Simulation-Based Inference with WALDO: Perfectly Calibrated Confidence Regions Using Any Prediction or Posterior Estimation Algorithm

Luca Masserano  
Department of Statistics and Data Science  
Carnegie Mellon University  
lmassera@andrew.cmu.edu

Tommaso Dorigo  
INFN  
Sezione di Padova  
tommaso.dorigo@cern.ch

Rafael Izbicki  
Department of Statistics  
Federal University of São Carlos  
rafaelizbicki@gmail.com

Mikael Kuusela  
Department of Statistics and Data Science  
Carnegie Mellon University  
mkuusela@andrew.cmu.edu

Ann B. Lee  
Department of Statistics and Data Science  
Carnegie Mellon University  
annlee@stat.cmu.edu

Abstract

The vast majority of modern machine learning targets prediction problems, with algorithms such as Deep Neural Networks revolutionizing the accuracy of point predictions for high-dimensional complex data. Predictive approaches are now used in many domain sciences to directly estimate internal parameters of interest in theoretical simulator-based models. In parallel, common alternatives focus on estimating the full posterior using modern neural density estimators such as normalizing flows. However, an open problem in simulation-based inference (SBI) is how to construct properly calibrated confidence regions for internal parameters with nominal conditional coverage and high power. Many SBI methods are indeed known to produce overly confident posterior approximations, yielding misleading uncertainty estimates. Similarly, existing approaches for uncertainty quantification in deep learning provide no guarantees on conditional coverage. In this work, we present WALDO, a novel method for constructing correctly calibrated confidence regions in SBI. WALDO reframes the well-known Wald test and uses Neyman inversion to convert point predictions and posteriors from any prediction or posterior estimation algorithm to confidence sets with correct conditional coverage, even for finite sample sizes. As a concrete example, we demonstrate how a recently proposed deep learning prediction approach for particle energies in high-energy physics can be recalibrated using WALDO to produce confidence intervals with correct coverage and high power.

1 Introduction

The vast majority of modern machine learning targets prediction problems, with algorithms such as Deep Neural Networks (DNN) being particularly successful with point predictions of a target variable $Y \in \mathbb{R}$ given complex feature vectors or image data $x \in X$. In many science applications,
however, one is more interested in uncertainty quantification (UQ) than in point estimation per se.

Domain sciences, particularly the physical sciences, often seek to constrain parameters of interest using theoretical (or simulation) models together with observed (experimental) data. For example, a primary goal in cosmological analysis is to increase the discovery potential of new physics by better constraining parameters of the Λ-CDM model in Big Bang cosmology, using higher-resolution simulation models and more precise survey data [2]. In particle physics, a central goal of collider experiments is to constrain parameters of theoretical models of particle interactions [8]. Climate scientists, on the other hand, are interested in constraining uncertain climate model parameters using atmospheric observations; e.g., [30] uses aircraft, ship, and ground observations to constrain parameters describing the impact of aerosols on Earth’s radiative balance. The above-mentioned inferential tasks are all inverse problems, meaning that the parameters of interest are not directly observed but are the causes of the observed data x. We will henceforth denote internal parameters by θ to distinguish the problem of using x to infer an internal fixed parameter θ ∈ Θ from the “predictive problem” of using x to predict an observable random variable Y. UQ for inverse problems entails constructing confidence regions (rather than prediction regions) for θ.

Simulations can be used to predict how systems will behave in a variety of circumstances, and can help to constrain parameters when Gaussian or other parametric likelihood models become questionable at the level of precision of modern scientific data. Stochastic simulators are able to produce observable data x under different parameter settings, but they encode the likelihood function $L(θ; x) := p(x|θ)$ only implicitly. Statistical inference when $L(θ; x)$ is intractable is often dubbed likelihood-free inference (LFI), and is a form of simulation-based inference (SBI). LFI has undergone a revolution in terms of the complexity of problems that can be tackled (see [15] for a recent review). Many LFI methods are, however, known to be overly confident, meaning that they yield confidence sets with empirical coverage smaller than the desired nominal coverage [27], hence leading to potentially misleading results. In parallel to LFI methods, DNNs (such as convolutional neural networks [35]) are now used in many domain sciences to directly estimate internal parameters of interest in statistical models (e.g., [34, 22, 36, 28]). DNNs are often easier to train than neural density estimators (such as normalizing flows), but the problem of producing reliable uncertainty estimates of internal parameters from DNN point predictions has not yet found a solution.

