boldsymbol{\theta}^{(j)})\}_{j=1}^{B'}$
22: Estimate critical values $C_{\boldsymbol{\theta},\alpha}$ via quantile regression of $\widehat{\tau}^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta})$ on $\boldsymbol{\theta}$

23: **// Neyman inversion**
24: **if** prediction algorithm **then**
25:     Predict $\widehat{\mathbb{E}}[\boldsymbol{\theta}|D]$ and $\widehat{\mathbb{V}}[\boldsymbol{\theta}|D]$
26: **else if** posterior estimator **then**
27:     Draw N samples from $\widehat{p}(\boldsymbol{\theta}|D)$
28:     $\widehat{\mathbb{E}}[\boldsymbol{\theta}|D] \approx \frac{\sum_i \boldsymbol{\theta}_i}{N}$
29:     $\widehat{\mathbb{V}}[\boldsymbol{\theta}|D] \approx \frac{\sum_i (\boldsymbol{\theta}_i - \widehat{\mathbb{E}}[\boldsymbol{\theta}|D])(\boldsymbol{\theta}_i - \widehat{\mathbb{E}}[\boldsymbol{\theta}|D])^T}{N-1}$
30: **end if**
31: Predict $\widehat{C}_{\boldsymbol{\theta}_0,\alpha} \; \forall \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_{grid}$
32: Initialize $\mathcal{R}(D) \leftarrow \emptyset$
33: **for** $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_{grid}$ **do**
34:     **if** $\widehat{\tau}^{\text{WALDO}}(D; \boldsymbol{\theta}_0) \leq \widehat{C}_{\boldsymbol{\theta}_0;\alpha}$ **then**
35:         $\mathcal{R}(D) \leftarrow \mathcal{R}(D) \cup \{\boldsymbol{\theta}_0\}$
36:     **end if**
37: **end for**

38: **return** confidence set $\mathcal{R}(D)$

---

introduced in Dalmasso et al. (2021). In detail: simulate a set $\mathcal{T}'' = \{(\boldsymbol{\theta}^{(j)}, \mathcal{D}^{(j)})\}_{j=1}^{B''}$ and construct a confidence region for each $\mathcal{D}^{(j)} \in \mathcal{T}''$. Then model $\mathbb{1}\{\boldsymbol{\theta}^{(j)} \in \mathcal{R}(\mathcal{D}^{(j)})\}$ as a function of $\boldsymbol{\theta}^{(j)}$ adopting a suitable probabilistic classification method. By definition, this will estimate $\mathbb{E}[\mathbb{1}\{\boldsymbol{\theta} \in \mathcal{R}(\mathcal{D})|\boldsymbol{\theta}] = \mathbb{P}[\boldsymbol{\theta} \in \mathcal{R}(\mathcal{D})|\boldsymbol{\theta}]$ across the whole parameter space. Note that this module is completely *independent* from (i) and (ii). As such, it can be used to to check the empirical conditional coverage of any uncertainty estimate, as

illustrated in Section 3.4 for Neyman confidence sets where critical values are estimated via Monte Carlo sampling, in Section 4.1 for posterior credible regions, and in Section 4.2 for prediction sets from the output of a CNN.

3.3 Statistical Properties: Coverage and Power

We now show that the coverage guarantees of WALDO are independent from the prior distribution, which can also be chosen to increase power. We do so through univariate Gaussian examples with analytically computable solutions. Since $p = 1$, we use simple prediction algorithms to estimate $\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}]$ and $\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]$. See Appendix C.1 for details.

PROPERTY I: WALDO guarantees conditional coverage across Θ, regardless of the specified prior. Scientists sometimes have domain-specific knowledge that can guide inference through the elicitation of a prior distribution over the parameters of interest. The goal is to introduce a bias to help quantifying the uncertainty, but if the prior happens to be at odds with the data, then this bias can be harmful and cause posteriors to be overconfident and smaller than they should be (Hermans et al., 2021). Ideally, we would want the coverage guarantees of any estimated parameter region to be preserved under this bias. In this example, we assume $\theta \sim \mathcal{N}(0, 2)$, $\mathcal{D}|\theta \sim \mathcal{N}(\theta, 1)$. As Figure 2 shows, confidence sets for θ (left panel) constructed through Neyman inversion of a series of Wald tests guarantee the correct conditional coverage (right panel), since Wald is only influenced by the likelihood. Conversely, prediction sets ($\mathbb{E}[\theta|\mathcal{D}] \pm 1.96\sqrt{\mathbb{V}[\theta|\mathcal{D}]}$) are influenced by the prior through the bias induced in the point predictions, which increases with the distance from the prior mean and results in strong under-coverage. WALDO exploits the same inputs of prediction sets ($\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}]$ and $\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]$), but corrects this problem by calibrating the critical values via quantile regression, hence guaranteeing conditional coverage. Note that we only use a single observation ($n = 1$) for each confidence set.
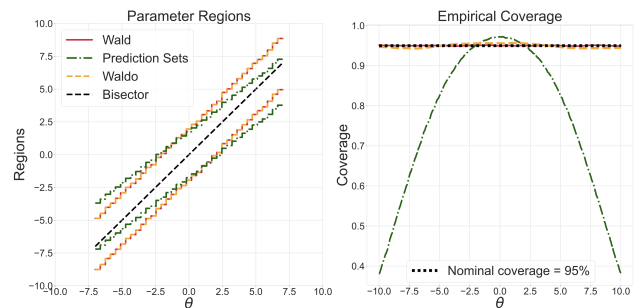

