Classification under Prior Probability Shift in Simulator-Based Inference: Application to Atmospheric Cosmic-Ray Showers

Alex Shen
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
ajshen@andrew.cmu.edu

Luca Masserano
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
lmassera@andrew.cmu.edu

Rafael Izbicki
Department of Statistics
Universidade Federal de São Carlos
São Paulo, Brazil 13565
rizbicki@ufscar.br

Tommaso Dorigo
Istituto Nazionale di Fisica Nucleare
Rome, Italy 00186
dorigo@pd.infn.it

Michele Doro
Department of Physics and Astronomy
Università di Padova
Padova, Italy 35122
michele.doro@unipd.it

Ann B. Lee
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
annlee@andrew.cmu.edu

Abstract

High-energy cosmic rays are informative probes of astrophysical sources in our galaxy. A main challenge is to separate gamma showers (extremely rare events of interest) from the vast majority of hadron showers, when we have access to realistic simulations of the shower production (forward process) but the prior distribution on the shower parameters is unknown. Direct classification of the showers using output data leads to biased predictions and invalid uncertainty estimates, since the prior is chosen by design and is different from the true distribution. We overcome these biases by proposing a new method that casts classification as a hypothesis testing problem under nuisance parameters. The main idea is to estimate ROC curves as a function of all nuisances, devising selection criteria that are valid under a generalized prior probability shift over both shower label and nuisance parameters. Our method yields a set-valued classifier that returns valid confidence sets for all levels α simultaneously without having to retrain the classifier for each level.

1 Introduction

Problem Set-Up. Simulator-based inference (SBI) refers to inference in a setting where the likelihood function $\mathcal{L}(\mathbf{x};\theta)$ – often associated with a "theory" about a phenomenon, e.g. in the physical sciences – is intractable. Whereas, it may be more feasible to simulate observable data $\mathbf{X} \in \mathcal{X}$ and generate large data sets $\mathcal{T}_B = \{(\theta_1, \mathbf{X}_1), \dots, (\theta_B, \mathbf{X}_B)\} \sim r(\theta)\mathcal{L}(\mathbf{x};\theta)$. The likelihood $\mathcal{L}(\mathbf{x};\theta)$ is implicitly encoded by the "theory" or mechanistic model $F_\theta : \theta \mapsto \mathbf{X}$, but the prior over parameters $r(\theta) = \mathbb{P}_{\text{train}}(\theta)$ is often *chosen by design* and different from the true distribution $\pi(\theta) = \mathbb{P}_{\text{target}}(\theta)$, hence causing a possibly harmful bias. When the unknown parameter of interest is a categorical

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

variable $\mu \in \mathcal{Y} = \{0, 1, \ldots, K\}$, and K is the number of classes, this difference in the joint distribution of $(\theta, \mathbf{X})$ between train and target data is referred to as prior probability shift or label shift [7, 8, 10]. Here we consider a generalized prior shift (GPS), where we assume a shift is happening not only in the distribution of the label μ (the parameter of interest), but also in a range of other parameters (nuisances or latent variables) $\nu \in \mathcal{N}$ of the mechanistic model, where $\theta = (\mu, \nu)$. In addition, we explicitly consider the case where the distribution over μ changes as a function of ν.

Motivating Application. High-energy cosmic rays both charged and neutral, are extremely informative probes of astrophysical sources in our galaxy and beyond. Neutral cosmic rays (gamma rays) come from a specific direction and target in the sky; charged cosmic rays (hadrons, which are mostly protons) come from any direction and are deflected by galactic magnetic fields. A vital step in analyzing gamma-ray sources using ground-based detector arrays is to separate gamma(G)-induced showers from the vast majority (>99.9%) of hadron(H)-induced showers based on ground measurements $\mathbf{x}$. The G/H separation problem is a challenging rare event detection problem under GPS, where the true distribution $\pi(\mu, \nu)$ of both the shower type μ and the shower parameters ν might be misspecified in simulated data.

Challenge. If one directly classifies ground measurements $\mathbf{x}$ using a classifier trained on $\mathcal{T}_B$, both the predictions and the associated uncertainty estimates will be biased, regardless of the amount of training data and the capacity of the classifier. The bias occurs because the posterior probability $\mathbb{P}'(\mu = 1|\mathbf{x})$ induced by the design prior $r(\theta)$ is not the same as the true class probability $\mathbb{P}(\mu = 1|\mathbf{x})$ induced by $\pi(\theta)$. This discrepancy can result in poor estimates of standard metrics, like true and false positive rates (TPR and FPR), and sub-optimal classification results. This problem is illustrated in Figure 1 (left) for the cosmic shower-ray application.

Our Approach and Contribution. By casting classification under GPS as a hypothesis testing problem with nuisance parameters, we are able to estimate TPR and FPR as continuous functions of *all* nuisance parameters via monotone regression and compute ROC curves as a function of $\nu \in \mathcal{N}$. We then derive selection criteria that are valid under GPS and obtain a set-valued classifier that returns valid $(1-\alpha)$ confidence sets for all levels α simultaneously, given an arbitrary observation $\mathbf{x}$, without any additional re-training of the base classifier. To the best of our knowledge, this is the first work that estimates ROC curves across the entire parameter space. Rather than using a surrogate likelihood or likelihood ratio (see references in [1]), we base our results directly on $\mathbb{P}'(\mu = 1|\mathbf{x})$. The ROC calibration framework of Section 2.2 has some similarities to [4, 12], which use monotone regression to estimate the CDF of probability integral transforms in predictive inference. The construction of set-valued classifiers of Section 2.4 is inspired by [3, 6, 9].

