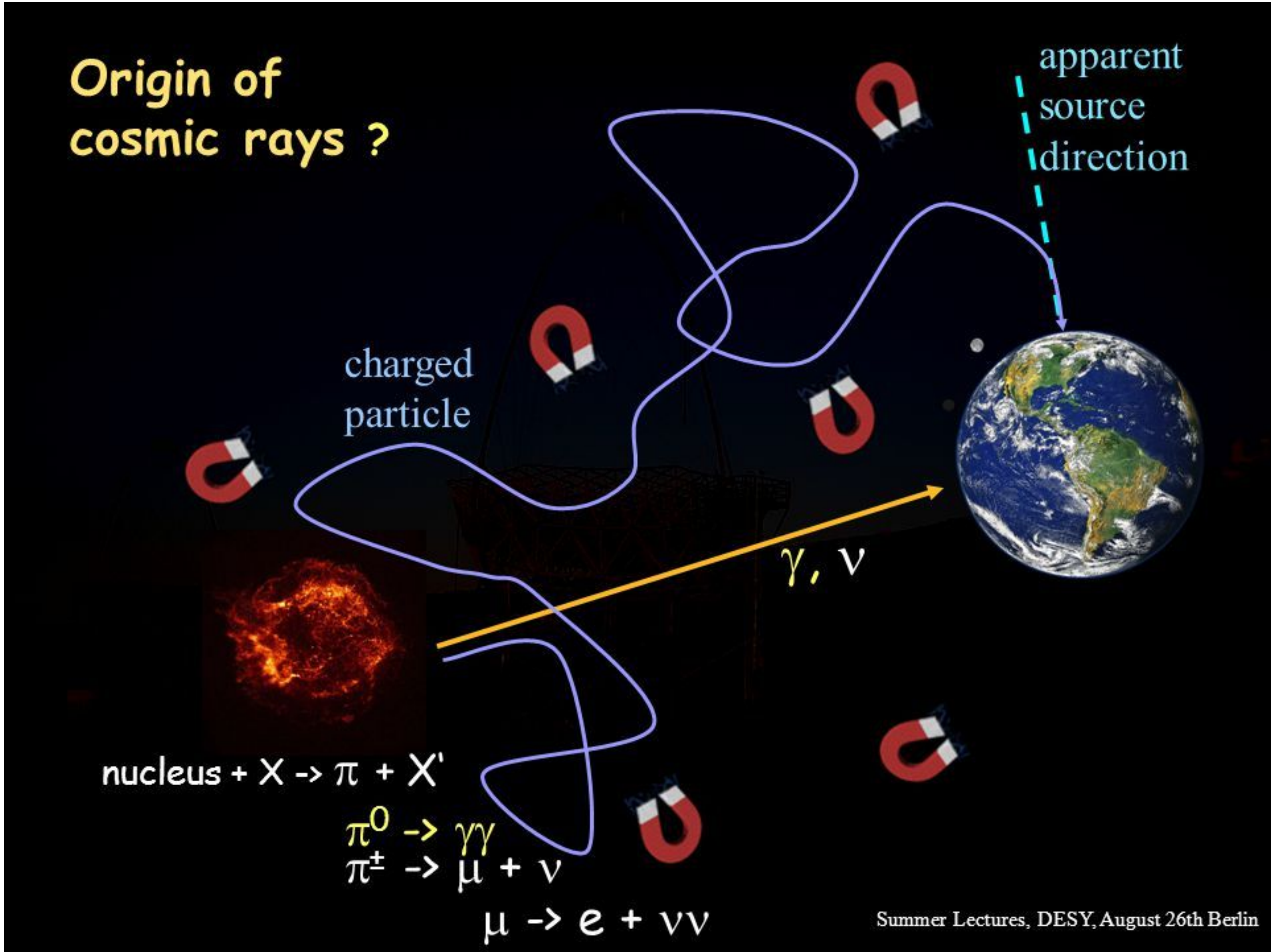
Machine learning model for discrimination of simulated muonic traces in water cherenkov detectors

Erick Richard Berazain Mallea

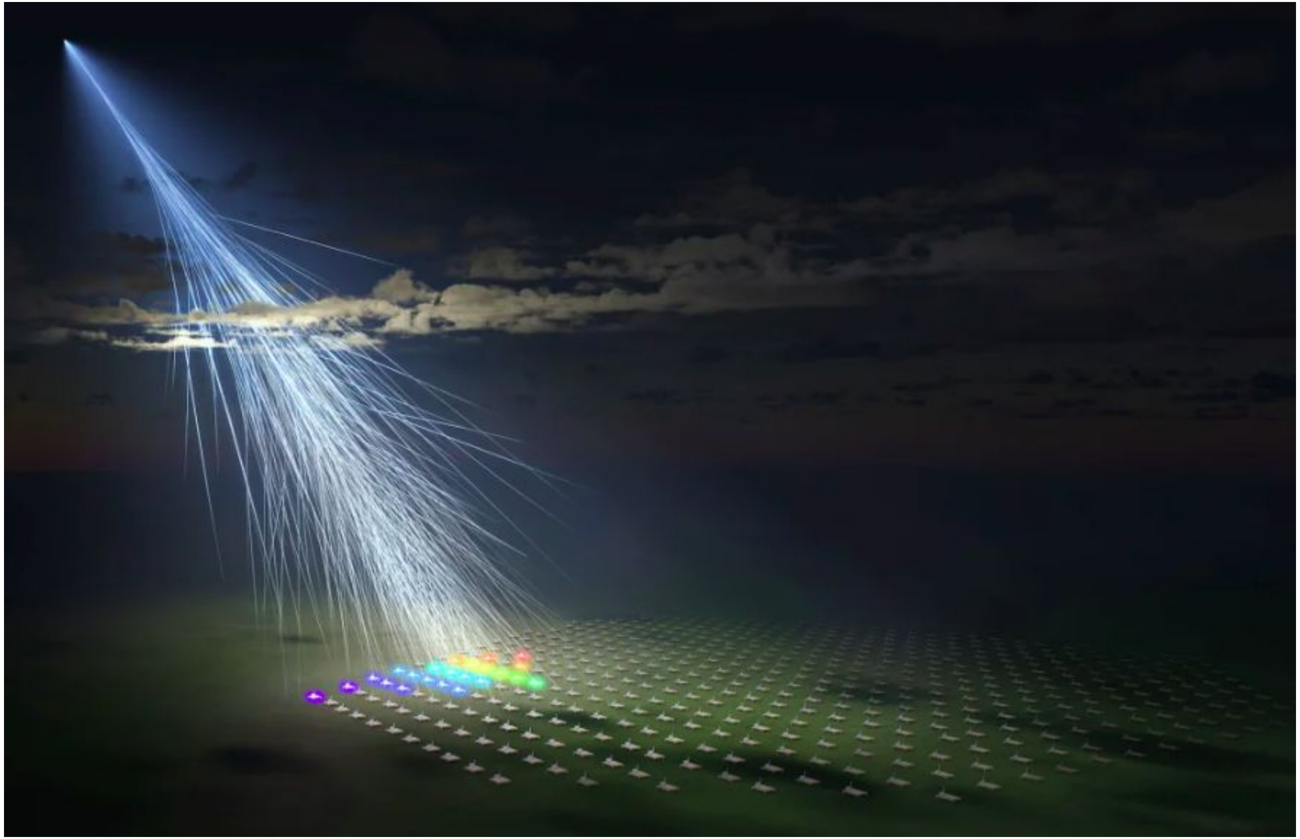
Co-authors
Martín Alfonso Subieta Vasquez Ph.D
Hugo Marcelo Rivera Bretel Ph.D

Introduction

Cosmic Rays


Origin of cosmic rays ?
apparent source direction
charged particle
γ, ν
nucleus + X -> π + X'
π0 -> γγ
π± -> μ + ν
μ -> e + νν
Summer Lectures, DESY, August 26th Berlin

EAS and its detection



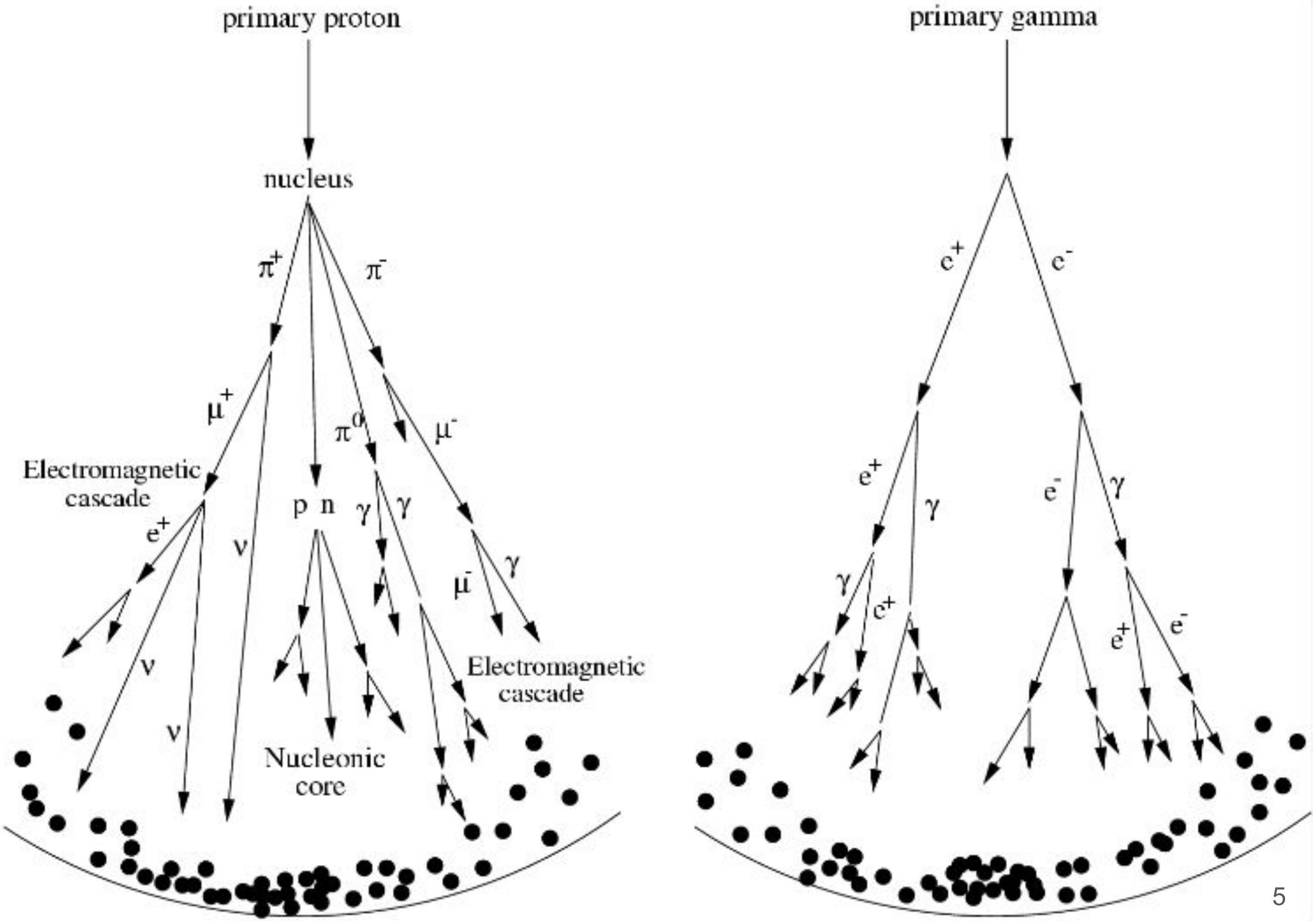
What can we estimate?