Figure 2: **PROPERTY I: WALDO guarantees conditional coverage across Θ, regardless of the specified prior.** Prior: $\theta \sim \mathcal{N}(0, 2)$. Likelihood: $\mathcal{D}|\theta \sim \mathcal{N}(\theta, 1)$. *Left:* median of upper/lower bounds of constructed parameter regions. *Right:* empirical coverage computed numerically using 100,000 samples for each θ over a fine grid in Θ (i.e., not using coverage diagnostics).
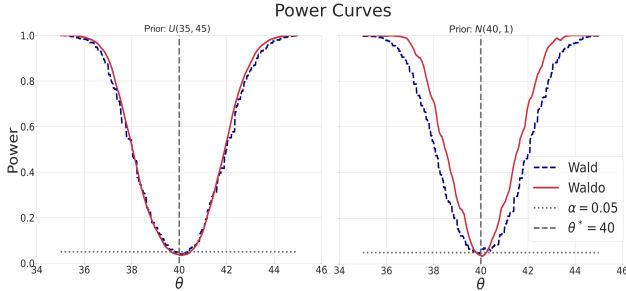
Figure 3: **PROPERTY II: WALDO exploits prior information and achieves higher power.** Power curves computed by recording the number of times a wrong value of θ is correctly outside the confidence set over 1,000 repetitions. Likelihood: $\mathcal{D} \sim \mathcal{N}(40, 1)$. *Left:* Wald and WALDO are equivalent when $\theta \sim \mathcal{U}(35, 45)$. *Right:* WALDO has higher power when $\theta \sim \mathcal{N}(40, 1)$.

PROPERTY II: WALDO exploits prior information and achieves higher statistical power. When the prior is correctly specified, we would like to leverage the induced bias to increase the power of the inverted tests and produce tighter constraints on the parameters, while retaining conditional coverage. Here we simulate data from a unique "true" Gaussian likelihood $\mathcal{D}|\theta \sim \mathcal{N}(\theta = 40, 1)$, and investigate the effect that the informativeness of the prior has on the power of the resulting tests. As Figure 3 shows, WALDO and Wald coincide when the prior is uninformative ($\theta \sim \mathcal{U}(35, 45)$; left panel), but the former has higher power when the prior is instead correctly specified ($\theta \sim \mathcal{N}(40, 1)$; right panel), thereby leading to smaller confidence sets.

3.4 Computational Properties

Scaling with high-dimensional parameters As mentioned in Section 3.2, WALDO exploits a simulated set sampled uniformly[1] over Θ to estimate critical values via quantile regression and guarantee coverage across the whole parameter space. While this might seem a daunting requirement, the only alternative to guarantee conditional coverage is to resort to Monte Carlo approaches that sample many times *at each* $\theta \in \Theta$. As Figure 4 shows, WALDO requires several orders of magnitude less simulations to achieve the correct calibration. This is true already when $p = 1$, and is even more evident when $p = 10$.

Quality of models WALDO relies on two estimation procedures (**(i)** and **(ii)** below) to construct the confidence set itself. The accuracy of the results relies on the estimation quality of these models and on the number of simulations B and B' that are available. In addition, there is a diagnostics procedure **(iii)** to estimate the conditional coverage of the final confidence sets, as a separate check that Equation 1 indeed holds.

---

[1]Technically, we only need to sample from a distribution that places mass on all Θ.

(i) Test statistic. The quality of prediction algorithms and posterior estimators is positively correlated with the power of the resulting tests. As the precision in the estimates of $\mathbb{E}[\theta|\mathcal{D}]$ and $\mathbb{V}[\theta|\mathcal{D}]$ decreases, the variance of the test statistics increases, which implies more conservative critical values and larger confidence regions. A good prior distribution will clearly help in achieving more precise estimates in regions of interest in the parameter space.

(ii) Critical values. As we prove in Appendix A, conditional coverage is achieved as long as the quantile regressor is well estimated. In practice, we observe that little hyper-parameter optimization is needed and that the number of simulations required to achieve well-calibrated critical values is usually a small fraction of those needed for the test statistic.

(iii) Diagnostics. The quality of the probabilistic classifier used to check the empirical coverage probability affects only the reliability of the diagnostics. Note that this module is completely independent of the others, and we can check its quality by inspecting the cross-entropy loss, and the standard errors and confidence bands on the estimates that common statistical packages provide (e.g., `MGCV` in `R`).

4 RESULTS

We assess the performance of WALDO on two challenging experiments. In the first example (Section 4.1), we show how to use a posterior distribution estimated via normalizing flows to compute valid confidence regions, and how prior information can improve precision. The second example (Section 4.2) tackles a complex particle energy reconstruction problem in high-energy physics: we leverage predictions from a custom CNN to construct confidence intervals with correct coverage and high power.

4.1 Confidence Sets from Neural Posteriors

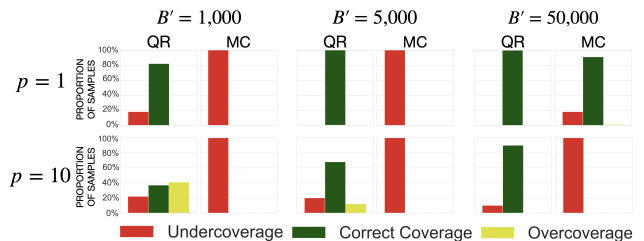This inference task was introduced in Sisson et al. (2007) and has become a standard benchmark in the SBI literature