2 Methodology

2.1 Hypothesis Testing with the Bayes Factor Test Statistic: General Notation

Let $\mu = 0$ denote hadron-induced showers and $\mu = 1$ denote gamma-ray induced showers. Our goal is to discriminate atmospheric showers based on ground-based measurements $\mathbf{x} \in \mathcal{X}$. However, rather than directly classifying $\mathbf{x}$ based on a learned classifier $\mathbb{P}'(\mu = 1|\mathbf{x})$ and a cutoff C derived from $\{(\mu_i, \mathbf{x}_i)\}_{i=1}^{B}$, we reformulate the gamma/hadron discrimination problem as a composite-versus-composite hypothesis test:

$$H_{0,\mu_0} : \theta \in \Theta_0 \;\; \text{versus} \;\; H_{1,\mu_0} : \theta \in \Theta_1 \tag{1}$$

where $\Theta_0 = \{\mu_0\} \times \mathcal{N}$, $\Theta_1 = \{\mu_0\}^c \times \mathcal{N}$, and $\mu_0 \in \{0, 1\}$. As test statistic, we exploit

$$\tau_{\mu_0}(\mathbf{x}) = \frac{\mathbb{P}'(\mu = \mu_0|\mathbf{x})\; \mathbb{P}'(\mu \neq \mu_0)}{\mathbb{P}'(\mu \neq \mu_0|\mathbf{x})\; \mathbb{P}'(\mu = \mu_0)}. \tag{2}$$

which is equivalent to the Bayes factor for test 8; see Appendix B. This quantity can be estimated directly from a *pre-trained* classifier based on $\mathcal{T}_B$; there is no need for an extra step to, e.g., try to learn the likelihood function $\mathcal{L}(\mathbf{x}; \mu, \nu)$ or the associated likelihood ratio statistic from simulated data as done in [1] and references therein.

2.2 Estimating the ROC Curves Across the Entire Parameter Space: Gamma/Hadron Separation

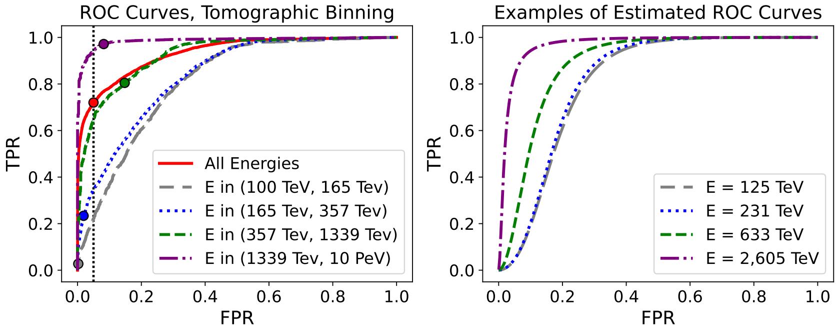

Figure 1: **ROC dependence on shower parameters.** *Left:* Empirical ROC curve for the entire calibration set (gray; "All Energies") and four tomographic bins over the calibration set according to true energy. Each dot on each ROC curve represents the same numeric cutoff on $\widehat{\tau}(\mathbf{x})$; this cutoff is chosen to achieve 0.05 FPR (vertical gray line) according to the "All Energies" curve. *Right:* Estimated ROC curves at energy levels roughly matching the bins in the left plot. These curves represent *single energy values* instead of ranges over bins.

For the case where $H_0 : \mu = 0$ versus $H_1 : \mu = 1$ (that is, $\mu_0 = 0$ in Equation 1), the estimated test statistic $\tau_{\mu_0}(\mathbf{x})$ defined by Equation 2 becomes

$$T(\mathbf{x}) = \frac{\mathbb{P}'(\mu = 0|\mathbf{x})\ \mathbb{P}'(\mu = 1)}{\mathbb{P}'(\mu = 1|\mathbf{x})\ \mathbb{P}'(\mu = 0)}. \tag{3}$$

We reject H_0 (that is, classify the shower as gamma-induced) for small values of $T(\mathbf{x}) := (1 - p(\mathbf{x}))p_1/(p(\mathbf{x})(1-p_1))$ where $p(\mathbf{x}) := \widehat{\mathbb{P}}'(\mu = 1|\mathbf{x})$ denotes the classifier output, and p_1 is the proportion of $\mu = 1$ instances in the train set.

To choose the optimal cutoff to reject H_0, we need some knowledge of how the classifier performs for different values ν of the nuisance parameters. For each cutoff $C \in \mathbb{R}$, let $I_C(\mathbf{x}) := \mathbb{I}(T(\mathbf{x}) \leq C)$ be the event that H_0 is rejected (and the shower labeled as gamma-induced). The key insight is that we can map out the receiver operating characteristic (ROC) of our test procedure over the *entire* parameter space through a monotone regression that estimates the rejection probability functions

$$\texttt{FPR}(C;\nu) := \mathbb{P}\left(T(\mathbf{X}) \leq C \mid \mu = 0, \nu\right) = \mathbb{E}_{\mathbf{X}|\mu=0,\nu}\left(I_C(\mathbf{X}) \mid \mu = 0, \nu\right) \tag{4}$$

$$\texttt{TPR}(C;\nu) := \mathbb{P}\left(T(\mathbf{X}) \leq C \mid \mu = 1, \nu\right) = \mathbb{E}_{\mathbf{X}|\mu=1,\nu}\left(I_C(\mathbf{X}) \mid \mu = 1, \nu\right), \tag{5}$$

for all $C \in \mathbb{R}$ and all $\nu \in \mathcal{N}$. At fixed ν, the ROC curve is defined as the true positive rate (TPR) vs false positive rate (FPR) over the space of cutoffs C. Appendix C and Algorithm 1 detail our procedure for estimating the rejection probability curves using calibration data $\mathcal{T}'_B = \{(\theta'_1, \mathbf{X}'_1), \ldots, (\theta'_B, \mathbf{X}'_{B'})\} \sim r(\theta)\mathcal{L}(\mathbf{x};\theta)$.