- Direction
- Energy
- Composition

Composition

Three components:

- Hadronic
- Electromagnetic
- Muonic



ALPACA Proyect

Andes

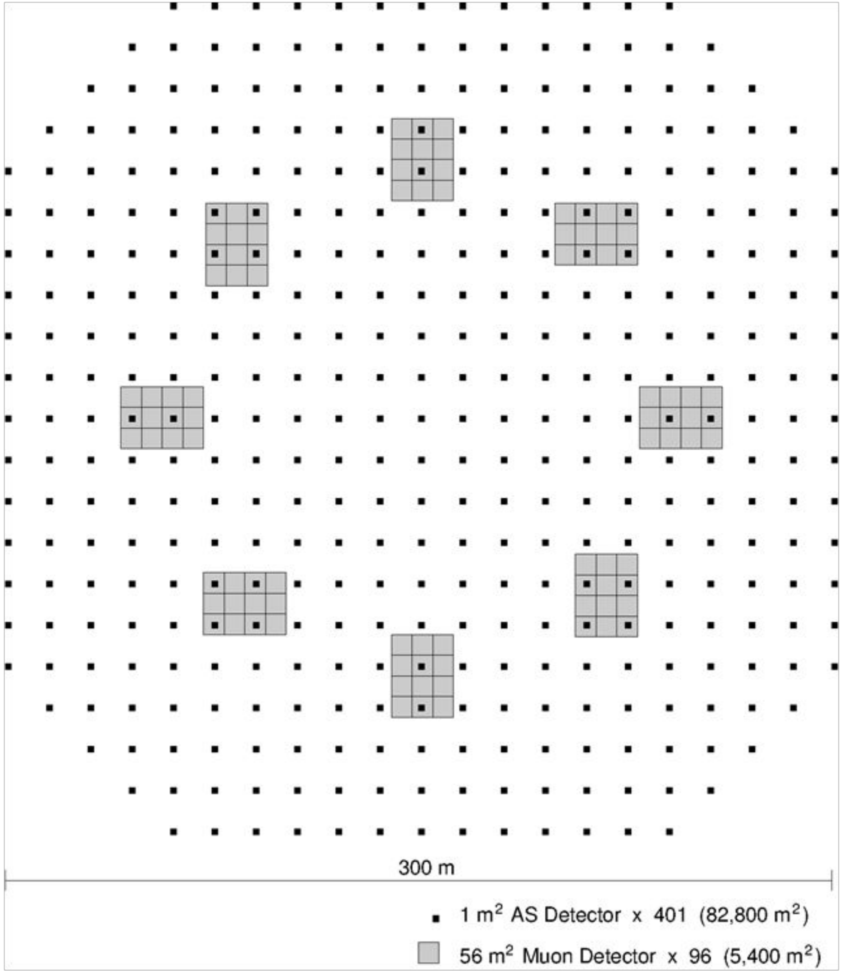
Large area

PArticle detector for

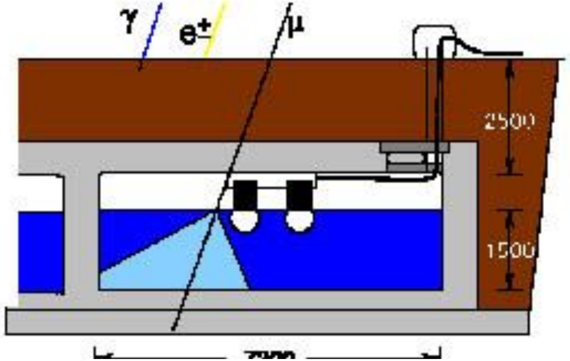
Cosmic ray physics and

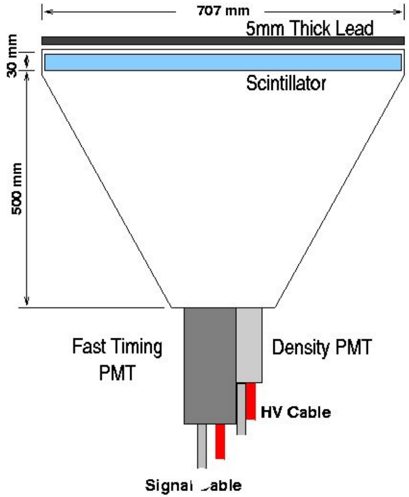
Astronomy

Detector set up

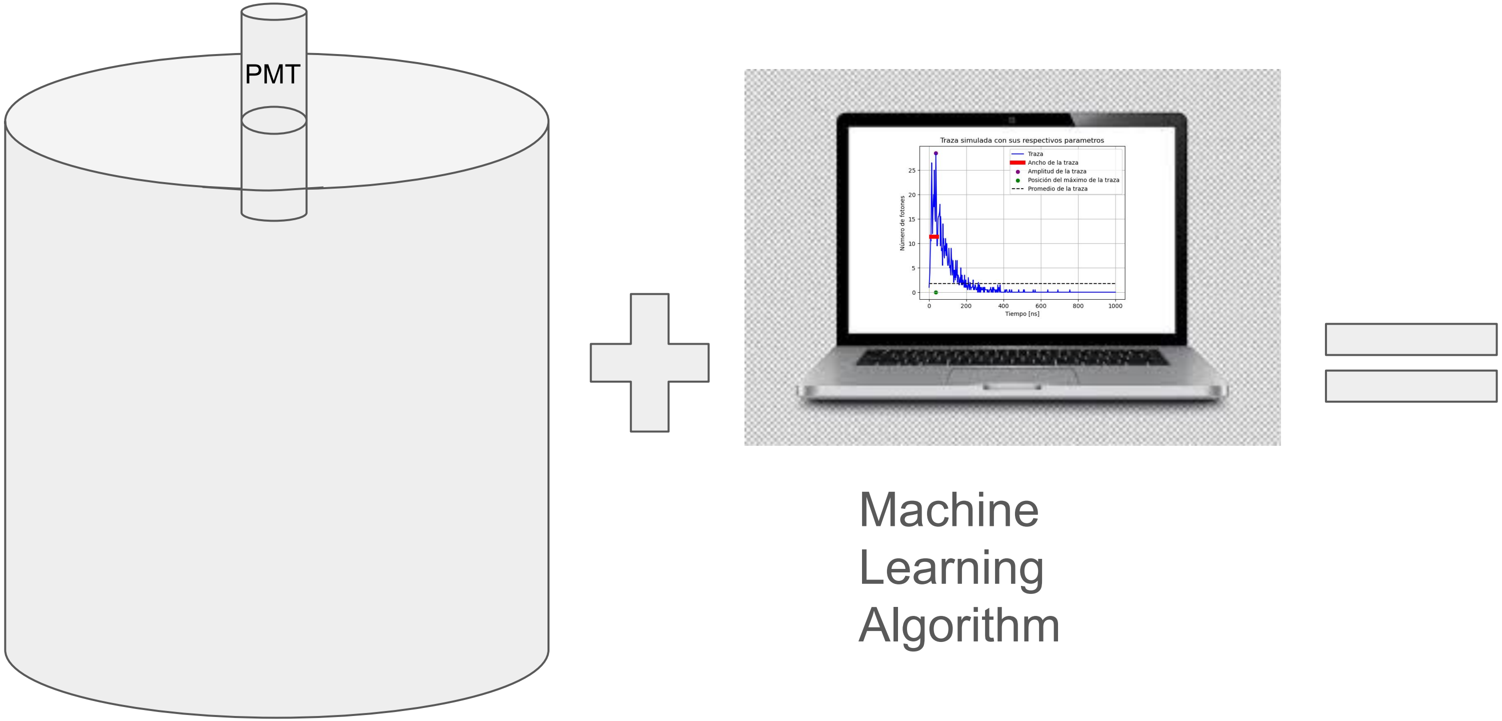

300 m
1 m^2 AS Detector x 401 (82,800 m^2)
56 m^2 Muon Detector x 96 (5,400 m^2)