Figure 4: **Quantile regression (QR) is orders of magnitude more efficient than Monte Carlo (MC) in terms of the number of simulations B' required to achieve correct coverage.** Each panel shows the fraction of samples (out of 1,000 total) for which the selected method to estimate critical values achieves approximately correct coverage ($\mathbb{P}(\theta \in \mathcal{R}(\mathcal{D})|\theta) \in [0.95 \pm 0.03]$). Prior: $\theta \sim \mathcal{N}(0, 0.1 \cdot \mathbf{I})$. Likelihood: $\mathcal{D}|\theta \sim \mathcal{N}(\theta, 0.1 \cdot \mathbf{I})$. In both cases, we used normalizing flows to estimate the posterior.
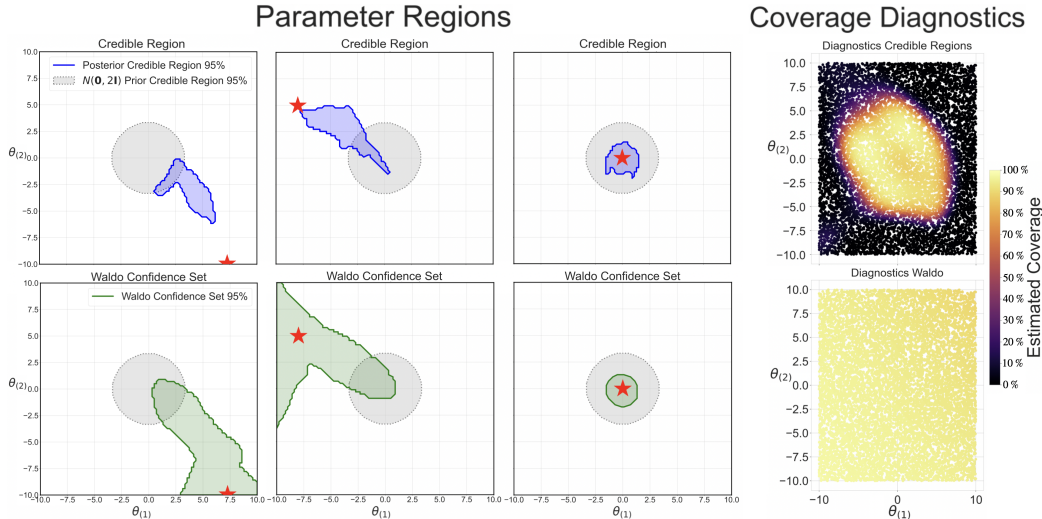
Luca Masserano, Tommaso Dorigo, Rafael Izbicki, Mikael Kuusela, Ann B. Lee

Figure 5: **WALDO converts posterior distributions into confidence regions with correct conditional coverage and high power**. *Left Panel - Top:* Examples of 95% credible regions (blue) from posteriors estimated with normalizing flows and a Gaussian $\mathcal{N}(\mathbf{0}, 2\mathbf{I})$ prior (gray) for different values of the true unknown parameter $\boldsymbol{\theta}^*$ (red star). *Right Panel - Top:* Credible regions have conditional coverage close to the nominal level only in a neighborhood of the prior, and severely undercover everywhere else. *Left Panel - Bottom:* Corresponding 95% WALDO confidence sets (green), derived from the same posterior estimates used for the top row. *Right Panel - Bottom:* Conditional coverage for WALDO confidence sets achieves the nominal 1-α level everywhere, where $\alpha = 0.05$.

(Clarté et al., 2021; Toni et al., 2009; Simola et al., 2021; Lueckmann et al., 2021). It consists of estimating the (common) mean of the components of a two-dimensional Gaussian mixture, with one component having much broader covariance: $\mathcal{D}|\boldsymbol{\theta} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\theta}, \mathbf{I}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\theta}, 0.01 \cdot \mathbf{I})$, where $\boldsymbol{\theta} \in \mathbb{R}^2$ and $n = 1^2$. We estimate $p(\boldsymbol{\theta}|\mathcal{D})$ using the implementation of Neural Posterior Estimators (NPE) of Durkan et al. (2020) through the `SBI` software package (Tejero-Cantero et al., 2020), and report results obtained with two different priors: $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, 2 \cdot \mathbf{I})$ and $\boldsymbol{\theta} \sim \mathcal{U}([-10, 10]^2)$ (the latter in Appendix B.2). We estimate the critical values with a 2-layer neural network minimizing the quantile loss. Simulated datasets used for training are of the following sizes: $B = 100{,}000$, $B' = 30{,}000$ when using a Gaussian prior. Conditional mean and variance were approximated with 50,000 Monte Carlo samples from the neural posterior.

The first four panels on the left of Figure 5 show examples of 95% credible regions (top) and WALDO confidence sets (bottom) obtained from the same posterior distribution, when the true parameter is far from the prior. If the data is at odds with the prior, then the induced bias leads to credible regions that severely undercover across the parameter space, as it is shown at the top of the rightmost panel, where the coverage probability for credible regions reaches values as low as 0-10%. WALDO can correct for this bias and output larger confidence sets which account for the added uncertainty, thereby leading to correct conditional coverage everywhere (bottom of rightmost panel). This is the same

behaviour seen in the first example of Section 3.3, although for a more complex setting and for a posterior estimator.

Conversely, when the prior is consistent with the data (Figure 5, right two panels of "Parameter Regions"), WALDO is not overly conservative and leverages the additional information to tighten the constraints on the parameters, closely tracking the size of the posterior credible region. In Appendix B.2, we also show that, over many independent observations, the average size of WALDO confidence sets is indeed smaller when using an informative prior than when using a Uniform over $\boldsymbol{\Theta}$. These results closely mimic those seen in the second example of Section 3.3.

4.2 Confidence Sets for Muon Energies using CNN Predictions

We now discuss the performance of WALDO on an application of interest to fundamental research: estimating the energy of muons at a future particle collider. Muons are a heavier replica of electrons; they are produced in subnuclear reactions involving electroweak interactions. Muons are also excellent probes of new phenomena: in fact, their detection and measurement has been key to several crucial discoveries in the past decades, including the Higgs boson (Augustin et al., 1974; Herb et al., 1977; CDF Collaboration, 1995; Aad et al., 2012; Chatrchyan et al., 2012). Traditionally, the energy of a muon is determined from the curvature of its trajectory in a magnetic field, but at energies above a few TeV this methods breaks down as trajectories become indistinguishable from straight paths even within the strongest practically achievable fields. Searching for viable alterna-