2.3 Selecting the Optimal Cutoff C under Generalized Prior Shift (GPS)

Once we have estimated the ROC curves as in Section 2.2, we can find the cutoff C for a new test point that either controls type-I error (FPR), or guarantees a minimum recall (TPR), or maximizes some merit function of choice that depends on both FPR and TPR. To achieve type-I error control at some pre-specified level $\alpha \in [0, 1]$, one could for example choose $C_{\alpha,0} = \inf_{\nu\in\mathcal{N}} \texttt{FPR}^{-1}(\alpha;\nu)$. Such a cutoff however is often overly conservative. An alternative approach, which leads to tighter constraints, is to first compute a $(1-\gamma)$ confidence set $R(\mathbf{x};\gamma)$ of ν at some level $\gamma \in [0, \alpha]$, e.g. using the techniques in [2, 3]. Then choose

$$C^*_{\alpha,0}(\mathbf{x}) = \inf_{\nu\in R(\mathbf{x};\gamma)} \texttt{FPR}^{-1}(\beta;\nu), \tag{6}$$

where $\beta = \alpha - \gamma$. This cutoff guarantees a FPR of at most α for any $\nu \in \mathcal{N}$, whereas *directly predicting* μ from $\mathbf{x}$ would not. Similarly, choosing a cutoff $\widetilde{C}^*_{\alpha,1}(\mathbf{x}) = \sup_{\nu\in R(\mathbf{x};\gamma)} \texttt{TPR}^{-1}(\beta;\nu)$ guarantees that the recall (TPR) is at least α. See Lemma 1 and proof in Appendix D.

2.4 Constructing Set-Valued Classifiers at Arbitrary Confidence Levels

Rather than just outputting a single label (G/H) for a particular shower, our framework provides valid measures of uncertainty under GPS and allows one to (i) rigorously quantify the evidence in favor of a shower being gamma-induced, and (ii) determine for which $\mathbf{x}$ the predicted labels are ambiguous. As mentioned, the probabilities $\mathbb{P}'(Y = 1|\mathbf{x})$ can lead to misleading answers under GPS. Hence, we instead report the output of a *set-valued classifier* $\mathbf{H} : \mathbf{x} \mapsto \{\emptyset, 0, 1, \{0, 1\}\}$, where the classifier guarantees user-defined levels of coverage $(1 - \alpha)$ or confidence (the probability that the true label is included in the set), no matter what the true class μ and the nuisance parameters ν are. At high levels of confidence, the classifier will report $\{0, 1\}$ for ambiguous instances $\mathbf{x}$, instead of forcing a 0 or 1 answer that has a high chance of being wrong. Section 3.2 shows illustrative examples, and Appendix D includes proofs of conditional coverage for the set-valued classifier given by

$$\mathbf{H}(\mathbf{x}; \alpha) = \left\{\mu_0 \in \{0, 1\} \mid \widehat{\tau}_{\mu_0}(\mathbf{x}) > C^*_{\alpha,\mu_0}(\mathbf{x})\right\}, \tag{7}$$

where $C^*_{\alpha,\mu_0}(\mathbf{x}) = \inf_{\nu \in R(\mathbf{x};\gamma)} W^{-1}_{\mu_0}(\beta; \nu)$ with the power function W_{μ_0} defined as in Equation 10 of Appendix C with $\mu = \mu_0$, and $\beta = \alpha - \gamma$.

3 Results: Gamma/Hadron Separation of Atmospheric Cosmic-Ray Showers

Data and Simulations We simulate cosmic ray showers using CORSIKA [5]. The shower parameters $\theta = (\mu, E, Z, A)$ include the identity μ, energy E, and the incident angles (Z, A) of the cosmic ray (primary particle). The output $\mathbf{x}$ denotes the measurements of all secondary shower particles that reach the ground. In our analysis, we treat the incident angles as known since they are easily reconstructed from sensor measurements; that is, the energy E is the only unknown latent variable $(\nu = E)$. From a total of 40,000 simulated showers, we construct the train set (to estimate $\tau_{\mu_0}(\mathbf{x})$), the calibration set (to estimate $\widehat{W}_{\mu}(C; \nu)$) and the test set (for evaluation and diagnostics) with sizes 20,000, 10,000 and 10,000, respectively. We fit $\widehat{\tau}_{\mu_0}(\mathbf{x})$ and $\widehat{W}_{\mu}(C; \nu)$ using shallow multi-layer perceptrons (MLP), where the model for $\widehat{W}_{\mu}(C; \nu)$ enforces monotonicity in C [11].

3.1 ROC Across the Entire Parameter Space

Nuisance parameters affect the performance of the classifier and the relative merits of different classifiers. Figure 1, *left*, shows how the ROC curve and subsequent FPR/TPR of the classifier depends on other shower parameters such as energy E. If we select the cutoff C based on $\mathbb{P}'(Y = 1|\mathbf{x})$ alone without regard to E, then we fail to achieve the nominal FPR or TPR conditional on particular energy ranges. The procedure detailed in Section 2.2 allows us to estimate an ROC curve for any $\nu \in \mathcal{N}$ (see *right* panel), and hence choose the correct cutoff C according to Section 2.4. Appendix E shows evidence that our ROC curves are indeed well calibrated.