Detectors





Proposal



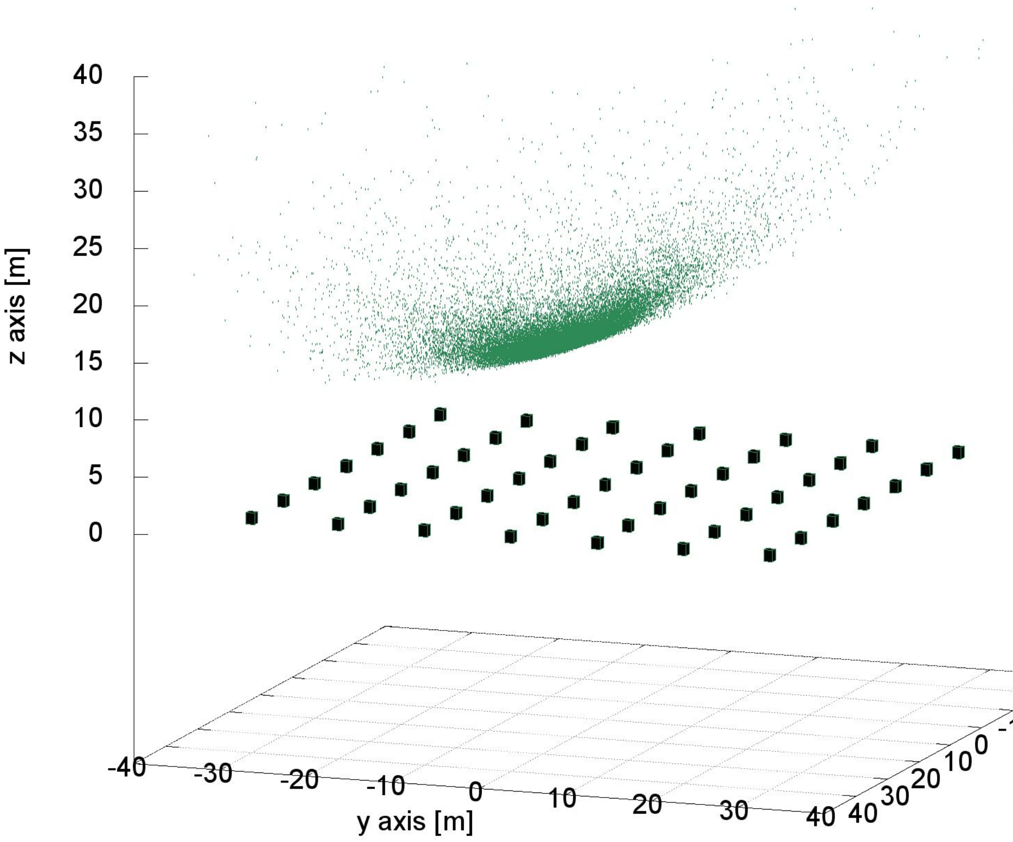
Machine Learning Algorithm

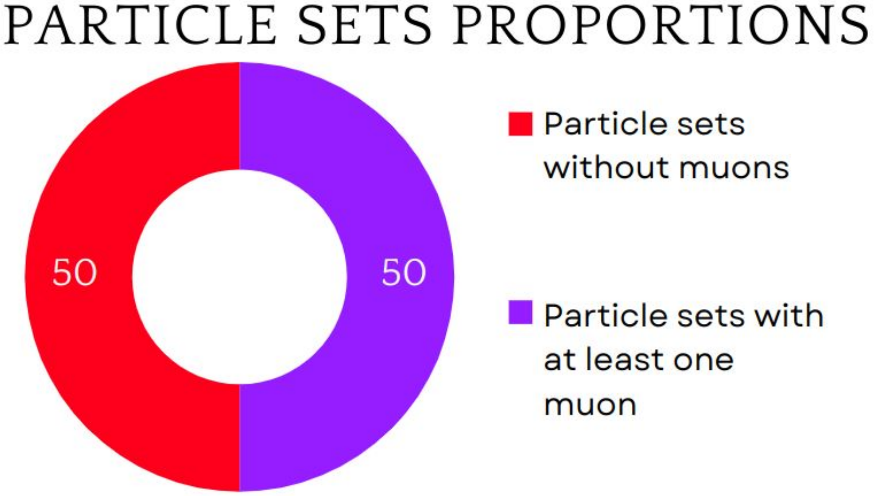
Probability to detect a muon on the WCD

What is the efficiency of a logistic classification algorithm to separate traces containing muons from traces without muons in a simulated WCD?