---

[2]WALDO works for an observed sample of any size, but we had to use $n = 1$ because the `SBI` Python library we used to estimate the posterior does not yet support larger sample sizes for NPE.
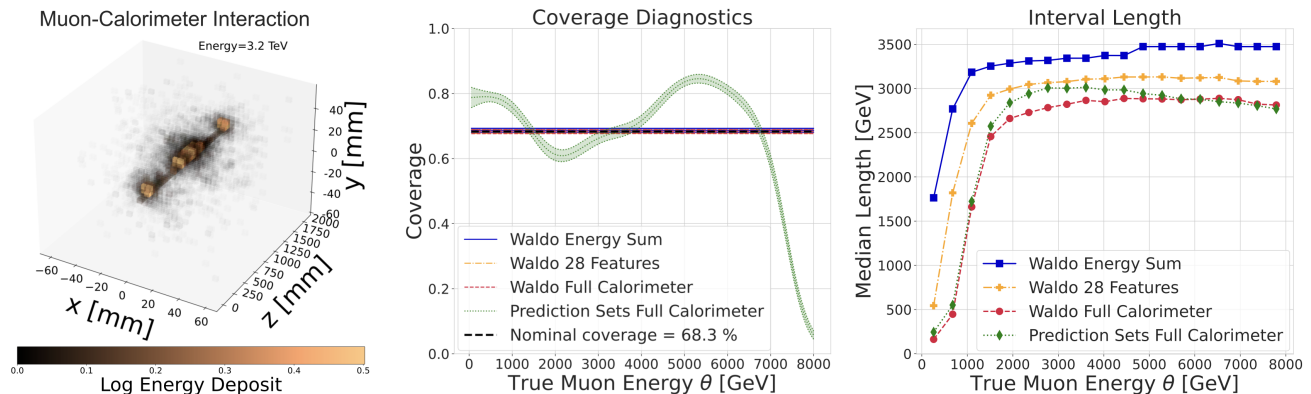
Figure 6: **WALDO guarantees the nominal coverage level, and yields smaller confidence intervals (more precise estimates of muon energy) with the higher-granularity ("full") calorimeter data.** *Left:* Energy deposited by a $\theta \approx 3.2$ TeV muon entering a calorimeter with $32 \times 32 \times 50$ cells. *Center:* WALDO (blue, orange, red in the right two panels) guarantees nominal coverage (68.3%), while 1σ prediction intervals (green) under- or over-cover in different regions of Θ. *Right:* Median lengths of constructed intervals: shorter intervals imply higher precision in the estimates. Prediction sets are on average wider than the corresponding confidence sets, using the same data.

tives, it has been observed (Kieseler et al., 2022; Dorigo et al., 2022) that both the pattern and the magnitude of small radiative energy losses that muons withstand in traversing dense and finely segmented calorimeters can be used to infer the incident muon energy, leveraging the capacity of modern deep learning architectures. Nonetheless, the above work also clearly showed that predictions of θ suffered from a strong bias, mainly due to the high nonlinearity of the response at very high energies. Motivated by this problem, we pose two questions: **(i)** Can we construct confidence sets with correct coverage of the true energy of muons using the information contained in the pattern and magnitude of radiative deposits in a dense calorimeter? **(ii)** Is it possible to extract additional information from finer segmentations of the calorimeter to allow for tighter constraints (i.e., smaller confidence sets with correct coverage) on muon energy estimates? Quantifying the latter would allow scientists to optimize their detector designs, since manufacturing very small calorimeter cells is expensive.

We have available 886,716 3D input "images" $\mathbf{x}$ and scalar muon energies θ obtained through GEANT4 (Agostinelli et al., 2003), a high-fidelity stochastic simulator. See Figure 6 (left panel) for an illustration of one simulated $\mathbf{x}_i$ for a particular θ_i. The data are available in Kieseler et al. (2021). As the interest is on constraining muon energies as much as possible while guaranteeing conditional coverage, we use three versions of the same dataset with increasing dimensionality: a 1D input equal to the sum over all calorimeter cells with deposited energy $E > 0.1$ GeV, for each muon; 28 custom features extracted from the spatial and energy information of the calorimeter cells (see Kieseler et al. (2022)); and the full calorimeter measurements ($\mathbf{x}_i \in \mathbb{R}^{51,200}$). For the first two datasets, we estimate $\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}]$ and $\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]$ via Gradient Boosting (Chen and Guestrin, 2016). For the full calorimeter data, we rely on the CNN developed by Kieseler et al. (2022). We use Gradient Boosting for quantile regres-

sion (Pedregosa et al., 2011).

Answering **(i)** affirmatively, Figure 6 (center) shows that confidence sets constructed with WALDO achieve exact conditional coverage (68.3%) regardless of the dataset used. The corresponding 1σ prediction intervals ($\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] \pm \sqrt{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$) using full calorimeter data, instead, exhibit over- or under-coverage in different regions over Θ, which in the latter case means that prediction sets contain the true value with much lower probability than anticipated. As for question **(ii)**, we make two observations (see Figure 6; right panel): First, using the raw higher-dimensional energy deposits with WALDO allows to reduce the uncertainty around muon energies. Second, confidence sets constructed with WALDO are even shorter than the corresponding prediction intervals, while also guaranteeing conditional coverage.

5 DISCUSSION

We presented WALDO, a novel method to construct confidence sets with correct finite-n conditional coverage by leveraging prediction algorithms and posterior estimators for inverse problems. WALDO relies on a regression-based Neyman construction, which requires orders of magnitude fewer simulations than traditional Monte Carlo approaches to be well calibrated across the parameter space (see Section 3.4). Nonetheless, our method still needs a simulator that is both high-fidelity – to draw inferences that reflect the true data-generating process – and fast – to simulate sufficiently large training sets to accurately learn the key quantities of WALDO: the test statistics, the critical values, and the coverage diagnostics, as discussed in Section 3.4. WALDO disentangles the *coverage* guarantees of the confidence region from the choice of the prior distribution. To increase *power*, one may be able to leverage domain-specific knowledge (see Sections 3.3 and 4.1), or take advantage of the internal structure of the simulator (Brehmer et al., 2020),

with the guarantee that the confidence sets always contain the true parameter with the desired proability. One could also adaptively simulate more data in specific regions of interest in the parameter space. Active learning strategies, and a more formal treatment of the relation between power and priors, are promising areas for future studies.