3.2 Set-Valued Classification

Figure 2 shows the results of the set-valued classifier $\mathbf{H}(\mathbf{x}; \alpha)$ defined in Equation 7. For simplicity, in this work we set $\gamma = 0$ (i.e., we do *not* attempt to estimate ν given $\mathbf{x}$, hence the cutoff C_α in Equation 6 is computed by taking the infimum over the entire nuisance parameter space $\mathcal{N}$). The output is compared to results from the optimal Bayes classifier h_B, which yields $h_B(\mathbf{x}) = 1$ if $\widehat{\mathbb{P}}'(\mu = 1|\mathbf{x}) > \mathbb{P}'(\mu = 1)$, and $h_B(\mathbf{x}) = 0$ otherwise. Figure 2, *left* panel demonstrates how $\mathbf{H}(\mathbf{x}; \alpha)$ labels points close to the Bayes decision boundary (i.e. ambiguous points) as $\{0, 1\}$, with more points labeled as ambiguous, as the confidence level increases. The *center* panel shows that the set-valued classifier performs much better than the optimal Bayes counterpart by achieving higher precision (positive predictive value; blue curve) and lower false discovery rate (orange curve). This is partly because the set classifier opts to categorize instances as “ambiguous”, rather than incorrectly label difficult cases. The *right* panel shows that the set classifier also achieves a lower FNR (miss rate; orange curve) than optimal Bayes, but the TPR (recall; blue curve) is lower at higher confidence levels.

Appendix G verifies that our set-valued classifiers guarantee nominal coverage, even when we are not estimating ν (that is, we set $\gamma = 0$), whereas set-valued classifiers that ignore GPS fail to control

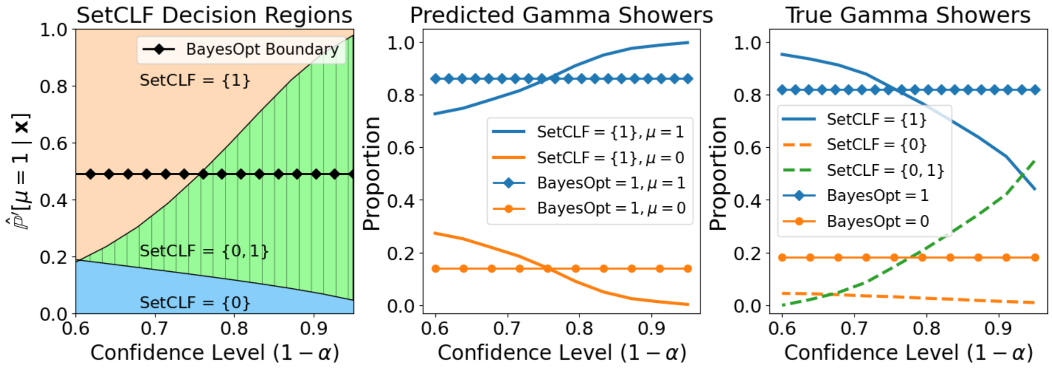

Figure 2: **Set classification output $\mathbf{H}(\mathbf{x};\alpha)$ (denoted "SetCLF") for different confidence levels, compared to optimal Bayes classifier $h_B(\mathbf{x})$ (denoted "BayesOpt").** *Left:* Decision regions for $\mathbf{H}(\mathbf{x};\alpha)$ as a function of confidence level. The $h_B(\mathbf{x})$ decision boundary is shown as reference, where $h_B(\mathbf{x}) = 1$ above the boundary. *Center:* Proportion of actual G-showers ($\mu = 1$) and H-showers ($\mu = 0$) among instances labeled or "predicted" as gamma showers. As the confidence level increases, precision (blue) increases and FDR (orange) decreases, reflecting the fact that SetCLF can choose not to output singleton predictions if the data are ambiguous. *Right:* Proportion of different classifier outputs for true gamma showers ($\mu = 1$). We achieve a lower miss rate (orange) at all confidence levels, but don't perform as well in terms of recall (blue) compared to BayesOpt at higher confidence levels.

type I errors. However, our classification results tend to be overly conservative, hence the lower TPR (or lower power). In future work, we can increase the power with larger train and calibration sets, which would also allow us to constrain ν with $(1-\gamma)$ confidence sets for $\gamma > 0$, while still guaranteeing user-defined levels of coverage; see Lemma 1 in Appendix D.

Broader Impact Statement. Systematic uncertainties due to model mis-specifications and nuisance parameters is a challenging problem for classification and rare event detection problems, especially in the physical sciences. This paper introduces a new method for handling prior probability shift of both label and nuisance parameters in simulation-based inference with a high-fidelity mechanistic model. We demonstrate a new technique for estimating the ROC across the entire parameter space. We also show how we can create set-valued classifiers that have a guaranteed user-specified probability $(1-\alpha)$ of including the true label (parameter of interest), for all levels $\alpha \in [0,1]$ simultaneously, without having to retrain the model for every α. These set-valued classifiers are valid, no matter what the true label and unknown nuisance parameters are, whereas classifiers that ignore GPS are not. We only consider one nuisance parameter in the paper, but the proposed framework scales to high-dimensional nuisance parameter spaces.

Acknowledgments. This work is supported in part by NSF DMS-2053804, NSF PHY-2020295, FAPESP-2023/07068-1 and CNPq-2019/11321-9.

References

[1] 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.

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

[3] Niccolò 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.

[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.

[5] Dieter Heck, Johannes Knapp, JN Capdevielle, G Schatz, T Thouw, et al. Corsika: A monte carlo code to simulate extensive air showers. *Report fzka*, 6019(11), 1998.