Let $D := (x_1, \ldots, x_n)^T$ denote observable data, where the “sample size” n refers to the number of observations from a fixed configuration of the parameters θ. The goal of this work is to present a practical LFI procedure that can leverage any high-capacity prediction algorithm or neural posterior estimator to construct confidence regions for θ with correct conditional coverage, that is

$$P(θ ∈ R(D)|θ) = 1 − α, \forall θ ∈ Θ,$$

where $α ∈ (0, 1)$ is a prespecified miscoverage level. Correct conditional coverage implies correct marginal coverage, $P(θ ∈ R(D)) = 1 − α$, 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 Θ. The concept of conditional coverage is also more widely applicable than marginal coverage: the latter only applies when θ is treated as a random variable, while the former also applies when θ is a non-random fixed parameter, as in all the above-mentioned scientific applications. In addition to calibration, we want to construct confidence regions that are informative and can constrain parameters as much as possible; that is, with high statistical power. Finally, for our method to be useful in practice, we need diagnostics to verify whether we indeed achieve nominal conditional coverage across the entire parameter space Θ. Standard diagnostic tools for SBI [53] assess marginal coverage only, and cannot easily identify parameter values for which confidence sets are unreliable.

To address these problems, we introduce WALDO, a novel method to construct correctly calibrated confidence regions in an LFI setting. WALDO reframes the Wald test [57] and uses the Neyman construction [53] to convert point predictions and posterior distributions from any prediction algorithm or posterior estimator into confidence regions with correct conditional coverage. It does so by exploiting estimates of the conditional mean $E[θ|D]$ and conditional variance $V[θ|D]$, but it is indifferent to the source of these quantities. WALDO stems from a recent likelihood-free frequentist inference (LFI) framework proposed in [16], which showed it to construct valid confidence regions using likelihood estimates in a SBI setting. However, the authors also demonstrated that the statistical power of the resulting sets could easily degrade due to numerical approximations when using a likelihood-based test statistic. By recalibrating predictions or posteriors, WALDO is able to leverage recent developments in AI, which allow to effectively handle complex high-dimensional data.
Figure 1: Schematic diagram of WALDO. Left: To estimate the conditional mean $\mathbb{E}[\theta|D]$ and variance $\mathbb{V}[\theta|D]$ and construct the test statistic $\hat{\tau}_{WALDO}$, we use a first train set $T$ and a prediction algorithm (e.g., DNN) or posterior estimator (e.g., normalizing flow). Center: Then, we simultaneously estimate critical values $C_{\theta_0,\alpha}$ for all parameter values $\theta_0 \in \Theta$, via quantile regression on a second train set $T'$. Bottom: For observed data $D$, the two pieces are combined to construct a confidence region via Neyman inversion. Right: We compute coverage diagnostics for constructed confidence regions of any parameter $\theta \in \Theta$ via spline regression on a third train set $T''$. See Section 3.2 for details.

data. Figure 1 summarizes the different components of WALDO, and the full procedure is detailed in Section 3.2. This approach 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 correctly specified, and (ii) are guaranteed to have the nominal conditional coverage even in finite samples, regardless of the correctness of the prior. WALDO is amortized, meaning that once the procedure has been trained it can be evaluated on any number of observations. We lay out the statistical properties of WALDO, providing synthetic examples to support our claims. Finally, we show the effectiveness of this method with a real-world application in particle physics: inferring the energy of muons at a future particle collider using deep convolutional neural networks for high-dimensional 3D images from a particle detector. The results we obtain for this problem (see Figure 6) are of scientific interest by themselves, as a rigorous and powerful estimate of the uncertainty around observed muon energies is essential in the search of new physics. The code used to produce our experiments is available on Github.

Notation: We refer to parameters of interest as $\theta \in \Theta \subset \mathbb{R}^p$ and to input data as $D = (x_1, \ldots, x_n)^T$, with $x_i \in \mathcal{X} \subset \mathbb{R}^d$ and possibly $p \neq d$. Throughout the paper, $n$ denotes the size of a sample of dimension $d$ from a specific configuration of the parameters $\theta$, and is distinct from $B, B'$ and $B''$, i.e., the number of simulations in each of the datasets $T, T'$ and $T''$ required at different steps of our method. In addition, let $F_\theta$ represent the stochastic simulator that encodes the likelihood function of $\theta$. We distinguish between observable data and actual observations by denoting the latter as $D$. The true value of a parameter is always denoted as $\theta^*$. Finally, we refer to confidence regions as $\mathcal{R}(D)$. The terms “set”, “region” and (when $p = 1$) “interval” are used interchangeably.

2 Relation To Other Work

Existing approaches for uncertainty quantification in SBI (see [15] for a recent review) are either based on Bayesian posterior estimation [6, 41, 45, 38, 25, 11, 29] or on inversion of likelihood ratio tests [9, 8, 16]. WALDO differs from the former by aiming for conditional coverage instead of marginal coverage, and from the latter by taking advantage of the power of predictive ML algorithms. Similarly, existing approaches for deep learning uncertainty quantification (see [21] for a recent
review), such as Monte Carlo drop out [20] and conformal inference [44], construct prediction sets instead of confidence sets. As such, these approaches, again, control marginal coverage instead of conditional coverage. 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.

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. Examples of such corrections range from particle physics [18] to cosmology [28] and remote sensing [32]. Usually the goal of these corrections is to reduce the impact of the prior specification, but in contrast to WALDO, these approaches do not provide formal coverage guarantees.