Domain sciences, especially the physical sciences, routinely seek to constrain parameters of interest using both theoretical (or simulation) models and experimental data. WALDO provides reliable constraints that can be used to deduce trustworthy scientific conclusions when other uncertainty quantification methods are either unavailable, unreliable or inefficient.

Acknowledgments

References

Georges Aad, Tatevik Abajyan, B Abbott, J Abdallah, S Abdel Khalek, Ahmed Ali Abdelalim, R Aben, B Abi, M Abolins, OS AbouZeid, et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1):1–29, 2012.

Sea Agostinelli, John Allison, K Amako, John Apostolakis, H Araujo, Pedro Arce, Makoto Asai, D Axen, Swagato Banerjee, G Barrand, et al. GEANT4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research A*, 506(3):250–303, 2003.

J-E Augustin, Adam M Boyarski, Martin Breidenbach, F Bulos, JT Dakin, GJ Feldman, GE Fischer, D Fryberger, G Hanson, B Jean-Marie, et al. Discovery of a narrow resonance in e+ e- annihilation. *Physical Review Letters*, 33(23):1406, 1974.

M. J. Bayarri and J. O. Berger. The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004. doi: 10.1214/088342304000000116.

James Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006. doi: 10.1214/06-BA115.

Rongmon Bordoloi, Simon J Lilly, and Adam Amara. Photo-z performance for precision cosmology. *Monthly Notices of the Royal Astronomical Society*, 406(2):881–895, 2010.

Denis Boyda, Gurtej Kanwar, Sébastien Racanière, Danilo Jimenez Rezende, Michael S Albergo, Kyle Cranmer, Daniel C Hackett, and Phiala E Shanahan. Sampling using su (n) gauge equivariant flows. *Physical Review D*, 103(7):074504, 2021.

Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10):5242–5249, 2020. doi: 10.1073/pnas.1915980117.

The CDF Collaboration. Observation of top quark production in Pbar-P collisions. *arXiv preprint hep-ex/9503002*, 1995.

MT Chao. The asymptotic behavior of Bayes' estimators. *The Annals of Mathematical Statistics*, 41(2):601–608, 1970.

Serguei Chatrchyan, Vardan Khachatryan, Albert M Sirunyan, Armen Tumasyan, Wolfgang Adam, Ernest Aguilo, Thomas Bergauer, M Dragicevic, J Erö, C Fabjan, et al. Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc. *Physics Letters B*, 716(1):30–61, 2012.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Grégoire Clarté, Christian P Robert, Robin J Ryder, and Julien Stoehr. Componentwise approximate Bayesian computation via Gibbs-like steps. *Biometrika*, 108(3):591–607, 2021.

Robert D. Cousins. Lectures on statistics in theory: Prelude to statistics in practice, 2018. URL https://arxiv.org/abs/1807.05996.

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, Luca Masserano, David Zhao, Rafael Izbicki, and Ann B Lee. Likelihood-Free Frequentist Inference: Confidence Sets with Correct Conditional Coverage. *arXiv preprint arXiv:2107.03920*, 2021.

Gauri Sankar Datta and Trevor J. Sweeting. Probability matching priors. In D.K. Dey and C.R. Rao, editors, *Bayesian Thinking*, volume 25 of *Handbook of Statistics*,

pages 91–114. Elsevier, 2005. doi: https://doi.org/10.1016/S0169-7161(05)25003-4.

Biprateep Dey, David Zhao, Jeffrey A Newman, Brett H Andrews, Rafael Izbicki, and Ann B Lee. Calibrated predictive distributions via diagnostics for conditional coverage. *arXiv preprint arXiv:2205.14568*, 2022.

Tommaso Dorigo, Sofia Guglielmini, Jan Kieseler, Lukas Layer, and Giles C Strong. Deep regression of muon energy with a k-nearest neighbor algorithm. *arXiv preprint arXiv:2203.02841*, 2022.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. nflows: Normalizing flows in PyTorch, November 2020. URL https://doi.org/10.5281/zenodo.4296287.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

Florian Gerber and Douglas Nychka. Fast covariance parameter estimation of spatial Gaussian process models using neural networks. *Stat*, 10(1):e382, 2021.

JK Ghosh. Higher order asymptotics for the likelihood ratio, rao's and wald's tests. *Statistics & probability letters*, 12 (6):505–509, 1991.

JK Ghosh and RV Ramamoorthi. Preliminaries and the finite dimensional case. *Bayesian Nonparametrics*, pages 9–55, 2003.

JK Ghosh, BK Sinha, and SN Joshi. Expansions for posterior probability and integrated Bayes risk. *Statistical Decision Theory and Related Topics III*, 1:403–456, 1982.

SW Herb, DC Hom, LM Lederman, JC Sens, HD Snyder, JK Yoh, JA Appel, BC Brown, CN Brown, WR Innes, et al. Observation of a dimuon resonance at 9.5 GeV in 400-GeV proton-nucleus collisions. *Physical Review Letters*, 39(5):252, 1977.

Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, and Gilles Louppe. Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*, 2021.

Matthew Ho, Markus Michael Rau, Michelle Ntampaka, Arya Farahi, Hy Trac, and Barnabás Póczos. A robust and efficient deep learning method for dynamical mass

measurements of galaxy clusters. *The Astrophysical Journal*, 887(1):25, 2019.

Robert E. Kass and Larry Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996. doi: 10.1080/01621459.1996.10477003.

Matthäus Kiel, Christopher W O'Dell, Brendan Fisher, Annmarie Eldering, Ray Nassar, Cameron G MacDonald, and Paul O Wennberg. How bias correction goes wrong: Measurement of X_{CO_2} affected by erroneous surface pressure estimates. *Atmospheric Measurement Techniques*, 12(4):2241–2259, 2019.

Jan Kieseler, Giles Chatham Strong, Filippo Chiandotto, Tommaso Dorigo, and Lukas Layer. Preprocessed dataset for "Calorimetric measurement of multi-TeV muons via deep regression", August 2021. URL https://doi.org/10.5281/zenodo.5163817.