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

[7] Felipe Maia Polo, Rafael Izbicki, Evanildo Gomes Lacerda Jr, Juan Pablo Ibieta-Jimenez, and Renato Vicente. A unified framework for dataset shift diagnostics. *Information Sciences*, page 119612, 2023.

[8] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[9] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

[10] Afonso Fernandes Vaz, Rafael Izbicki, and Rafael Bassi Stern. Quantification under prior probability shift: The ratio estimator and its extensions. *The Journal of Machine Learning Research*, 20(1):2921–2953, 2019.

[11] Antoine Wehenkel and Gilles Louppe. Unconstrained monotonic neural networks. *Advances in neural information processing systems*, 32, 2019.

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

A Experimental Set-Up with Ground-Based Detector Arrays

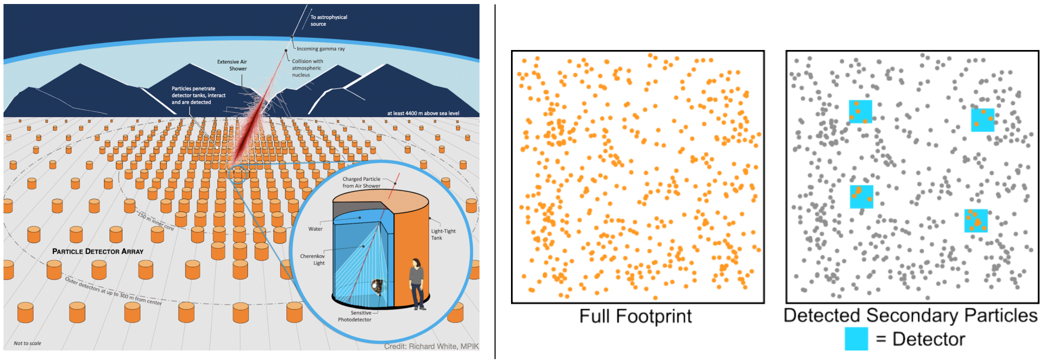
To astrophysical source
Incoming gamma ray
Collision with atmospheric nucleus
Extensive Air Shower
Particles penetrate detector tanks, interact and are detected
at least 4400 m above sea level
150 m inner core
Particle Detector Array
Outer detectors at up to 300 m from center
Charged Particle from Air Shower
Water
Light-Tight Tank
Cherenkov Light
Sensitive Photodetector
Not to scale
Full Footprint
Detected Secondary Particles
= Detector
Credit: Richard White, MPIK

Figure 3: *Left:* Artistic representation of the SWGO array. The inlay shows the individual detector unit. *Right:* Although we have access to all secondary particles in our simulated cosmic ray showers, we only include the particles that hit our simulated detector setup (blue rectangles) in the analysis. This layout pictured here is an illustrative example.

The data used in this paper are generated via the CORSIKA cosmic ray simulator [5]. CORSIKA is a Monte Carlo simulation program that models the interactions of primary cosmic rays with the Earth's atmosphere. Given values of the parameters μ, E, Z, A, which define the primary cosmic ray identity, energy, zenith and azimuth angle, respectively, CORSIKA outputs the identities, momenta, positions, and arrival times of all secondary particles generated in the atmospheric shower, that eventually reach the ground and that are mostly muons, electrons and photons at gamma-ray energies with minor abundance of heavier particles.

The measured data $\mathbf{x}$ in our analysis does not incorporate the full shower footprint, as this level of information cannot be captured in any realistic scenario. Instead, we simulate a simple 6×6 detector grid, where each detector covers a 2×2 m^2 area, with 48 m detector spacing. Information for a secondary particle of a particular shower footprint is incorporated into the analysis only if that secondary particle lands within the area of a detector. See Figure 3 (right) for a simplified representation of the detector grid.

We assume 100% detector efficiency and that all secondary particles types are detectable. We also assume that showers always originate at the center of the detector grid. Finally, we assume that both the zenith and azimuth angles Z and A are known due to the relative ease with which they can be estimate from observed footprint data. Thus, our only nuisance parameter for inference on μ is the energy E of the cosmic ray.

The data used to estimate the test statistic are drawn according to the following distribution (which may be different from that of actual astrophysical sources):

1. Gamma ray to Hadron ratio 1:1 (whereas actual observed ratios are in the range 1:1,000 – 1:100,000)
2. Energy between 100 TeV and 10 PeV, with probability density proportional to E^{-1} for gamma rays and E^{-2} for hadrons (with standard astrophysical sources closer to between -2:-4)
3. Zenith uniformly distributed between 0 and 65 degrees
4. Azimuth uniformly distributed between -180 and 180 degrees

To derive $\mathbf{x}_i$, we first define four secondary particle groups: photons (neutral); electrons and positrons; muons (charged); and all other secondary particle types. Then for each simulated detector, we record the count of particles in each group that hit the detector. This results in a vector of length $4 \cdot 36 = 144$ for each primary cosmic ray that represents the detector data. We construct $\mathbf{x}_i$ by concatenating the detector data with Z_i and A_i.

For the calibration and test sets, we use the same reference distribution.