Posterior inferences do not control conditional coverage even for correctly specified priors [47]. WALDO addresses this by recalibrating the posterior using Neyman inversion. An alternative approach is to choose the prior so that the conditional coverage of posterior inferences is improved [5, 7, 31, 50, 17]. However, these modified prior distributions need to be created using properties of the likelihood instead of actual prior information, a limitation that is not present in WALDO.

3 WALDO

We discuss the foundational tools needed for WALDO, describe the method in detail, and outline its statistical properties, while offering supporting numerical examples.

3.1 Foundations

Neyman construction A key ingredient of WALDO is the equivalence between hypothesis tests and confidence sets, which was formalized by Neyman in 1937 [43]. The idea is to invert a series of two-sided size $\alpha$ hypothesis tests of the form

$$H_0: \theta = \theta_0 \quad \text{vs.} \quad H_1: \theta \neq \theta_0,$$

for all $\theta_0 \in \Theta$. After observing a sample $D$, we then define a confidence region $R(D)$ by taking all $\theta_0$’s that were not rejected by the test above. By construction, the set $R(D)$ satisfies Equation (1), i.e., has the correct conditional coverage. Although powerful and intuitive, the Neyman construction is hard to use in practice because it requires estimating critical values $C_{\theta_0, \alpha}$ that define the size $\alpha$ acceptance region for each hypothesis we invert. In the context of LFI, this is usually done by resorting to Monte Carlo or bootstrap approaches [40, 56], which quickly become computationally prohibitive as the dimensionality of the parameter space increases.

The Wald test Since any test that controls type I error at level $\alpha$ may be used, we can resort to the one introduced by Wald [57], which leads to a uniformly most powerful test in many settings. It measures the agreement of the data with the null hypothesis for $\theta$ using the squared distance between an unrestricted estimator $\hat{\theta}$ of $\theta$ and the hypothesized value $\theta_0$. In the one-dimensional case, it is based on the following test statistic:

$$\tau_{\text{Wald}}(D; \theta_0) := \left( \frac{\hat{\theta}_{\text{MLE}} - \theta_0}{\sqrt{V(\hat{\theta}_{\text{MLE}})}} \right)^2,$$

where $\hat{\theta}_{\text{MLE}}$ is the maximum-likelihood estimator of $\theta$ and $V(\hat{\theta}_{\text{MLE}})$ can be any consistent estimator of its variance. In our setting, where we do not have access to the likelihood and we cannot resort to assumptions on the distribution of $\tau_{\text{Wald}}(D; \theta_0)$ nor to asymptotic regimes, it becomes hard not only to estimate $\hat{\theta}_{\text{MLE}}$ and its variance effectively, but also to estimate critical values $\tilde{C}_{\theta, \alpha}$.

3.2 Methodology

From Wald to WALDO: Recalibrating posteriors and point predictions. WALDO replaces $\hat{\theta}_{\text{MLE}}$ and its variance in the Wald statistic with the typical output of standard LFI or prediction algorithms. We define the WALDO test statistic for parameters of arbitrary dimensionality $p$ as

$$\tau_{\text{WALDO}}(D; \theta_0) = (E[\theta | D] - \theta_0)^T V[\theta | D]^{-1} (E[\theta | D] - \theta_0),$$

where $E[\theta | D]$ is the posterior mean of $\theta$ given the data $D$, and $V[\theta | D]$ is the posterior covariance matrix.
where \( E[\theta | D] \) and \( V[\theta | D] \) are the conditional mean and covariance matrix of \( \theta \) given the data \( D \), respectively. Existing literature on the asymptotic behaviour of Bayes estimators (e.g., [10, 23, 24, 27]) has shown convergence results which may explain the large-sample equivalence of WALDO and Wald observed empirically. Specifically, it has been proved that

\[
\left( E[\theta | D] - \hat{\theta} \right) = o_p(n^{-1/2}) \quad \text{and} \quad \left( V[\theta | D] - \frac{1}{n} H^{-1}(\hat{\theta}) \right) = o_p(n^{-1}),
\]

where \( H^{-1}(\hat{\theta}) \) is the negative inverse Fisher information matrix evaluated at \( \hat{\theta} \). In this case, WALDO would also enjoy the asymptotic properties typical of the Wald test, which would also make it a pivotal test statistic.