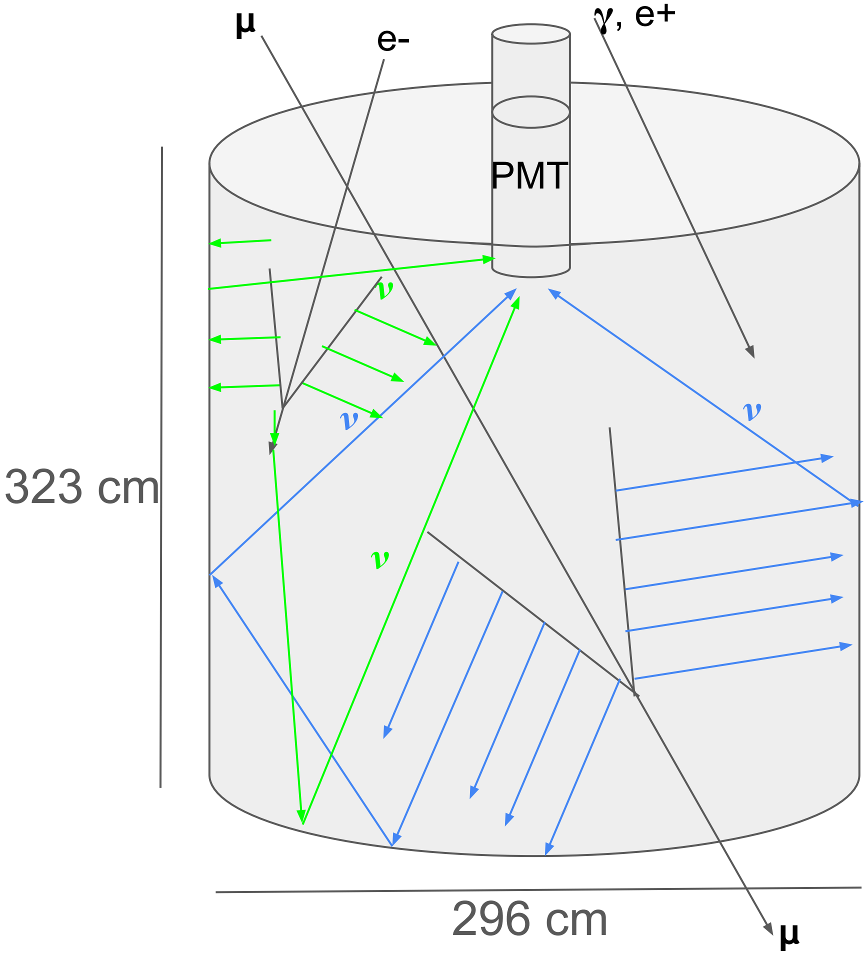
Simulations

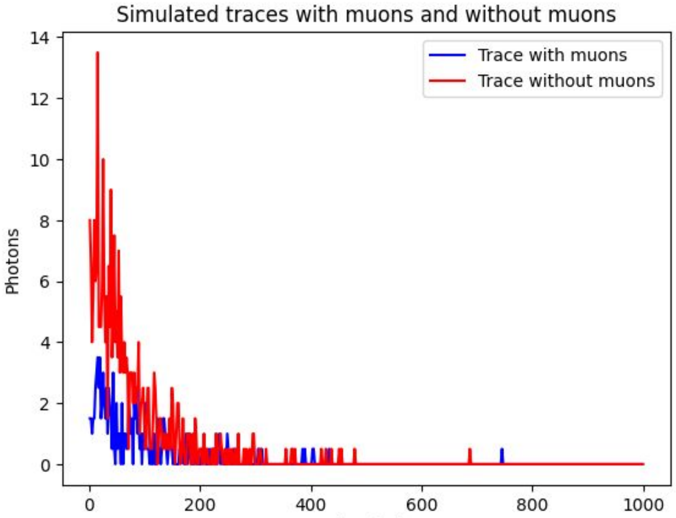
Particle simulations

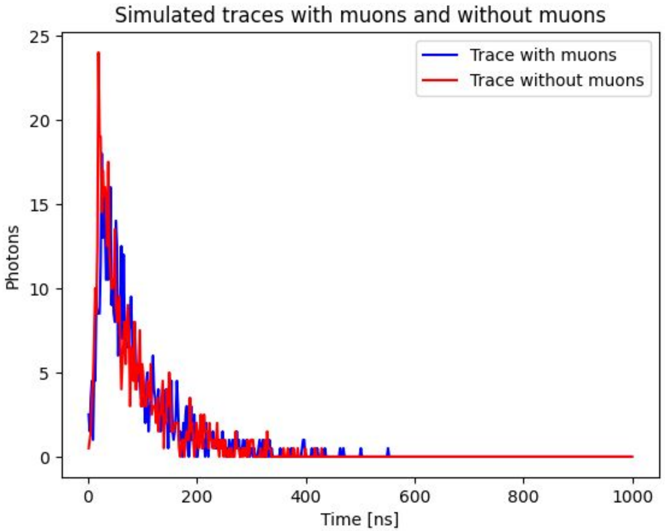




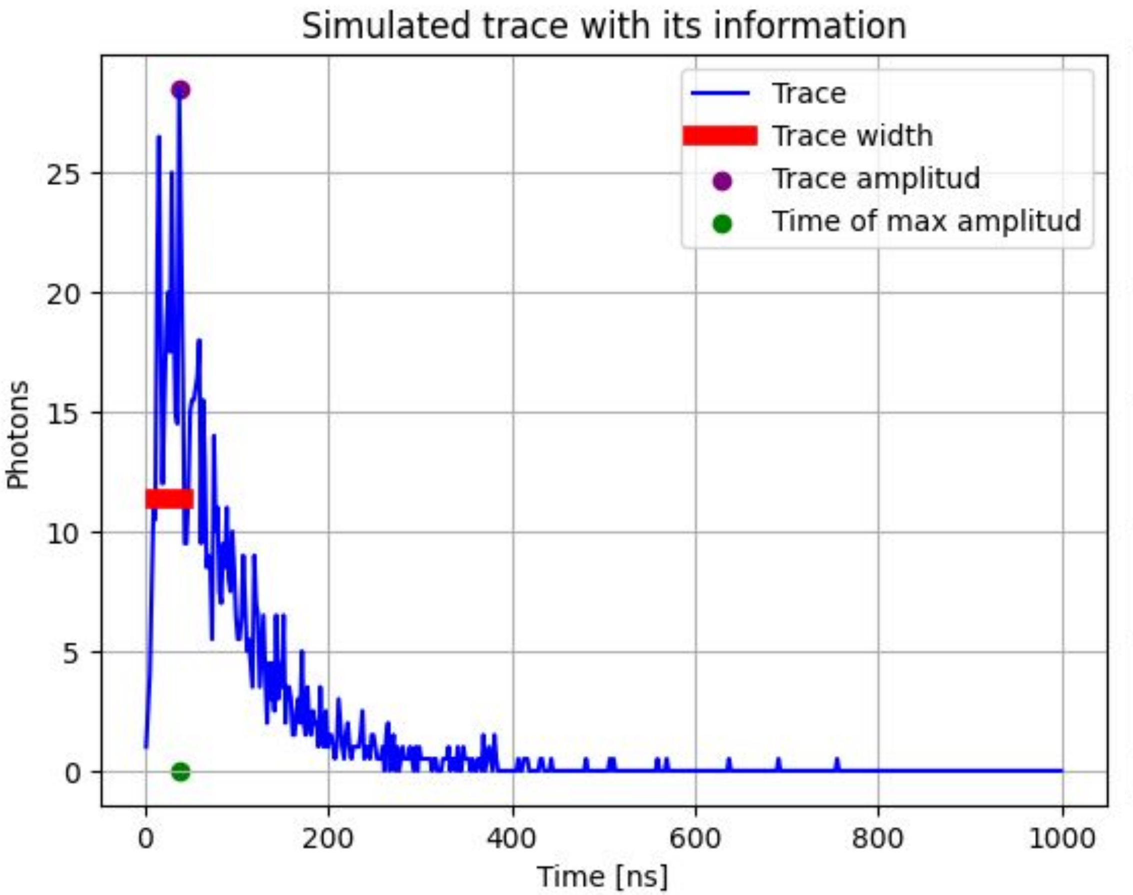
Detector and traces simulations





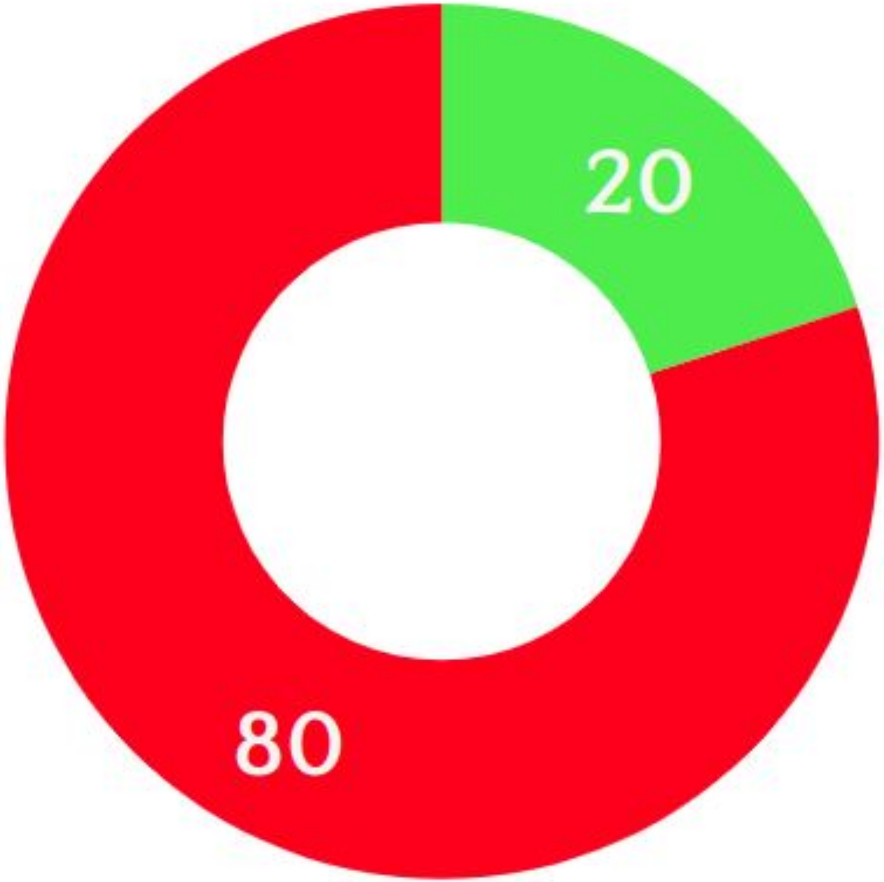


Information from each trace



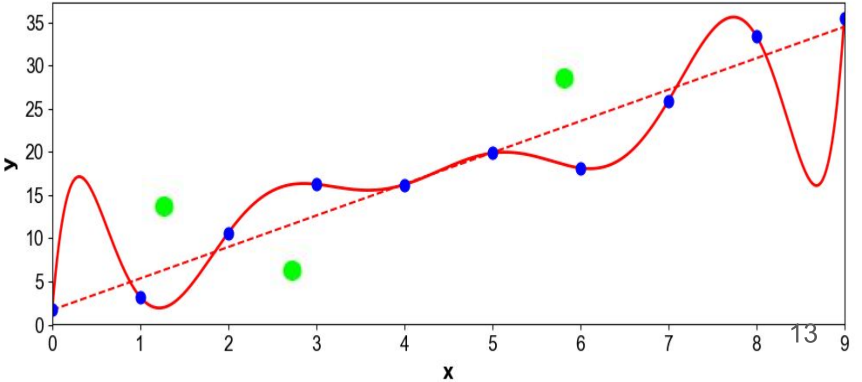
- Integral: Total number of photons that arrived to the PMT.
- Max amplitude: The maximum number of photons on the trace.
- Max amplitude time: Time when the trace reached the maximum number of photons.
- Trace width: Time interval since the number of photons got over the 40% of the max amplitude until it got under the 40% of the max amplitude.

DATA SETS



🟥 Training data set

🟩 Testing data set



Classification algorithm

Binary logistic regression

How can we determine the probability that a trace contains a muon?

- Trace width at 40 % ($X_{i,1}$)
- Max amplitud time ($X_{i,2}$)
- Log of trace integral ($X_{i,3}$)

Linear combination:

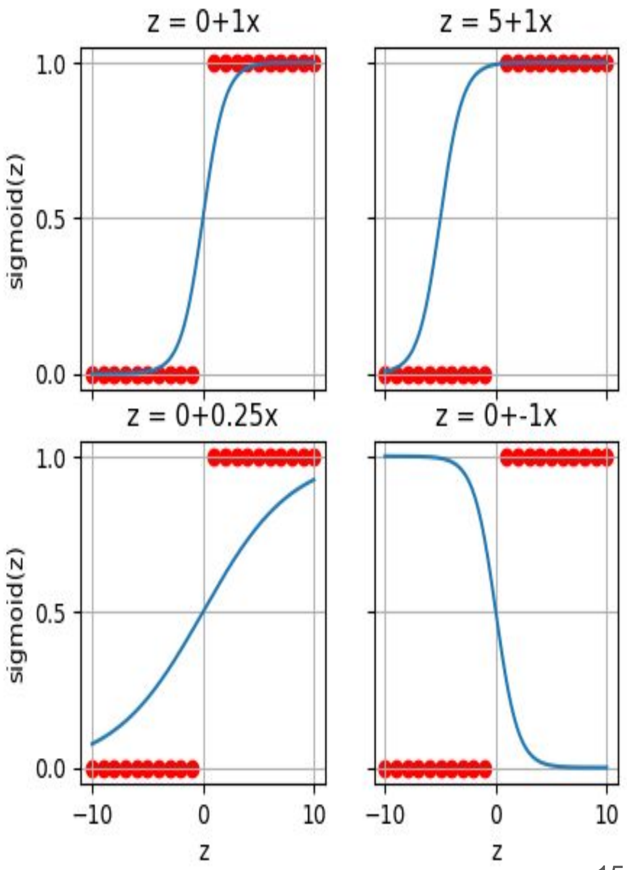
$$z_i = \vec{X}_i \vec{\theta} = \theta_0 + X_{i,1}\theta_1 + X_{i,2}\theta_2 + X_{i,3}\theta_3$$

Probability:

$$P = F(z_i) = \frac{1}{1+e^{-z_i}}$$

Coefficients:

$$\vec{\theta} = [\theta_0, \theta_1, \theta_2, \theta_3] = ???$$



How can we determine the coefficients?

$$\vec{\theta} = [\theta_0, \theta_1, \theta_2, \theta_3]$$

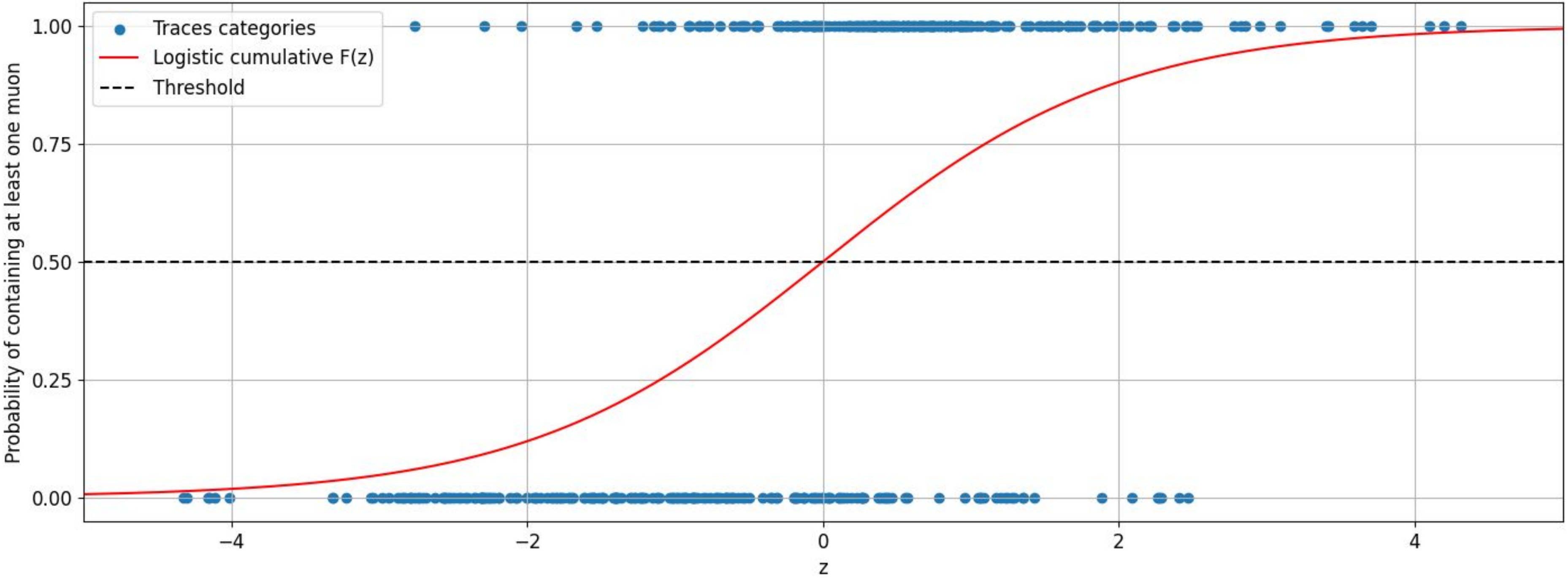
Coefficients	Estimate	Error
θ_0	-7.24	0.06
θ_1	0.004	0.001
θ_2	0.0048	0.0001
θ_3	2.49	0.02