B The Bayes Factor as a Frequentist Test Statistic

In this work, we treat the Bayes factor as a frequentist test statistic, similar to the Bayes Frequentist Factor (`BFF`) method in [3]. Consider the composite-versus-composite hypothesis test:

$$H_{0,\mu_0} : \theta \in \Theta_0 \text{ versus } H_{1,\mu_0} : \theta \in \Theta_1 \tag{8}$$

where $\Theta_0 = \{\mu_0\} \times \mathcal{N}$, $\Theta_1 = \{\mu_0\}^c \times \mathcal{N}$, and $\mu_0 \in \{0, 1\}$. The Bayes factor of the test is defined as

$$\tau_{\mu_0}(\mathbf{x}) := \frac{\mathbb{P}'(\mathbf{x}|H_{0,\mu_0})}{\mathbb{P}'(\mathbf{x}|H_{1,\mu_0})} = \frac{\int_{\mathcal{N}} \mathcal{L}(\mathbf{x}; \mu_0, \nu)\ r(\nu|\mu_0)\ d\nu}{\int_{\mathcal{N}} \mathcal{L}(\mathbf{x}; \mu \neq \mu_0, \nu)\ r(\nu|\mu \neq \mu_0)\ d\nu}$$

By Bayes theorem,

$$\begin{aligned}
\tau_{\mu_0}(\mathbf{x}) &= \frac{\int_{\mathcal{N}} \frac{p'(\mu_0,\nu|\mathbf{x})}{p'(\mu_0,\nu)} r(\nu|\mu_0)\ d\nu}{\int_{\mathcal{N}} \frac{p'(\mu\neq\mu_0,\nu|\mathbf{x})}{p'(\mu\neq\mu_0,\nu)} r(\nu|\mu \neq \mu_0)\ d\nu} \\
&= \frac{\int_{\mathcal{N}} \frac{p'(\mu_0,\nu|\mathbf{x})}{\mathbb{P}'(\mu=\mu_0)}\ d\nu}{\int_{\mathcal{N}} \frac{p'(\mu\neq\mu_0,\nu|\mathbf{x})}{\mathbb{P}'(\mu\neq\mu_0)}\ d\nu} \\
&= \frac{\mathbb{P}'(\mu = \mu_0|\mathbf{x})\ \mathbb{P}'(\mu \neq \mu_0)}{\mathbb{P}'(\mu \neq \mu_0|\mathbf{x})\ \mathbb{P}'(\mu = \mu_0)}.
\end{aligned} \tag{9}$$

However, unlike `BFF`, we are not estimating the likelihood or odds from simulated data, but instead directly evaluate a pretrained classifier $\mathbb{P}'(\mu = 1|\mathbf{x})$.

C Estimating the Power Function

Definition 1 (Power Function)**.** *Consider the composite-versus-composite hypothesis test* H_{0,μ_0} *(Equation 8) with the test statistic* $\tau_{\mu_0}(\mathbf{x})$ *(Equation 9). For* $\mu \in \{0, 1\}$*,* $\nu \in \mathcal{N}$*, and* $C \in \mathbb{R}$*, the power function of the test is defined as*

$$W_\mu(C; \nu) := \mathbb{P}_{\mu,\nu} \left(\tau_{\mu_0}(\mathbf{x}) \leq C\right). \tag{10}$$

We learn $W_\mu(C; \nu)$ using a monotone regression that enforces that the power is a non-decreasing function of C. For each point i ($i = 1, \ldots, n$) in the calibration sample $\mathcal{D}_\mu = \{(\nu_1, \mathbf{X}_1), \ldots, (\nu_n, \mathbf{X}_n)\}$ (generated at fixed μ), we choose a set of K cut-offs from a grid $G = \{C_1, \ldots, C_K\}$ of K uniformly spaced quantiles of $\lambda(\mathbf{X}_i)$ for $\mathbf{X}_i \in D_\mu$ Then, we regress the random variable

$$Y_{i,j} := \mathbb{I}\left(\widehat{\tau}_{\mu_0}(\mathbf{X}_i) \leq C_j\right) \tag{11}$$

on *both* ν_i and $C_{i,j}$ ($= C_j$) using the augmented calibration sample $\mathcal{D}'_\mu = \{(\nu_i, C_{i,j}, Y_{i,j})\}_{i,j}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, K$. See Algorithm 1 for details.

D Selecting the Cutoff C under the Presence of Nuisance Parameters

Let $T(\mathbf{x})$ be a test statistic for $H_0 : \mu = 0$, and let $\nu \in \mathcal{N}$ denote the nuisance parameters.

Definition 2 (Confidence set for nuisance parameters)**.** *Let* $\mu_0 \in \{0, 1\}$ *and let* $\gamma \in [0, 1]$*. The random set* $R_{\mu_0}(\mathbf{x}; \gamma)$ *is a valid* $(1 - \gamma)$ *level confidence set for* ν *at fixed* μ_0*, if*

$$\mathbb{P}_{\mu_0,\nu_0} \left(\nu_0 \in R_{\mu_0}(\mathbf{X}; \gamma)\right) = 1 - \gamma,\ \ \forall \nu_0 \in \mathcal{N}. \tag{12}$$

Definition 3 (Type I error control)**.** *For each* $\nu_0 \in \mathcal{N}$ *and level* $\beta \in [0, 1]$*, let* C_{ν_0} *be such that*

$$\mathbb{P}_{\mu_0=0,\nu_0} \left(T \leq C_{\nu_0}\right) = \beta.$$

Let

$$C^*(\mathbf{x}) = \inf_{\nu_0 \in R_0(\mathbf{x};\gamma)} \{C_{\nu_0}\}, \tag{13}$$