Jan Kieseler, Giles C Strong, Filippo Chiandotto, Tommaso Dorigo, and Lukas Layer. Calorimetric measurement of multi-TeV muons via deep regression. *The European Physical Journal C*, 82(1):1–26, 2022.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 2005.

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

Yong Li, Jun Yu, and Tao Zeng. Deviance information criterion for latent variable models and misspecified models. *Journal of Econometrics*, 216(2):450–493, 2020.

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.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of Machine Learning Research*, 7 (6), 2006.

Siddharth Mishra-Sharma and Kyle Cranmer. Neural simulation-based inference approach for characterizing the galactic center γ-ray excess. *Physical Review D*, 105 (6):063017, 2022.

Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937.

Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. Conformal prediction with neural networks. In *19th IEEE*

International Conference on Tools with Artificial Intelligence (ICTAI 2007), volume 2, pages 388–395. IEEE, 2007.

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.

Pratik Patil, Mikael Kuusela, and Jonathan Hobbs. Objective frequentist uncertainty quantification for atmospheric CO_2 retrievals. *SIAM/ASA Journal on Uncertainty Quantification*, 10(3):827–859, 2022. doi: 10.1137/20M1356403.

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.

Catia Scricciolo. Probability matching priors: A review. *Journal of the Italian Statistical Society*, 8:83–100, 1999. doi: 10.1007/BF03178943.

Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. 2010.

Umberto Simola, Jessi Cisewski-Kehe, Michael U Gutmann, and Jukka Corander. Adaptive Approximate Bayesian Computation tolerance selection. *Bayesian analysis*, 16 (2):397–423, 2021.

Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J. Gonçalves, David S. Greenberg, and Jakob H. Macke. `sbi`: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020. doi: 10.21105/joss.02505. URL `https://doi.org/10.21105/joss.02505`.

Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate Bayesian Computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482, 1943.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 03 1938. doi: 10.1214/aoms/1177732360.

A THEORETICAL RESULTS

We assume that the quantile regression estimator described in Section 3 is consistent in the following sense:

Assumption 1 (Uniform consistency) *Let $F(\cdot|\boldsymbol{\theta})$ be the cumulative distribution function of the test statistic $\tau(\mathcal{D};\boldsymbol{\theta}_0)$ conditional on $\boldsymbol{\theta}$, where $\mathcal{D} \sim F_{\boldsymbol{\theta}}$. Let $\widehat{F}_{B'}(\cdot|\boldsymbol{\theta})$ be the estimated conditional distribution function, implied by a quantile regression with a sample $\mathcal{T}'$ of B' simulations $\mathcal{D} \sim F_{\boldsymbol{\theta}}$. Assume that the quantile regression estimator is such that*

$$\sup_{\tau \in \mathbb{R}} |\widehat{F}_{B'}(\tau|\boldsymbol{\theta}_0) - F(\tau|\boldsymbol{\theta}_0)| \xrightarrow[B' \longrightarrow \infty]{\mathbb{P}} 0.$$

Assumption 1 holds, for instance, for quantile regression forests (Meinshausen and Ridgeway, 2006). Next, we show that step **(ii)** in Section 3.2 yields a valid hypothesis test as $B' \to \infty$.

Theorem 1 *Let $C_{B'} \in \mathbb{R}$ be the critical value of the test based on a strictly continuous statistic $\tau(\mathcal{D};\boldsymbol{\theta}_0)$ chosen according to step **(ii)** for a fixed $\alpha \in (0,1)$. If the quantile estimator satisfies Assumption 1, then,*

$$\mathbb{P}_{\mathcal{D}|\boldsymbol{\theta}_0,C_{B'}}(\tau(\mathcal{D};\boldsymbol{\theta}_0) \geq C_{B'}) \xrightarrow[B' \longrightarrow \infty]{a.s.} \alpha,$$

where $\mathbb{P}_{\mathcal{D}|\boldsymbol{\theta}_0,C_{B'}}$ denotes the probability integrated over $\mathcal{D} \sim F_{\boldsymbol{\theta}_0}$ and conditional on the random variable $C_{B'}$.

If the convergence rate of the quantile regression estimator is known (Assumption 2), Theorem 2 provides a finite-B' guarantee on how far the Type-I error of the test will be from the nominal level.

Assumption 2 (Convergence rate of the quantile regression estimator) *Using the notation of Assumption 1, assume that the quantile regression estimator is such that*

$$\sup_{\tau \in \mathbb{R}} |\widehat{F}_{B'}(\tau|\boldsymbol{\theta}_0) - F(\tau|\boldsymbol{\theta}_0)| = \mathcal{O}_p\left(\left(\frac{1}{B'}\right)^r\right)$$

for some $r > 0$.

Theorem 2 *With the notation and assumptions of Theorem 1, and if Assumption 2 also holds, then,*

$$|\mathbb{P}_{\mathcal{D}|\boldsymbol{\theta}_0,C_{B'}}(\tau(\mathcal{D};\boldsymbol{\theta}_0) \geq C_{B'}) - \alpha| = \mathcal{O}_p\left(\left(\frac{1}{B'}\right)^r\right).$$

Proofs of these results can be found in Dalmasso et al. (2021).