WALDO expands on the framework formalized in [16], which consists of a modular procedure to (i) estimate a likelihood-based test statistic via odds ratios, (ii) estimate critical values \( C_{\theta_0, \alpha} \) across the parameter space via quantile regression, and (iii) check that the constructed confidence sets achieve the desired coverage level for all \( \theta \in \Theta \). Here, we replace (i) and instead use posteriors or point predictions to compute \( \tau_{WALDO} \) in [4]. We now describe in detail the entire inferential machinery that is depicted in Figure [4]. Assuming we have access to a simulator \( F_\theta \) that can produce high-fidelity observable data \( D \) from a parameter configuration \( \theta \), we break down the construction of a confidence set (including diagnostics) in three steps:

(i) **Neural density estimators or prediction algorithms for the test statistic.** By simulating a train set \( T = \{(\theta^{(j)}, D^{(j)})\}_{j=1}^B \), where each \( \theta^{(j)} \) is itself drawn from a prior \( \pi_\theta \), we have two ways of estimating \( E[\theta | D] \) and \( V[\theta | D] \), which together define \( \tau_{WALDO} \). By leveraging a modern neural density estimator, such as normalizing flows [46], we can compute a posterior distribution from which we can efficiently draw a large number of samples [4]. The posterior mean and covariance matrix are then estimated via Monte Carlo sampling. Alternatively, we can leverage the fact that prediction algorithms output an estimate of the conditional mean of \( \theta \) given \( D \), when minimizing the squared error loss. More specifically, we can obtain \( E[\theta | D] \) by directly predicting \( \theta \) from \( D \), and \( V[\theta | D] \) by predicting \( (\theta - E[\theta | D])^2 \) from \( D \) again, as \( E[(\theta - E[\theta | D])^2 | D] = V(\theta | D) \).

(ii) **Quantile regression for critical values.** Once we have the building blocks of the test statistic from step (i), we can proceed to estimate critical values for \( \tau_{WALDO} \). We simulate a second train set \( T' = \{(\theta^{(j)}, D^{(j)})\}_{j=1}^{B'} \), drawing \( \theta \sim r_\theta \) uniformly over \( \Theta \) (\( r_\theta \equiv U(\Theta) \)) to allow calibration across the entire parameter space. By evaluating \( \tau_{WALDO} \) over each element of \( T' \), we obtain pairs \( \left( \tau_{WALDO}(D^{(j)}; \theta^{(j)}), \hat{C}_{\theta_0, \alpha}^{(j)} \right) \) that we can use to train a quantile regressor and estimate critical values \( \hat{C}_{\theta_0, \alpha} \) over a fine grid \( \theta_0 \in \Theta \). This procedure lets us test hypotheses of the form defined in Equation [4] at the desired level \( \alpha \).

(iii) **Putting it all together: Neyman inversion.** The final step is to construct the confidence set as Neyman outlined in [4]. Once \( D \) is observed, we evaluate \( \tau_{WALDO}(D; \theta_0) \) over a fine grid of \( \theta_0 \in \Theta \), and retain all \( \theta_0 \) for which the corresponding test does not reject the null. That is,

\[
R(D) = \{ \theta_0 \in \Theta : \tau_{WALDO}(D; \theta_0) \leq \hat{C}_{\theta_0, \alpha} \}. \quad (5)
\]

As we show in Appendix [X], step (ii) leads to valid size \( \alpha \) hypothesis tests as long as the quantile regression model is well estimated, which then implies that \( R(D) \) satisfies conditional coverage (Eq [1] at level \( 1-\alpha \), regardless of the true value of \( \theta \) and of the size \( n \) of the observable 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 introduced in [16]. For a third simulated train set \( T'' = \{(\theta^{(j)}, D^{(j)})\}_{j=1}^{B''} \), we construct a confidence region for each \( D^{(j)} \in T'' \) and then regress \( 1\{\theta^{(j)} \in R(D^{(j)})\} \) against \( \theta^{(j)} \) adopting a suitable regression method. By definition, this will estimate \( E[1\{\theta \in R(D)\} | \theta] = P(\theta \in R(D) | \theta) \) across the whole parameter space. Note that this technique can be used to check the empirical coverage of any uncertainty estimate, as illustrated in Section [4] for posterior credible regions and prediction sets.

Notes:
- Amortized neural density estimators are more suitable than their sequential counterparts, as we need to evaluate the posterior on more than one sample for steps (ii) and (iii).
- Note that the prediction approach is most suitable when \( p = 1 \) and \( n = 1 \), as standard algorithms are usually designed to handle pairs of data points of the form \( (\theta, x) \).
3.3 Statistical Properties

**PROPERTY I: WALDO guarantees coverage everywhere in Θ, regardless of the specified prior.**
In scientific inquiry, practitioners sometimes have domain-specific knowledge that could guide inference through the elicitation of a prior distribution over the parameters of interest. In practice, there is no way to verify whether this prior is correctly specified with respect to the data-generating process or not, a fact that could introduce a bias in the estimated posterior or point prediction. Ideally, we would want the statistical guarantees of any estimated parameter region to be preserved under this bias. Figure 2 shows that this is not the case even for a simple one-dimensional setting. Specifically, we assume $\theta \sim \mathcal{N}(0, 2)$, $D|\theta \sim \mathcal{N}(\theta, 1)$, with $p = 1$ and $n = 1$. We compute exact 95% confidence sets for the mean $\theta$ by inverting a series of Wald tests, the corresponding uncorrected 95% exact central prediction intervals around the conditional mean $\mathbb{E}[\theta|D] \pm 1.96\sigma$, and the fully estimated confidence sets with WALDO using $B = B' = 20,000$. The left panel clearly shows the bias of the uncorrected prediction sets increasing as we go further from the prior mean, which is then reflected in the poor coverage performance on the right panel; see [47] for further analysis of this effect. WALDO is able to recalibrate the uncertainty around the conditional mean and output a confidence set with the correct level of coverage across $\Theta$.