where $R_0(\mathbf{x}; \gamma)$ *is a* $(1 - \gamma)$ *level confidence set for* ν *when* $\mu_0 = 0$*.*

```
Algorithm 1 Learning the Power Function
Require: true class label $\mu \in \{0,1\}$; test statistic $\lambda$; calibration data $\mathcal{D}_\mu = \{(\nu_1, \mathbf{X}_1), \ldots, (\nu_n, \mathbf{X}_n)\}$; grid of cut-offs $G = \{C_1, \ldots, C_K\}$; evaluation points $\mathcal{V} \subset \mathcal{N}$
Ensure: Estimate of power function $W_\mu(C;\nu)$ for all $\nu \in \mathcal{V}$

// Learn power function from augmented calibration data $\mathcal{D}'$
Set $\mathcal{D}' \leftarrow \emptyset$
for $i$ in $\{1, ..., n\}$ do
    for $j$ in $\{1, ..., K\}$ do
        Compute $Y_{i,j} \leftarrow \mathbb{I}\left(\lambda(\mathbf{X}_i) \le C_j\right)$
        Let $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(\nu_i, C_j, Y_{i,j})\}$
    end for
end for
Use $\mathcal{D}'$ to estimate $W_\mu(C;\nu) := \mathbb{P}_{\mu,\nu}\left(\lambda(\mathbf{X}) \le C\right)$ via a regression of $Y$ on $\nu$ and $C$, which is
  monotonic w.r.t. $C$.
return estimated rejection probabilities $\widehat{W}_\mu(C;\nu)$, for $C \in G$ and $\nu \in \mathcal{V}$
```

Definition 4 (Minimum precision). *For each $\nu_0 \in \mathcal{N}$ and level $\beta \in [0,1]$, let $\widetilde{C}_{\nu_0}$ be such that*

$$\mathbb{P}_{\mu_0=1,\nu_0}\left(T \le \widetilde{C}_{\nu_0}\right) = \beta.$$

Let

$$\widetilde{C}^*(\mathbf{x}) = \sup_{\nu_0 \in R_1(\mathbf{x};\gamma)} \{\widetilde{C}_{\nu_0}\}, \tag{14}$$

where $R_1(\mathbf{x};\gamma)$ is a $(1-\gamma)$ level confidence set for ν when $\mu_0 = 1$.

Lemma 1. *Choose a threshold $\alpha \in [0,1]$ and $\gamma \in [0,\alpha]$. Let $R_{\mu_0}(\mathbf{x};\gamma)$ be a valid $(1-\gamma)$ level confidence interval for ν at fixed $\mu_0 \in \{0,1\}$ according to Equation 12. Let $\beta = \alpha - \gamma$, and define $C^*(\mathbf{x})$ according to Equation 13. Then, for all $\nu_0 \in \mathcal{N}$,*

$$\mathbb{P}_{\mu=0,\nu_0}(T \le C^*(\mathbf{X})) \le \alpha \quad \textit{(type I error control)}$$

Similarly, if we let $\beta = \alpha + \gamma$ and define $\widetilde{C}^*(\mathbf{x})$ according to Equation 14, then for all $\nu_0 \in \mathcal{N}$,

$$\mathbb{P}_{\mu=1,\nu_0}(T \le \widetilde{C}^*(\mathbf{X})) \ge \alpha \quad \textit{(minimum precision).}$$

Proof. Notice that

$$\begin{aligned}
\mathbb{P}_{\mu=0,\nu_0}(T \le C^*(\mathbf{x})) &= \mathbb{P}_{\mu=0,\nu_0}(T \le C^*(\mathbf{x}), \nu_0 \in R_0(\mathbf{X};\gamma)) + \mathbb{P}_{\mu=0,\nu_0}(T \le C^*(\mathbf{x}), \nu_0 \notin R_0(\mathbf{X};\gamma)) \\
&\le \mathbb{P}_{\mu=0,\nu_0}(T \le C_{\nu_0}) + \mathbb{P}_{\mu=0,\nu_0}(\nu_0 \notin R_0(\mathbf{X};\gamma)) \\
&\le \beta + \gamma = \alpha.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{P}_{\mu=1,\nu_0}(T \ge \widetilde{C}^*(\mathbf{x})) &= \mathbb{P}_{\mu=1,\nu_0}(T \ge \widetilde{C}^*(\mathbf{x}), \nu_0 \in R_1(\mathbf{X};\gamma)) + \mathbb{P}_{\mu=1,\nu_0}(T \ge \widetilde{C}^*(\mathbf{x}), \nu_0 \notin R_1(\mathbf{X};\gamma)) \\
&\le \mathbb{P}_{\mu=1,\nu_0}(T \ge \widetilde{C}_{\nu_0}) + \mathbb{P}_{\mu=1,\nu_0}(\nu_0 \notin R_1(\mathbf{X};\gamma)) \\
&\le 1 - \beta + \gamma = 1 - \alpha,
\end{aligned}$$