B ADDITIONAL EXPERIMENTS

B.1 PROPERTY III: Estimating the Conditional Variance Matters

We complete the exposition of the statistical properties of WALDO (Section 3.3) by demonstrating the importance of estimating the conditional variance in the test statistic τ^{WALDO}. Recall that in principle any test statistic defined in an LFI setting could be used for our framework. One could then define a simpler "unstandardized" test statistic $\tau^{\text{WALDO-NOVAR}}(\mathcal{D};\boldsymbol{\theta}_0) = (\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^T(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)$ which does not require estimation of $\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]$. It turns out that estimating $\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]$ and using τ^{WALDO} is actually of crucial importance, as it leads to confidence regions of smaller or equal expected size, especially in settings where the conditional variance varies significantly as a function of $\boldsymbol{\theta}$. Consider, for example, the problem of estimating the shape of a Pareto distribution with fixed scale $x_{\min} = 1$ and true unknown shape $\theta^* = 5$, which yields a strongly right-skewed data distribution. Figure 7 shows that τ^{WALDO} has much higher power than $\tau^{\text{WALDO-NOVAR}}$ for inferring θ. Dividing by the conditional variance effectively stabilizes the test statistic and makes its distribution over $\mathcal{D}$ pivotal, i.e., independent of θ. This implies that the critical values will be relatively constant over θ (see top right panel for WALDO), which yields tighter parameter regions due to the curvature of the test statistic.
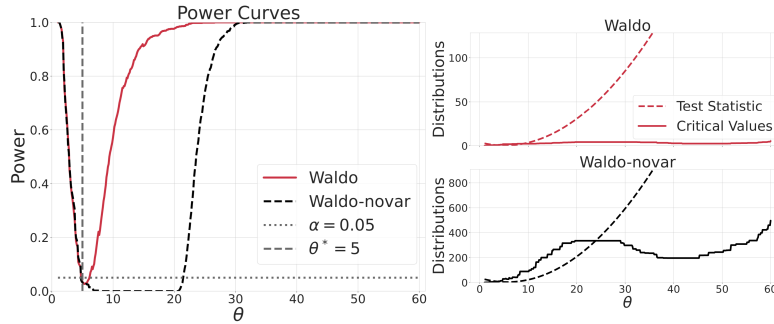
Figure 7: **PROPERTY III: Estimating the conditional variance matters.** *Left:* Power curves at 95% confidence level when the true Pareto shape $\theta^* = 5$, implying a very skewed data distribution. *Right:* Test statistics and critical values as a function of θ. ($n = 10$).

B.2 Confidence Sets from Neural Posteriors: Two-Dimensional Gaussian Mixture

The results of Figure 5 in the main text showed that WALDO is able to leverage an estimated posterior to construct conditionally valid confidence regions, even when the prior is at odds with the data. On the other side, when no prior information is available, it is common to sample θ according to a uniform distribution over the parameter space. In this case, we observe that confidence sets and posterior credible regions largely overlap. Nonetheless, if the latter happen to suffer from approximation errors, as is common for neural posteriors in high dimensions, this could hinder the statistical reliability of the estimated region. WALDO can correct even for this problem and guarantee conditional coverage, as we can see from panel *a)* in Figure 8.

Figure 9 shows the output of the diagnostics procedure when using a uniform prior to train the posterior estimator (compare with Figure 5, right column, in the main text, which used a Gaussian prior). We achieve correct conditional coverage for WALDO but not for credible regions even though the prior is is uniform, due to estimation and approximation errors in the posterior, which WALDO can correct using quantile regression to calibrate the test statistics.

B.3 Confidence Sets for Muon Energies using CNN Predictions

Figure 10 compares confidence sets and prediction sets for the full calorimeter data, showing clearly the bias in the prediction sets and the correction applied by Waldo. These results explain the observed patterns in Figure 6 in the main text: prediction sets are centered around the point prediction, which is downward biased at high energies, mainly due to the nonlinearity of the response at high energies.
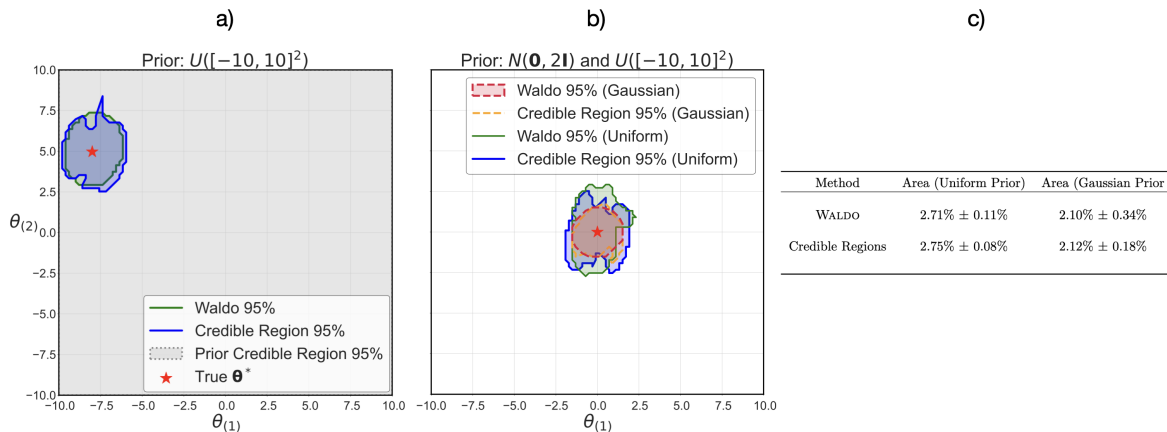

Figure 8: **a) When the prior is uninformative, WALDO can still correct for possible approximation errors in the estimated posterior. b)-c) When the prior is consistent with the data, WALDO tightens the confidence sets, improving the precision with respect to the case using a Uniform prior.** *a)-b):* Posterior credible regions and WALDO confidence sets using different priors. *Right:* Average area of credible regions and WALDO confidence sets across 100 independent samples, reported as the percentage of points retained among those in the evaluation grid.
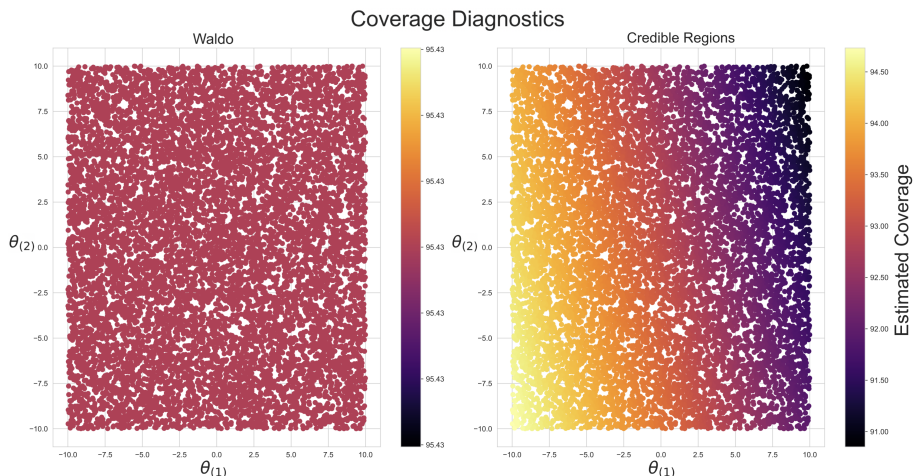
Figure 9: **Coverage diagnostics for Gaussian mixture model example with uniform prior**. We achieve correct conditional coverage for WALDO (left figure) but not for credible regions (right figure) even though the prior is is uniform, due to estimation and approximation errors, which WALDO can correct via recalibration.