**PROPERTY III: Estimating the conditional variance matters.** Finally, 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 $\tau_{\text{WALDO-NOVAR}}(D; \theta_0) = (\mathbb{E}[\theta|D] - \theta_0)^T (\mathbb{E}[\theta|D] - \theta_0)$. It

---

*Here we have everything available in closed form, so we can calculate the analytical versions of the Wald test statistics and of prediction sets.*
turns out that estimating \( \mathbb{V}[\theta|D] \) and using \( \tau^{\text{ALDO}} \) is actually of crucial importance, as it leads to confidence regions of smaller or equal expected volume, especially in settings where the conditional variance varies significantly as a function of \( \theta \). Consider, for example, the problem of estimating the shape of a Pareto distribution with fixed scale \( x_{\text{min}} = 1 \) and true unknown shape \( \theta^* = 5 \), which yields a strongly right-skewed data distribution. Figure 4 shows that \( \tau^{\text{ALDO}} \) has much higher power than \( \tau^{\text{ALDO-NOVAR}} \) for inferring \( \theta \). Dividing by the conditional variance effectively stabilizes the test statistic and makes its distribution over \( D \) more pivotal, i.e., less dependent on \( \theta \). This implies that the critical values will be relatively constant over \( \theta \) (see top right panel for \text{ALDO})

![Figure 4: 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 \( \theta \). \( n = 10 \).](image)

4 Experimental Results

We assess the performance of \text{ALDO} on two challenging experiments which complement the synthetic examples of Section 3. In the first example (Section 4.1), we show how to recalibrate a posterior distribution estimated via normalizing flows, especially when prior information is available. The second example (Section 4.2) tackles a complex particle energy reconstruction problem in high-energy physics: we convert predictions from a custom deep learning architecture to confidence intervals with correct coverage and high power.

4.1 Gaussian Mixture Model: Recalibrating a Posterior Estimated via Normalizing Flows

This inference task was introduced in \cite{12} and has become a standard benchmark in the SBI literature \cite{55,51,39}. It consists of estimating the (common) mean of the components of a two-dimensional Gaussian mixture, where one component has much broader covariance than the other: \( D|\theta \sim \frac{1}{2}\mathcal{N}(\theta, 1) + \frac{1}{2}\mathcal{N}(\theta, 0.01 \odot 1) \), where \( \theta \in \mathbb{R}^2 \) and \( n = 10 \). We estimate \( p(\theta|D) \) using the implementation of masked autoregressive flows available in \cite{19} through the SBI software package of \cite{54}, and report results obtained with two different priors: \( \theta \sim \mathcal{U}([-10, 10]^2) \) and \( \theta \sim \mathcal{N}(0, 2 \odot 1) \). We estimated the critical values with a 3-layer neural network minimizing a quantile loss. Simulated data sets used for training are of the following sizes: \( B = 100,000 \), \( B' = 10,000 \) when using a uniform prior and \( B' = 30,000 \) when using a Gaussian prior. Conditional mean and variance were approximated with 50,000 Monte Carlo samples from the neural posterior. From the results, we conclude the following:

(i) Absent prior knowledge, \text{ALDO} corrects for estimation errors while retaining high power.

When no prior information is available, it is common to sample \( \theta \) according to a uniform distribution over the parameter space. In this case, confidence sets and posterior credible regions will largely overlap, but the latter might suffer from estimation bias and approximation errors that could hinder the statistical reliability of the estimated region. \text{ALDO} can correct even for this problem and guarantee conditional coverage, as we can see from the top left plot in the left panel of Figure 5.

(ii) \text{ALDO} guarantees coverage everywhere even if the prior is misspecified.

Even when domain knowledge is available, there is no way in practice to check for misspecifications with respect to the truth. If the observed data happens to derive from a true parameter that is far from the bulk of the prior distribution, then the posterior will try to use information in the data to construct a credible region consistent with the misspecified prior. As we can see from the two right plots in the left panel

\footnote{While \text{ALDO} works for a sample made of any number of observations, we had to use \( n = 1 \) because the SBI Python library we used to estimate the posterior does not yet support larger sample sizes.}
of Figure 5, WALDO can correct for the bias in the estimated posterior. The right panel shows the output of the diagnostic procedure described in Section 3.2 while credible regions approximate the desired level of conditional coverage only where the prior puts non-negligible mass, WALDO yields confidence regions that are guaranteed to cover the true parameter everywhere with probability $1 - \alpha$.

(iii) When the prior is correctly specified, WALDO further constrains parameters of interest. Finally, we highlight that when prior knowledge is accurate, WALDO does not trade off coverage for power and yields confidence regions of smaller expected size with respect to those estimated using a non-informative prior, as we can see from the bottom left plot in the left panel of Figure 5.