Bernoulli:

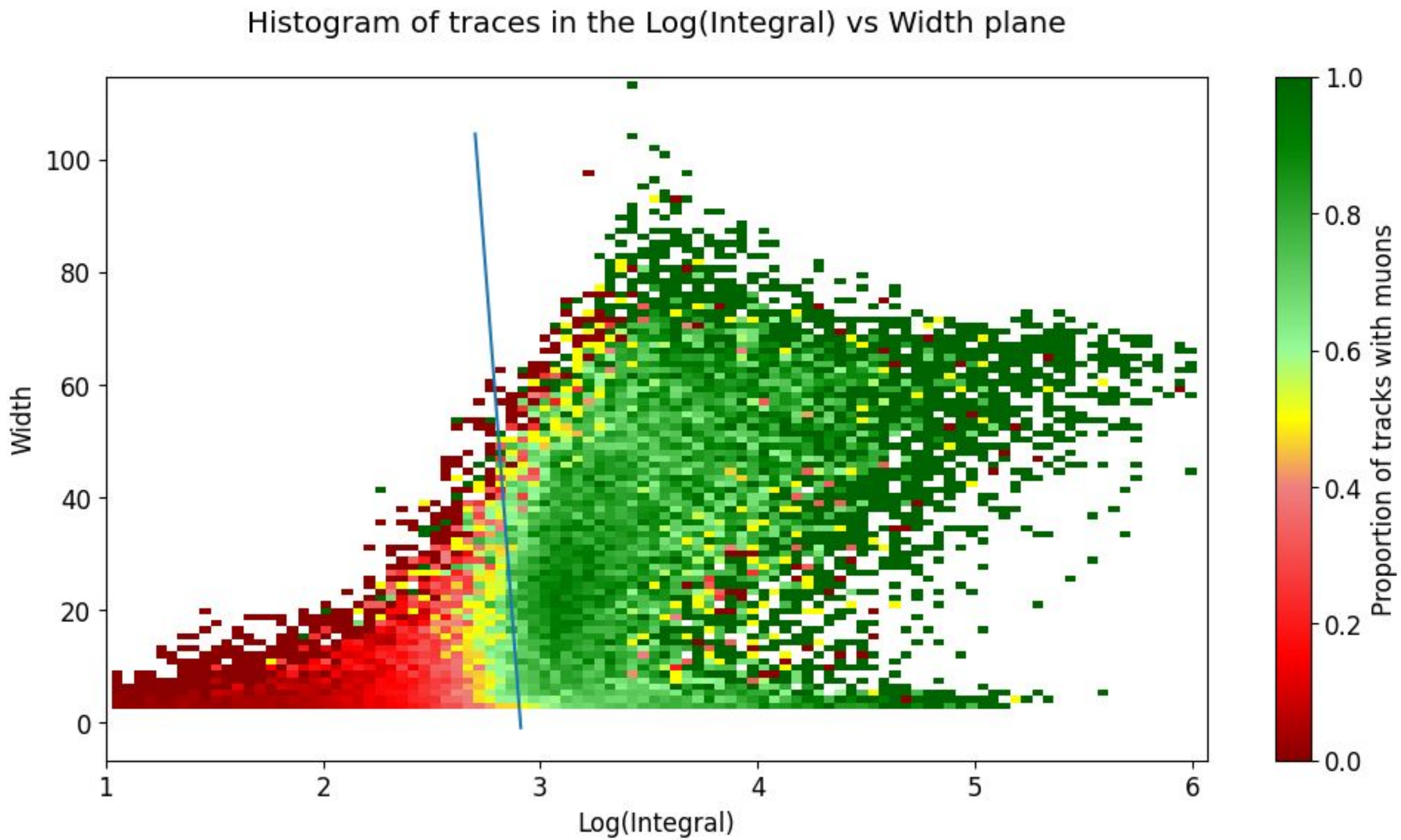
$$f(y) = P^y(1-P)^{1-y}$$

Likelihood:

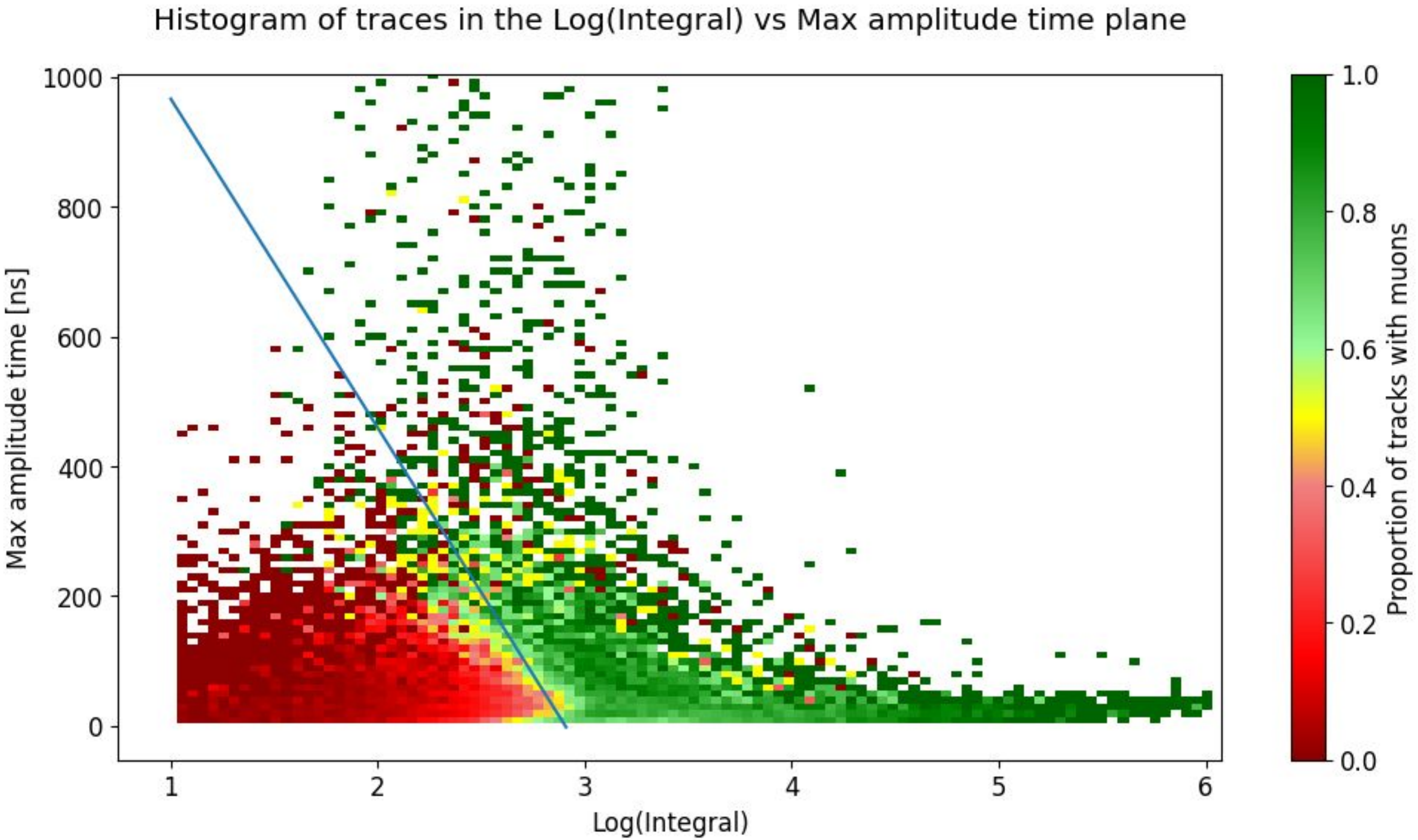
$$L(\vec{\theta}; X) = \prod_{i=1}^{m} \left(\frac{1}{1+e^{-\vec{X}_i\vec{\theta}}}\right)^{y_i} \left(\frac{e^{-\vec{X}_i\vec{\theta}}}{1+e^{-\vec{X}_i\vec{\theta}}}\right)^{1-y_i}$$