and therefore

$$\mathbb{P}_{\mu=1,\nu_0}(T \le \widetilde{C}^*(\mathbf{x})) \ge \alpha$$

□

E Diagnostics of Estimated ROC Curves

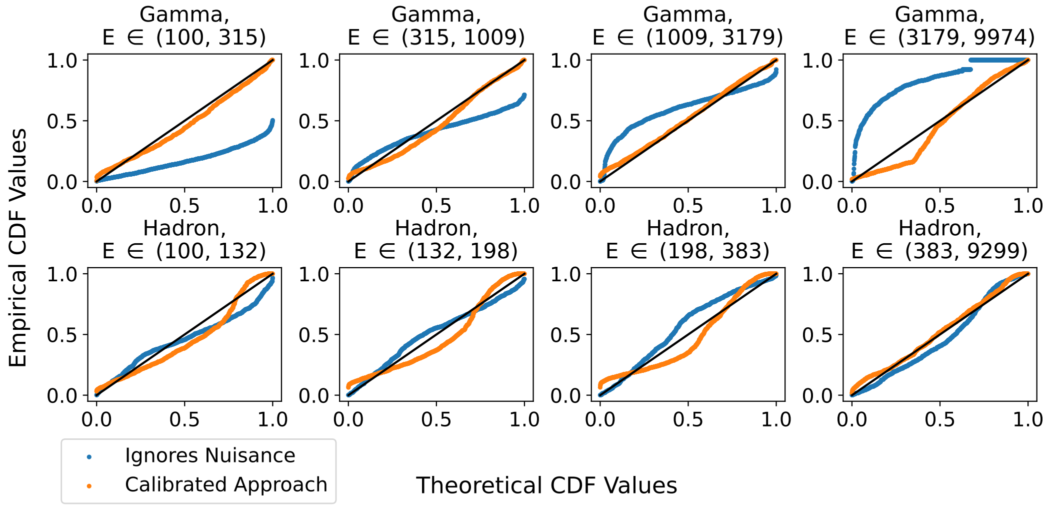

Figure 4: **Consistency check of estimated TPR and FPR functions for our approach, compared to an approach that ignores nuisance parameters.** Probability-Probability plots to assess how well the estimated power function $\widehat{W}_\mu(C;\nu)$ fits the test data for each shower type μ (rows) in tomographic energy bins (columns). Here, we only show results for the test statistic $\tau_{\mu_0=0}$. The blue curves are generated by estimating the marginal power function $W'_\mu(C)$ that averages over different values of ν the marginal power function $W'_\mu(C)$ that averages over different values of ν (we take the empirical CDF of $\tau_{\mu_0}(\mathbf{x})$ over the calibration set as an example). We then divide the test data into bins based on energy. On the y-axis, we plot the values of $\widehat{W}'_\mu(\tau(\mathbf{x}_i))$ for each $\mathbf{x}_i$ in that bin, and on the x-axis we plot the theoretical values of $W'_\mu(\tau(\mathbf{x}_i))$ if TPR and FPR did not depend on ν (which would then follow a uniform distribution). We see that this assumption is not supported by the data, in particular for gamma showers at very low or high energies. We repeat the same procedure for $\widehat{W}_\mu(\tau(\mathbf{x}_i);\nu_i)$ for each $(\mathbf{x}_i,\nu_i)$ in the test set and find that our approach (orange) better aligns with the observed data than the approach that ignores nuisance parameters (blue).

F Set-Valued Classifiers: Additional Results

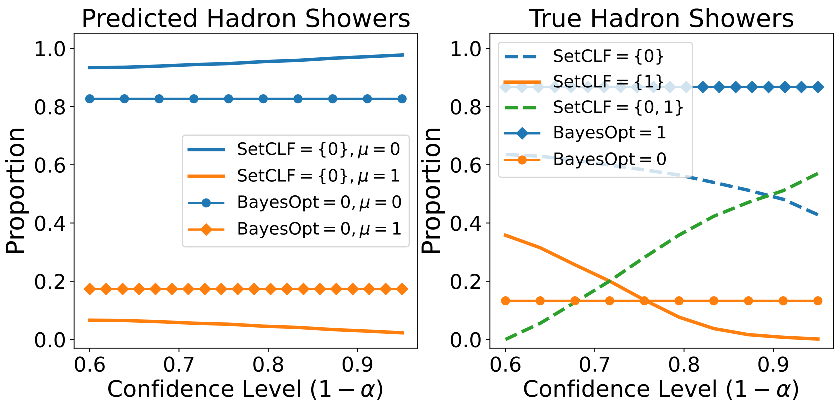

Figure 5: **Set classification output $\mathbf{H}(\mathbf{x};\alpha)$ (denoted "SetCLF") for different confidence levels, compared to optimal Bayes classifier $h_B(\mathbf{x})$ (denoted "BayesOpt").** *Left:* Proportion of actual G-showers ($\mu=1$) and H-showers ($\mu=0$) among instances labeled or "predicted" as hadron showers. *Right:* Proportion of different classifier outputs for true gamma showers ($\mu=1$).

G Diagnostics of Set-Valued Classifiers

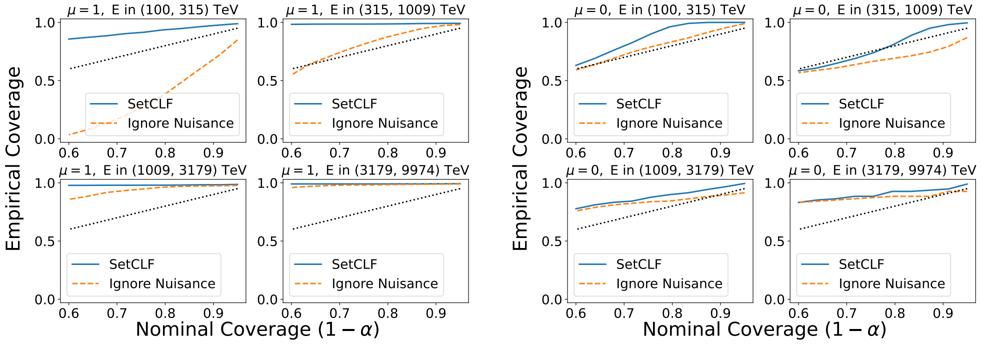

Figure 6: **Empirical coverage of set-valued classifiers**. For the true gamma showers (left panel) and true hadron showers (right panel), we divide each group into four bins according to true shower energy. We then check that $\mathbf{H}(\mathbf{x}; \alpha)$ (denoted "SetCLF" in the plots) achieves nominal coverage (black dotted lines) for both G/H-showers and at each energy bin. We compare this to an approach that attempts to control FPR without accounting for nuisance parameters (orange dashed line). We see that our approach is always valid (though sometimes also overly conservative), whereas ignoring nuisance parameters may lead to severe under-coverage, especially for gamma showers at lower energies