4.2 High-Dimensional 3D Image Data: Estimate of Muon Energy in a Granular Calorimeter

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 sub-nuclear 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 [4, 26, 13, 1]. The energy of a muon can be 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 alternatives, it has been observed [34, 18] that both the pattern and the size of small radiative energy losses that muons withstand in traversing finely segmented calorimeters can be used to infer the incident muon energy. This technique may preserve the power of muons as probes of new physics in future higher-energy colliders.

We tackle this problem by applying WALDO within a prediction framework. We have available 886,716 3D “images” and scalar true muon energies $\theta$ obtained through Geant4 [3], a high-fidelity stochastic simulator that encodes the physical process generating the calorimeter deposits. Data is available at [33]. As the interest is on constraining muon energies as much as possible while guaranteeing conditional coverage, we use three versions of the same data set, of increasing dimensionality: a 1D representation computed by summing over all calorimeter cells with deposited energy $E > 0.1$ GeV, for each muon; a 28-dimensional representation obtained by extracting 28 features from the spatial and energy information of the calorimeter cells [34]; and the full calorimeter measurements ($x_i \in \mathbb{R}^{51,200}$). For the first two datasets, we estimate the conditional mean and variance using Gradient Boosted Trees. For the full calorimeter data, we leverage the Deep 3D Convolutional Neural Network developed in [34]. In each case, we use Gradient Boosted Trees for quantile regression.

Figure 6 shows that confidence intervals constructed with WALDO achieve exact conditional coverage (68.3%) regardless of the data set used. The corresponding $1\sigma$ prediction intervals using full
Figure 6: Muon energy reconstruction with WALDO using calorimetric measurements of increasing dimensionality. WALDO (blue, orange, red) guarantees nominal coverage (68.3%), while 1σ prediction intervals (green) under- or over-cover in different regions of Θ and are wider on average than the corresponding WALDO intervals. Left: Energy deposited by a θ ≈ 3.2 TeV muon entering a calorimeter with 32 × 32 × 50 cells. Center: empirical coverage estimated via WALDO diagnostics. Right: Median lengths of constructed intervals.

calorimeter data, instead, exhibit over- or under-coverage in different regions of the parameter space. The latter is especially problematic because it implies that prediction sets for very high energies contain the true value with much lower probability than anticipated. This is due to prediction sets being centered around the point prediction, which is downward biased at high energies because of low signal-to-noise ratio in the calorimeter data for these kinds of muons. In terms of the length of the intervals, we make two key observations. First, using higher-dimensional representations of the particle-calorimeter interaction provides valuable information, with full calorimeter data yielding the tightest constraints on muon energies for the same level of coverage across the parameter space. Second, confidence intervals constructed with WALDO are on average even shorter than the corresponding prediction intervals, while also guaranteeing conditional coverage.

5 Discussion

We presented WALDO, a novel method to construct correctly calibrated confidence regions for parameters in inverse problems by recalibrating predictions and posteriors from prediction or posterior estimation algorithms. To increase power, one may be able to leverage Bayesian priors, as shown in Sections 3.3 and 4.1 or take advantage of the internal structure of the simulator as in [9]. Alternatively, one could adaptively simulate more data in specific regions of interest in the parameter space. The latter, and a more formal treatment of the relation between power and priors, are interesting directions for future studies. Note that the prediction approach is especially useful for settings with n = 1, p = 1 and large d (as seen in Section 4.2). The posterior approach, on the other hand, is particularly valuable for settings with multiple observations n and larger values of p. Finally, when dealing with nuisance parameters, standard (hybrid) approaches marginalize over them [13]. Hybrid methods do not formally control α, but offer a good approximation that can lead to robust results [48][49]. This approach can be easily incorporated into WALDO, and using the diagnostics procedure we can shed light on whether or not the final results have adequate conditional coverage, as was done in [16].

Acknowledgements We thank Niccolò Dalmasso for early feedback and discussions on this work, and for providing code previously written for LF2I. We are also indebted to Jan Kieseler and to Giles C. Strong for providing the muon energy data and the structure of the deep neural network employed for the studies described in Section 4.2 respectively. We also thank Michael Stanley for many valuable discussions on the details of WALDO. This work is supported in part by NSF DMS-2053804, NSF PHY-2020295, and the C3.ai Digital Transformation Institute. RI is grateful for the financial support of CNPq (309607/2020-5) and FAPESP (2019/11321-9). We are also grateful to Microsoft for providing Azure computing resources for this work.
References

[1] 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.

[2] Paul A Abell, Julius Allison, Scott F Anderson, John R Andrew, J Roger P Angel, Lee Armus, David Arnett, SJ Asztalos, Tim S Axelrod, Stephen Bailey, et al. LSST science book, version 2.0. *arXiv preprint arXiv:0912.0201*, 2009.

[3] 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.

[4] 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.

[5] 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.