Projection of the tracks on the integral and the width of the track plane

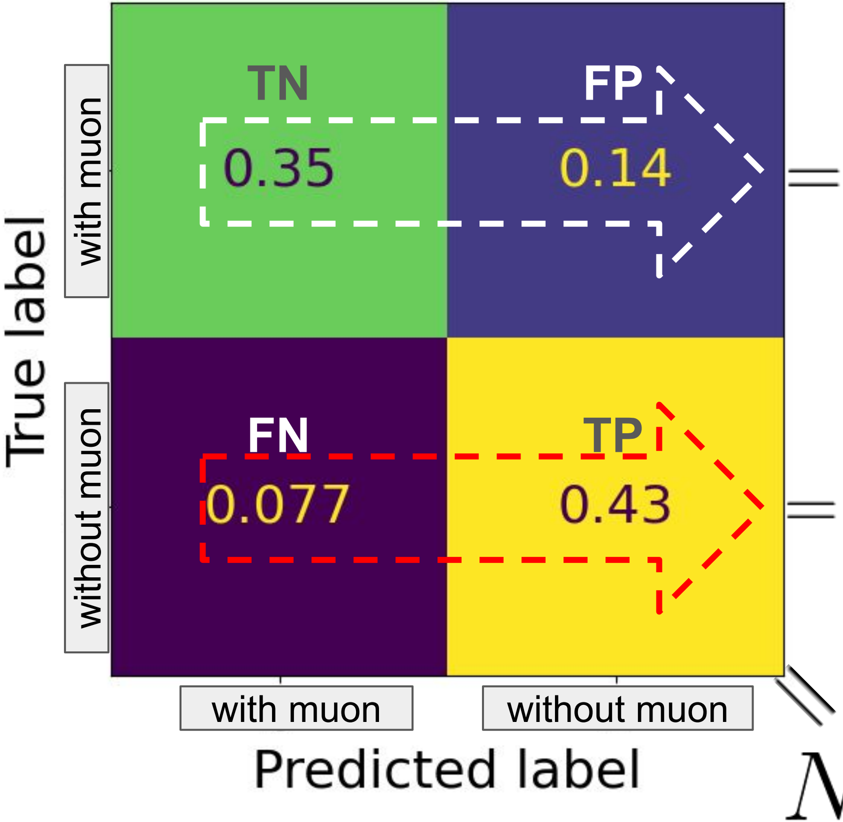


Proyección de las trazas en el plano de la integral y de la posición de la amplitud máxima



Analysis and results

Confusion matrix



$$= N_{EM}$$

$$= N_{\mu}$$

$$N_{success} = TN + TP$$

Efficiency:

$$\epsilon = \frac{N_{success}}{N}$$

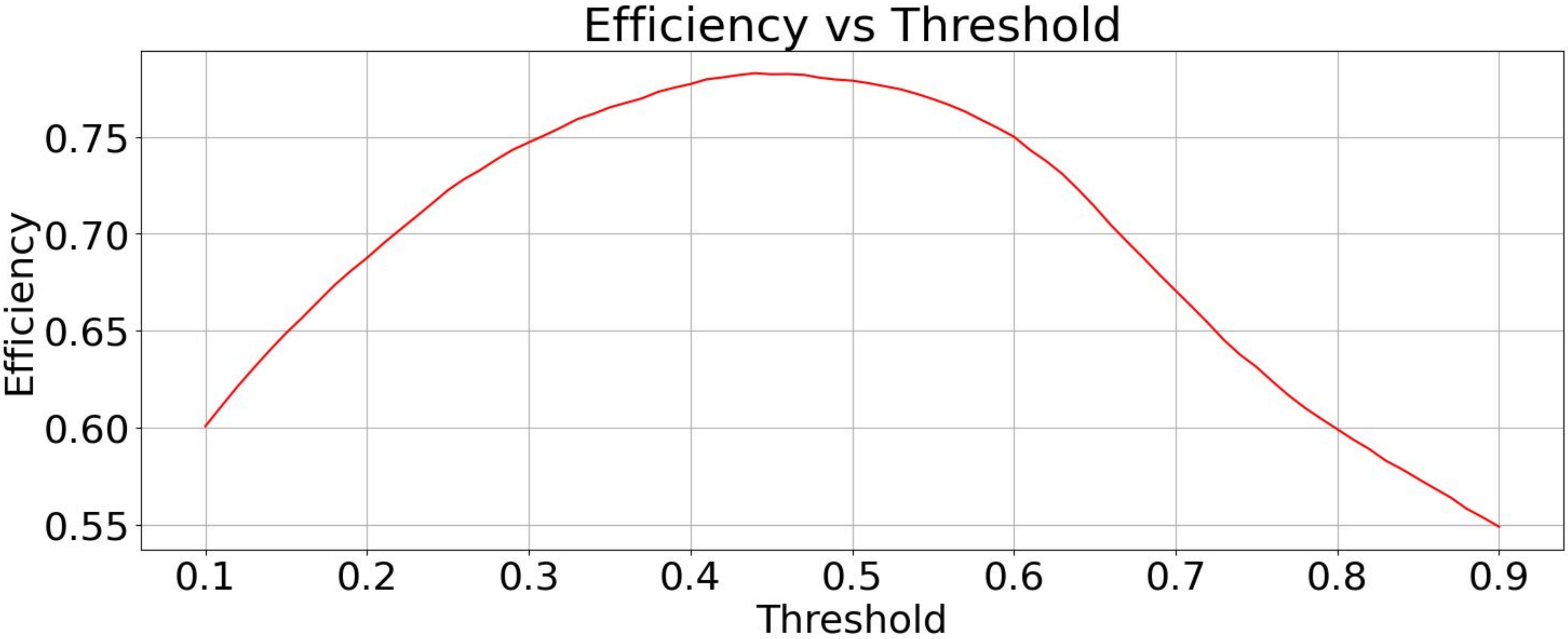
Sensibility:

$$S = \frac{TP}{N_{\mu}}$$

Specificity:

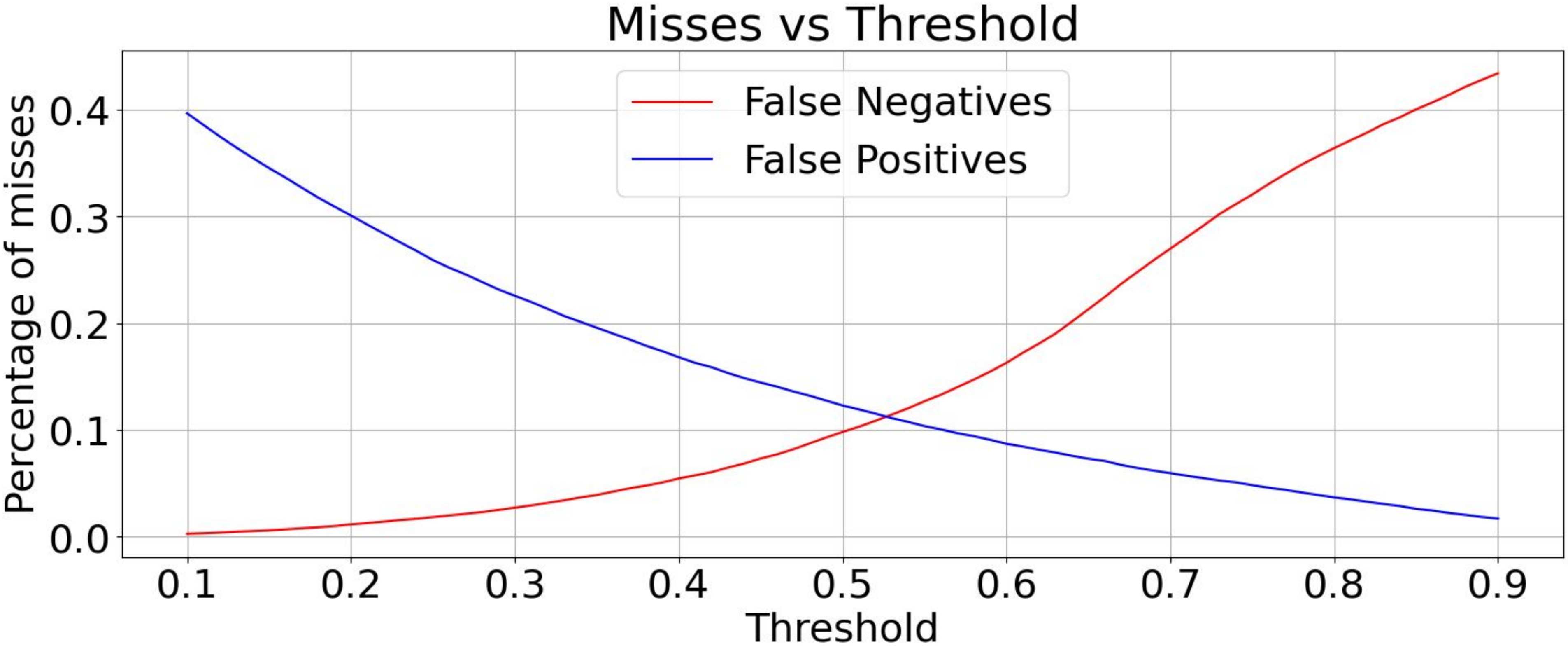
$$E = \frac{TN}{N_{EM}}$$

How to determine the threshold?