C DETAILS ON MODELS, TRAINING, AND COMPUTATIONAL RESOURCES

C.1 Synthetic Examples for Statistical Properties

See Section 3.3 in the main text and Appendix B.1 for descriptions of the experiments. For PROPERTY I and PROPERTY II, we used the implementation of local linear regression available in Seabold and Perktold (2010) to estimate conditional mean and conditional variance within a prediction setting, with $B = 20,000$. For PROPERTY III, instead, we used a simple neural network with one hidden layer and $B = 50,000$. In all cases, for quantile regression we used quantile Gradient Boosting Pedregosa et al. (2011), with $B' = 20,000$ for PROPERTY I and PROPERTY II, and $B' = 50,000$ for PROPERTY III. All models were trained on a MacBook Pro M1Pro (CPU only).
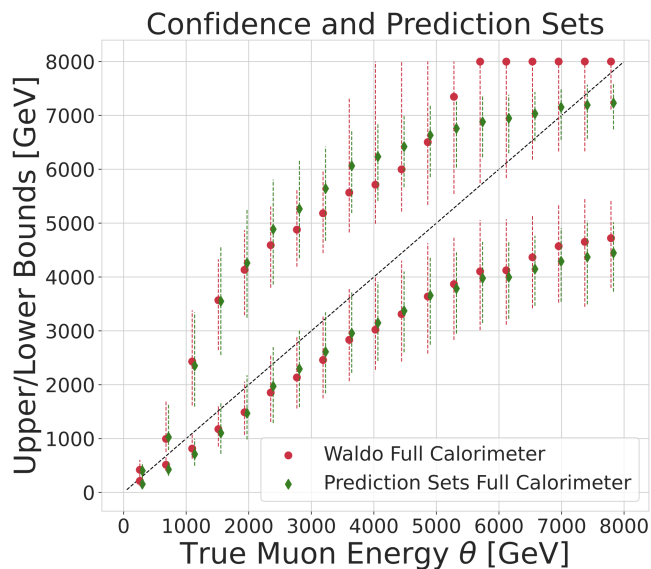

Figure 10: **Confidence and prediction sets for the muon energy reconstruction experiment**. Boxplots of the upper and lower bounds of prediction sets (green) versus WALDO confidence sets (red) for full the calorimeter data, all divided in 19 bins over true energy. We clearly see the bias occurring in the prediction sets (especially at high energies) and the correction applied by WALDO.

C.2   Synthetic Example for Computational Properties

See Section 3.4 in the main text for a description of the experiment. To compute the test statistic τ^{WALDO}, we approximated conditional mean and conditional variance through a posterior distribution estimated via normalizing flows (Tejero-Cantero et al., 2020), with $B = 20,000$ for $p = 1$ and $B = 200,000$ for $p = 10$. To construct the confidence sets, critical values were then estimated both via quantile regression using quantile Gradient Boosting (Pedregosa et al., 2011) with varying values of B', and via Monte Carlo by simulating many times for each $\boldsymbol{\theta}$ and retaining the $(1 - \alpha)$ quantile of the computed test statistics. The evaluation set was made of $1,000$ samples over $\boldsymbol{\Theta} = [-1, 1]^p$. To make the comparison fair, if quantile regression used $B' = 50,000$, then Monte Carlo had access to 50 simulations for each of the $1,000$ samples in the evaluation set. The estimated coverage probability for both methods was then estimated using the implementation of Generalized Additive Models (GAMs) with thin plate splines available in the `MGCV` package of `R`, with $B'' = 30,000$.

C.3   Confidence Sets from Neural Posteriors: Two-Dimensional Gaussian Mixture

See Section 4.1 in the main text and Appendix B.2 for descriptions of the experiments and details on the algorithms and sample sizes used. Training was done on a MacBook Pro M1Pro (CPU only); it took approximately 15–20 minutes to train the posterior estimator, and an additional $\sim$2 minutes for the quantile neural network to estimate the critical values. Note that the latter step requires computing the conditional mean, the conditional variance and the Waldo statistic over all sample points in $\mathcal{T}'$. The posterior was sampled multiple times for each $\mathbf{x} \in \mathcal{T}'$ to approximate $\mathbb{E}(\boldsymbol{\theta}|\mathbf{x})$ and $\mathbb{V}(\boldsymbol{\theta}|\mathbf{x})$ via Monte Carlo; this procedure took a total of $\sim$45 minutes (but could potentially be optimized through vectorizations in the future).

C.4   Confidence Sets for Muon Energies using CNN Predictions

See Section 4.2 and Appendix B.3 for descriptions of the experiment and details on the algorithms and sample sizes used. We had access to 886,716 simulated muons in total; roughly 200,000 muons were used to estimate the critical values, $\sim$24,000 muons to construct the final confidence sets and diagnostics, and the rest was used to estimate the conditional mean and variance via the custom 3D CNN from Kieseler et al. (2022). Training the latter CNN took approximately 20 hours for the conditional mean and another 20 hours for the conditional variance, using an NVIDIA V100 GPU on an Azure cloud computing machine. Estimating the critical values via quantile gradient boosted trees took approximately 2 minutes.