[6] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian Computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

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

[8] Johann Brehmer, Kyle Cranmer, Gilles Louppe, and Juan Pavez. A guide to constraining effective field theories with machine learning. *Physical Review D*, 98(5):052004, 2018.

[9] 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.

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

[11] Yanzhi Chen and Michael U. Gutmann. Adaptive Gaussian copula ABC. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1584–1592. PMLR, 16–18 Apr 2019.

[12] 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.

[13] Cdf Collaboration et al. Observation of top quark production in Pbar-P collisions. *arXiv preprint hep-ex/9503002*, 1995.

[14] Robert D. Cousins and Virgil L. Highland. Incorporating systematic uncertainties into an upper limit. *Nuclear Instruments and Methods in Physics Research A*, 320(1):331–335, 1992. doi: https://doi.org/10.1016/0168-9002(92)90794-5.

[15] 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.

[16] 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.

[17] 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.
[18] 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.

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

[20] 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.

[21] 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.

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

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

[24] 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.

[25] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 2404–2414, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[26] 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.

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

[28] 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.

[29] 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, Feb 2019. ISSN 1537-2715. doi: 10.1080/10618600.2018.1546594. URL http://dx.doi.org/10.1080/10618600.2018.1546594.

[30] J. S. Johnson, L. A. Regayre, M. Yoshioka, K. J. Pringle, S. T. Turnock, J. Browse, D. M. H. Sexton, J. W. Rostron, N. A. J. Schutgens, D. G. Partridge, D. Liu, J. D. Allan, H. Coe, A. Ding, D. D. Cohen, A. Atanacio, V. Vakkari, E. Asmi, and K. S. Carslaw. Robust observational constraint of uncertain aerosol processes and emissions in a climate model and the effect on aerosol radiative forcing. Atmospheric Chemistry and Physics, 20(15):9491–9524, 2020.

[31] 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.

[32] 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.
[33] 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.

[34] 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.

[35] 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.

[36] Amanda Lenzi, Julie Bessac, Johann Rudi, and Michael L Stein. Neural networks for parameter estimation in intractable models. arXiv preprint arXiv:2107.14346, 2021.

[37] 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.

[38] 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. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 1289–1299. Curran Associates, Inc., 2017.

[39] 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.

[40] James G MacKinnon. Bootstrap hypothesis testing. Handbook of computational econometrics, 183:213, 2009.

[41] Jean-Michel Marin, Louis Raynal, Pierre Pudlo, Mathieu Ribatet, and Christian Robert. ABC random forests for Bayesian parameter inference. Bioinformatics (Oxford, England), 35, 05 2016. doi: 10.1093/bioinformatics/bty867.

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

[43] 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.

[44] 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.

[45] George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with Bayesian conditional density estimation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 29, pages 1028–1036. Curran Associates, Inc., 2016.

[46] 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.

[47] Pratik Patil, Mikael Kuusela, and Jonathan Hobbs. Objective frequentist uncertainty quantification for atmospheric CO2 retrievals. arXiv preprint arXiv:2007.14975, 2022.

[48] X Qian, A Tan, JJ Ling, Y Nakajima, and C Zhang. The Gaussian CLs method for searches of new physics. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 827:63–78, 2016.

[49] Stefan T. Radev, Ulf K. Mertens, Andreas Voss, Lynton Ardizzone, and Ullrich Köthe. BayesFlow: Learning complex stochastic models with invertible neural networks. IEEE Transactions on Neural Networks and Learning Systems, pages 1–15, 2020. doi: 10.1109/TNNLS.2020.3042395.
[50] Catia Scricciolo. Probability matching priors: A review. Journal of the Italian Statistical Society, 8:83–100, 1999. doi: 10.1007/BF03178943.

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

[52] 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.

[53] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration. arXiv preprint arXiv:1804.06788, 2018.

[54] 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.

[55] 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.

[56] Valérie Ventura. Bootstrap tests of hypotheses. In Analysis of parallel spike trains, pages 383–398. Springer, 2010.

[57] 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.
A Theoretical Properties

We assume that the quantile regression estimator described in Section 3.2 is consistent in the following sense:

**Assumption 1 (Uniform consistency)** Let $F(\cdot|\theta)$ be the cumulative distribution function of the test statistic $\tau(D; \theta_0)$ conditional on $\theta$, where $D \sim F_\theta$. Let $\hat{F}_{B'}(\cdot|\theta)$ be the estimated conditional distribution function, implied by a quantile regression with a sample $T'$ of $B'$ simulations $D \sim F_\theta$. Assume that the quantile regression estimator is such that

$$\sup_{\tau \in \mathbb{R}} |\hat{F}_{B'}(\tau|\theta_0) - F(\tau|\theta_0)| \xrightarrow{p} 0 \text{ as } B' \to \infty.$$ 

Assumption 1 holds, for instance, for quantile regression forests [42]. 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 $\tau(D; \theta_0)$ chosen according to step (ii) for a fixed $\alpha \in (0, 1)$. If the quantile estimator satisfies Assumption 1 then,

$$P_{D|\theta_0, C_{B'}}(\tau(D; \theta_0) \geq C_{B'}) \xrightarrow{a.s.} \alpha,$$

where $P_{D|\theta_0, C_{B'}}$ denotes the probability integrated over $D \sim F_{\theta_0}$ 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

$$\sup_{\tau \in \mathbb{R}} |\hat{F}_{B'}(\tau|\theta_0) - F(\tau|\theta_0)| = O_p\left(\left(\frac{1}{B'}\right)^r\right)$$

for some $r > 0$.