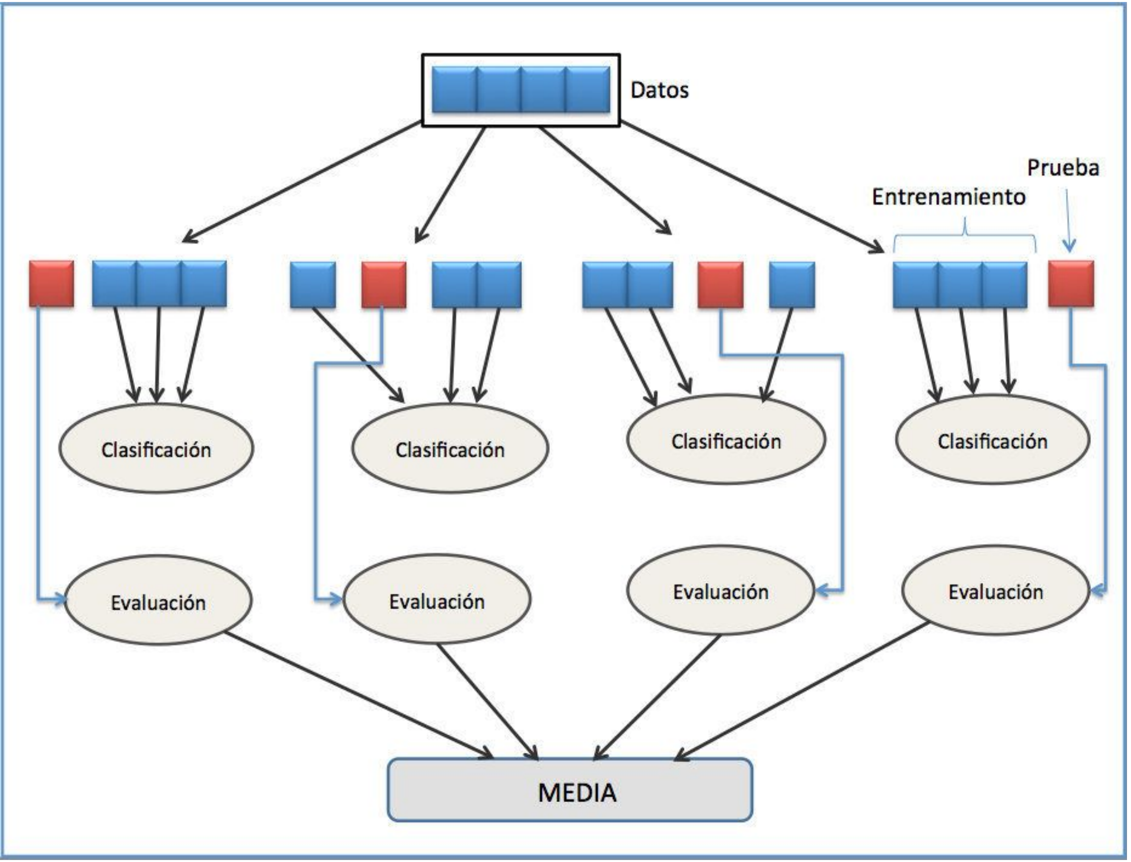


Threshold for maximum efficiency = 0.46

How to determine the threshold?



Cross validation



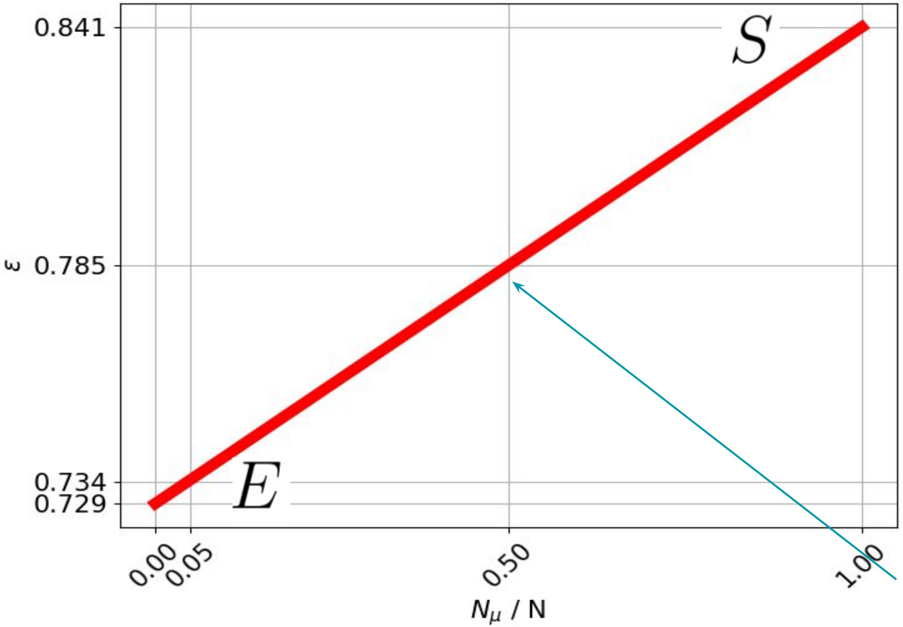
$$\epsilon = 0.785 \pm 0.003$$

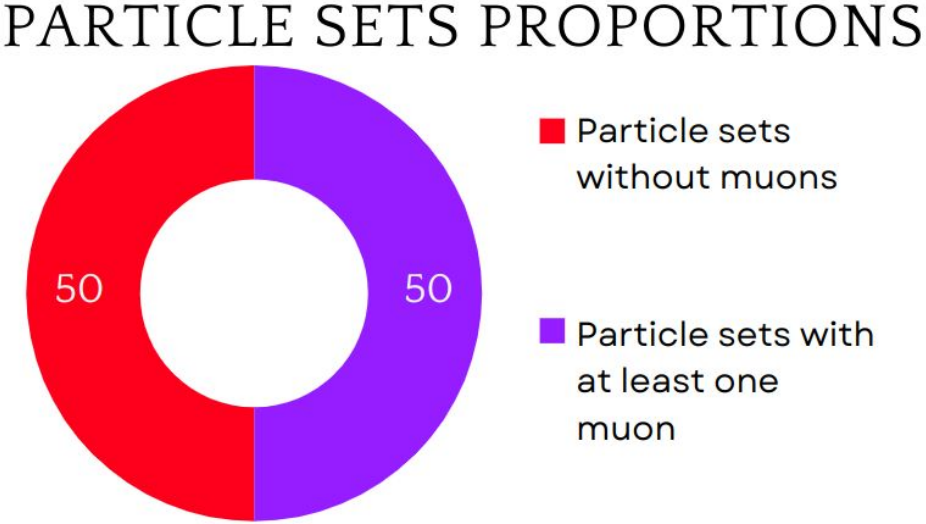
$$S = 0.841 \pm 0.003$$

$$E = 0.729 \pm 0.003$$

Variation of efficiency as a function of proportion

$$\epsilon = S \cdot \frac{N_\mu}{N} + E \cdot \frac{N_{EM}}{N}$$





$$\epsilon = S(0.5) + E(0.5) = 0.785 \pm 0.003$$

Summary

Summary

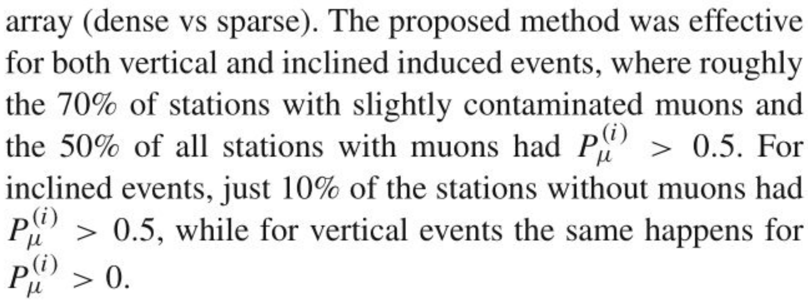
- It is possible to discriminate tracks with at least one muon from tracks without muons with an efficiency of ε = (78.5 ± 0.3)%, which forms the basis for discriminating gamma primaries in the ALPACA experiment.

- The efficiency is not a reliable metric because it changes as the proportion of tracks with muons varies.

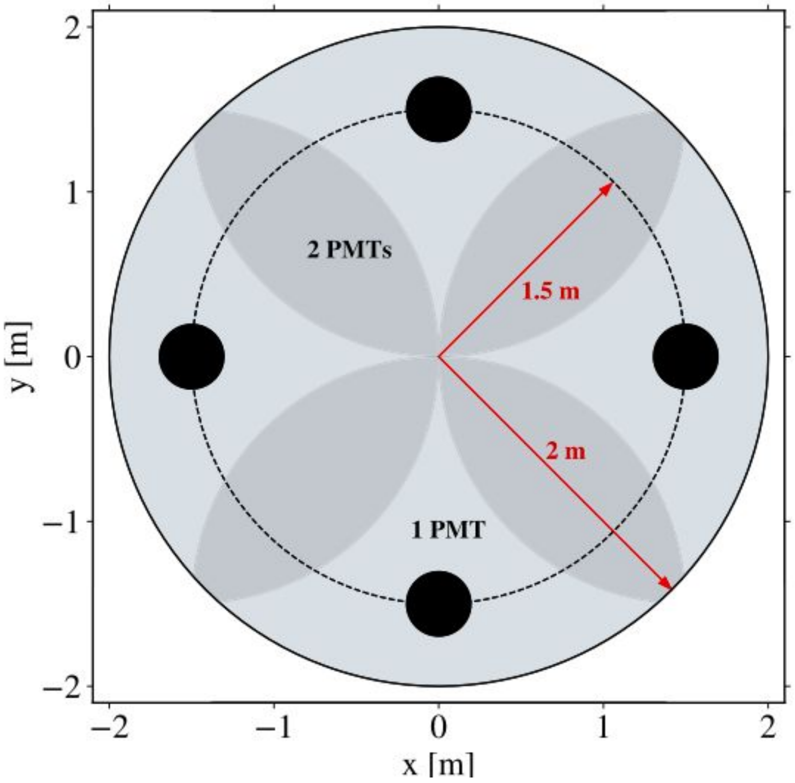
- The metrics to evaluate are sensitivity S = (84.1 ± 0.3)% and specificity E = (72.9 ± 0.3)%

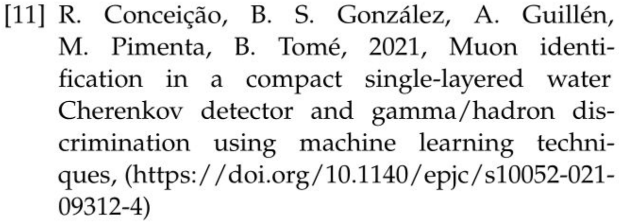
Summary

- We have a simple model and a low-cost detector with a single PMT, yet seemingly better results compared to other similar articles using more advanced techniques and a WCD with multiple PMTs.