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

$$|P_{D|\theta_0, C_{B'}}(\tau(D; \theta_0) \geq C_{B'}) - \alpha| = O_p\left(\left(\frac{1}{B'}\right)^r\right).$$

Proofs of these results can be found in [40].
Algorithm 1 Construct a confidence region for $\theta$ using WALDO

**Require:** Simulated training sets $T, T', T''$; observed sample $D$; prediction algorithm or posterior estimator; number of samples $N_\hat{p}$ to draw for Monte Carlo sampling when using posterior estimator; quantile regressor; grid of parameter values $\Theta_{N_{\text{grid}}}$ at which to perform Neyman inversion; desired coverage level $1 - \alpha$.

1: // Estimate Conditional Mean and Variance or Posterior Distribution
2: if prediction algorithm then
3: Use $T$ to learn $\hat{E}[\theta|D]$ under squared error loss
4: Use $T$ again to compute $(\theta - \hat{E}[\theta|D])^2$ and learn $\hat{V}[\theta|D] = \hat{E}[(\theta - \hat{E}[\theta|D])^2|D]$ 
5: else if posterior estimator then
6: end if
7: // Estimate Critical Values
8: if prediction algorithm then
9: Predict $\hat{E}[\theta|D]$ and $\hat{V}[\theta|D]$ for each $D$ that appears in $T'$
10: else if posterior estimator then
11: for each $D$ that appears in $T'$ do
12: Draw $N_\hat{p}$ samples from the posterior $\hat{p}(\theta|D)$
13: Approximate $\hat{E}[\theta|D] \approx \frac{1}{N_\hat{p}} \sum_i \theta_i$, $\hat{V}[\theta|D] \approx \frac{1}{(N_\hat{p}-1)} \sum_i (\theta_i - \hat{E}[\theta|D])(\theta_i - \hat{E}[\theta|D])^T$
14: end for
15: end if
16: Compute $\hat{\tau}_{WALDO}(D; \theta)$ for all $(D, \theta) \in T'$
17: Learn critical values $\hat{C}_{\theta,\alpha}$ using the quantile regressor
18: // Construct Confidence Region via Neyman Inversion
19: if prediction algorithm then
20: Predict $\hat{E}[\theta|D]$ and $\hat{V}[\theta|D]$
21: else if posterior estimator then
22: Draw $N_\hat{p}$ samples from the posterior $\hat{p}(\theta|D)$
23: Approximate $\hat{E}[\theta|D] \approx \frac{1}{N_\hat{p}} \sum_i \theta_i$, $\hat{V}[\theta|D] \approx \frac{1}{(N_\hat{p}-1)} \sum_i (\theta_i - \hat{E}[\theta|D])(\theta_i - \hat{E}[\theta|D])^T$
24: end if
25: Initialize $R(D) \leftarrow \emptyset$
26: for $\theta_0 \in \Theta_{N_{\text{grid}}}$ do
27: if $\hat{\tau}_{WALDO}(D; \theta_0) \leq \hat{C}_{\theta_0,\alpha}$ then
28: $R(D) \leftarrow R(D) \cup \{\theta_0\}$
29: end if
30: end for
31: return confidence region $R(D)$
C Details on the Experiments

C.1 Gaussian Mixture Model

See Section 4.1 for a description of the experiment. Training was done on a MacBook Pro M1Pro (CPU only); it took approximately 15–20 minutes to train the posterior estimator, and an additional ~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 $T'$. The posterior was sampled multiple times for each $x \in T'$ to compute $E(\theta | x)$ and $V(\theta | x)$ via Monte Carlo integration; this procedure took a total of ~45 minutes (but could potentially be optimized in the future).

Figure 7 shows the output of the diagnostics procedure when using a uniform prior to train the posterior estimator (compare with Figure 5, right column, 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, which WALDO can correct via recalibration.

![Coverage Diagnostics](image)

Figure 7: 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.2 Muon Energy Reconstruction

See Section 4.2 for a description of the experiment. We had access to 886,716 simulated muons in total; roughly 200,000 muons were used to estimate the critical values, ~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. \[34\]. 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 at an Azure cloud computing machine. Estimating the critical values via quantile gradient boosted trees took approximately 1 minute.

Figure 8 shows comparison of 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: prediction sets are centered around the point prediction, which is downward biased at high energies, because of low signal-to-noise ratio in this regime.
Figure 8: **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.