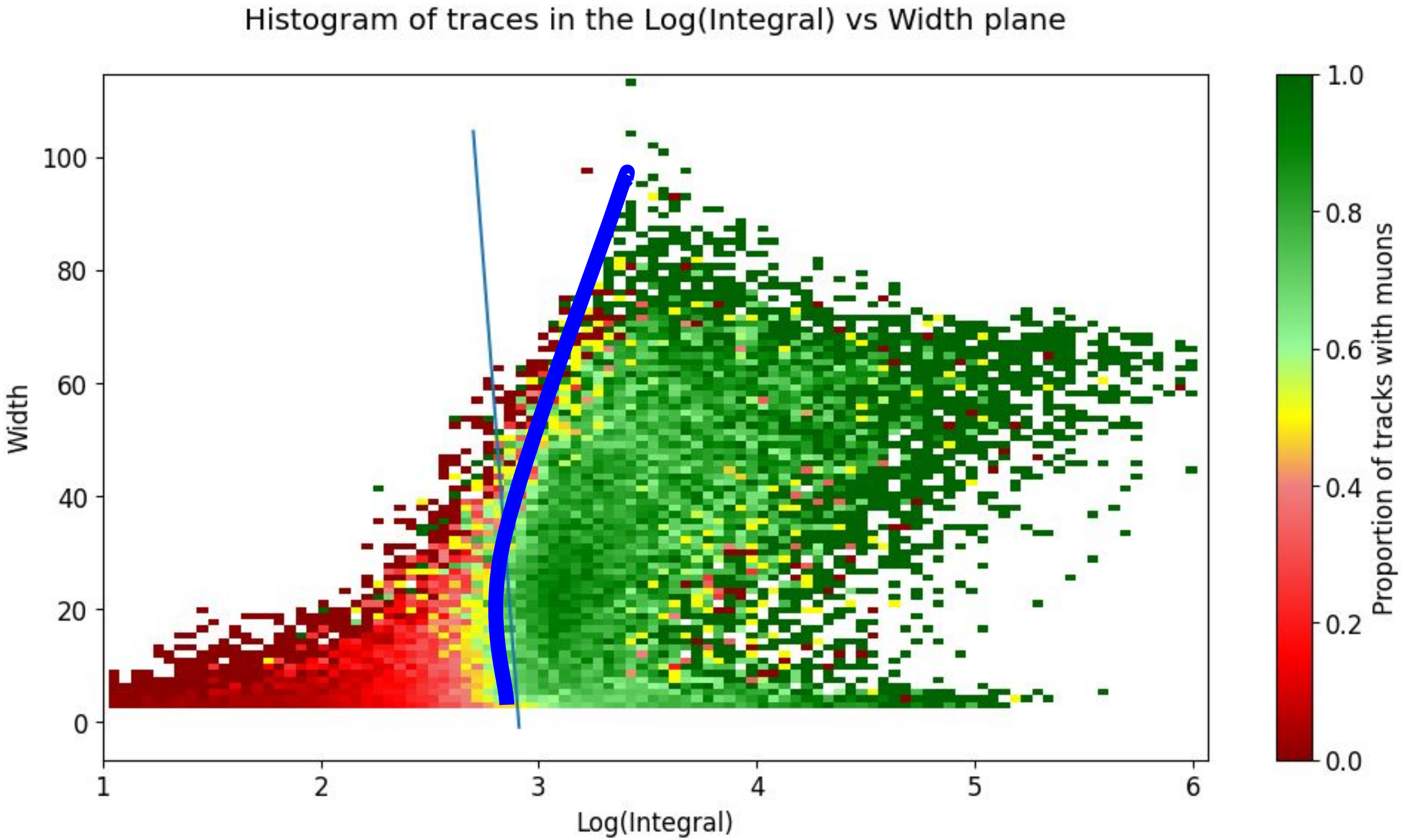
array (dense vs sparse). The proposed method was effective for both vertical and inclined induced events, where roughly the 70% of stations with slightly contaminated muons and the 50% of all stations with muons had $P_{\mu}^{(i)} > 0.5$. For inclined events, just 10% of the stations without muons had $P_{\mu}^{(i)} > 0.5$, while for vertical events the same happens for $P_{\mu}^{(i)} > 0$.





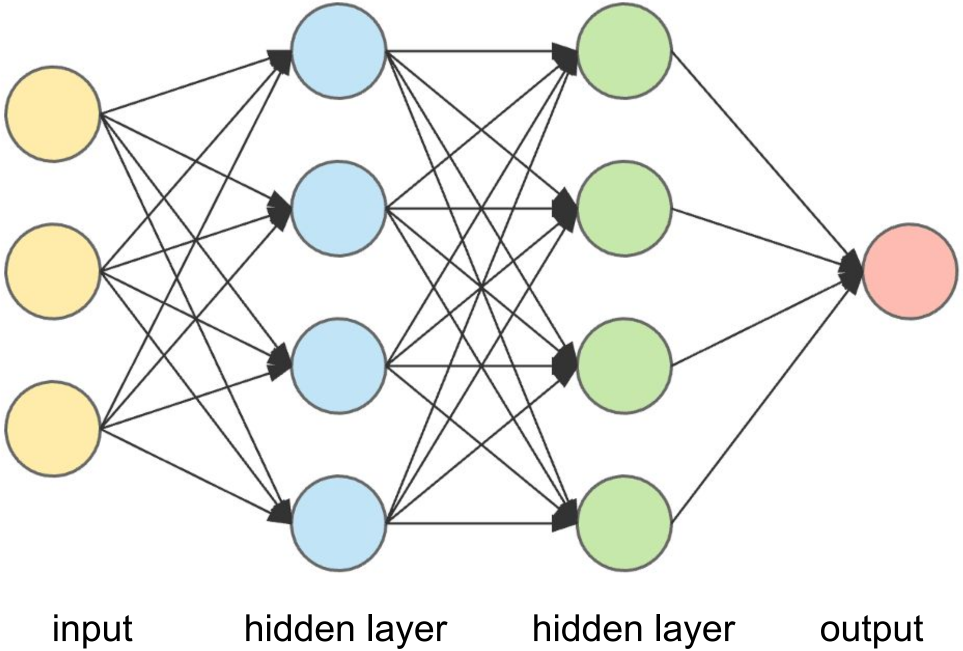
[11] R. Conceição, B. S. González, A. Guillén, M. Pimenta, B. Tomé, 2021, Muon identification in a compact single-layered water Cherenkov detector and gamma/hadron discrimination using machine learning techniques, (https://doi.org/10.1140/epjc/s10052-021-09312-4)

Perspectives

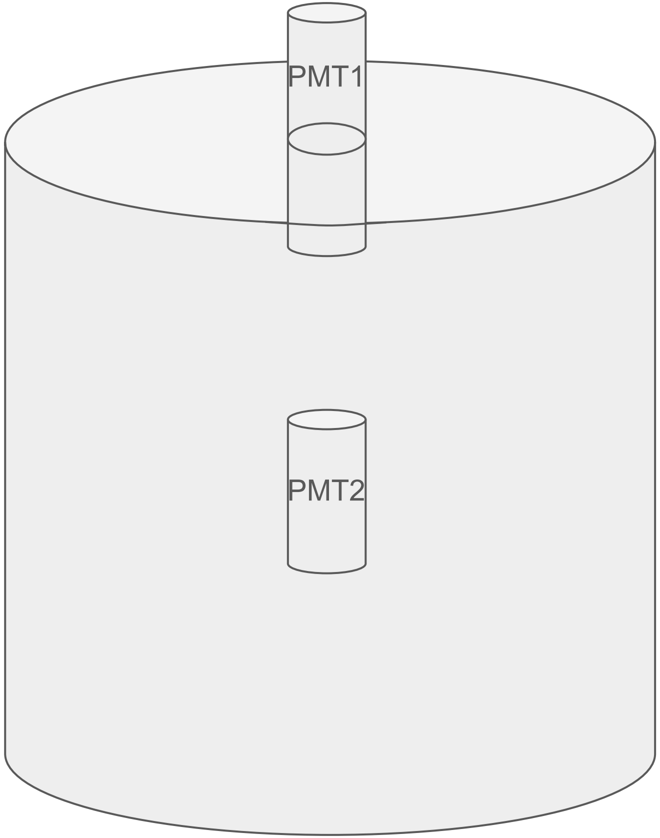


Perspectives

Neural networks



Add a second PMT



Thank you

Método de máxima verosimilitud

El método de máxima verosimilitud es una técnica fundamental en estadística que se utiliza para estimar los parámetros de un modelo probabilístico al encontrar los valores de los parámetros que maximizan la probabilidad de observar los datos que realmente se observaron.

$$L(\phi|y) = \prod_{i=1}^{m} f(y_i|\phi) \tag{6}$$

Donde llamaremos a $L(\phi)$ la función de verosimilitud y donde $f(y|\phi)$ es la función de densidad de probabilidad con parametros ϕ que ajustaremos al modelo

Para poder aplicar el método se deben cumplir las siguientes condiciones:

- Los errores deben ser independientes entre si.
- Los errores deben seguir una distribución $f(y|\phi)$
- Las variables independientes X deben ser independientes entre si.

Distribución de Bernoulli

La distribución de Bernoulli es una distribución de probabilidad discreta que modela el resultado de un experimento aleatorio que tiene dos posibles resultados: éxito (generalmente denotado como 1) o fracaso (generalmente denotado como 0). Es importante destacar que estos resultados deben ser mutuamente excluyentes y exhaustivos, es decir, solo puede ocurrir uno de los dos resultados posibles. Esta esta descrita de la siguiente forma:

$$f(y|p) = \begin{cases} p & \text{if } y = 1 \\ q = 1 - p & \text{if } y = 0 \end{cases}$$

o también:

$$f(y|p(\phi)) = (p(\phi))^y \cdot (1 - p(\phi))^{1-y} \tag{7}$$

donde $y = 0, 1$ y p es la función que describe la probabilidad de éxito para el modelo. Por ejemplo para una distribución normal, se usara su cumulativa.

Método de máxima verosimilitud

$$L(\phi|y) = \prod_{i=1}^{m} p(\phi)_i^y \cdot (1 - p(\phi))^{1-y_i}$$

$$l(\phi|y) = \sum_{i=1}^{m} [y_i \ln p(\phi) + (1 - y_i) \ln(1 - p(\phi))] \tag{8}$$

Distribución Logística

Densidad de probabilidad:

$$f(y_i, X_i|\theta, s) = \frac{e^{-\frac{(y-X\theta)}{s}}}{s(1+e^{-\frac{(y-X\theta)}{s}})^2}$$

Cumulativa:

$$p(X\theta) = \frac{1}{1+e^{-X\theta}}$$

Función de verosimilitud.

$$l(\theta|y, X) = \sum_{i=1}^{m}[y_i \ln \frac{1}{1+e^{-X\theta}} + (1-y_i)\ln(1-\frac{1}{1+e^{-X\theta}})]$$