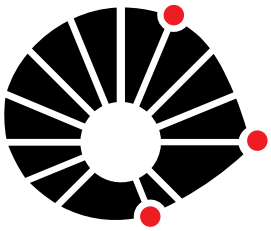


UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Física
Gleb Wataghin

JOÃO VÍTOR REGINATTO AKIM

Investigating the Performance of Machine Learning Algorithms for Muon Signal Identification at the Pierre Auger Observatory

Investigação do desempenho dos Algoritmos de Aprendizagem de Máquinas para Identificação de Sinais muonicos no Observatório Pierre Auger

Campinas

2023

João Vítor Reginatto Akim

Investigating the Performance of Machine Learning Algorithms for Muon Signal Identification at the Pierre Auger Observatory

Investigação do desempenho dos Algoritmos de Aprendizagem de Máquinas para Identificação de Sinais muonicos no Observatório Pierre Auger

Dissertação apresentada ao Instituto de Física Gleb Wataghin da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Física, na área de Física.

Dissertation presented to the Institute of Physics Gleb Wataghin of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Physics, in the area of Physics.

Supervisor: Carola Dobrigkeit Chinellato

Este trabalho corresponde à versão final da Dissertação defendida pelo aluno João Vítor Reginatto Akim e orientada pela Profa. Dra. Carola Dobrigkeit Chinellato.

Campinas

2023

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Física Gleb Wataghin
Maria Graciele Trevisan - CRB 8/7450

Ak52i

Akim, João Vítor Reginatto, 1998-
Investigating the performance of machine learning algorithms for muon signal identification at the Pierre Auger Observatory / João Vítor Reginatto Akim. – Campinas, SP : [s.n.], 2023.

Orientador: Carola Dobrigkeit Chinellato.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Física Gleb Wataghin.

1. Múons. 2. Observatório Pierre Auger. 3. Aprendizado de máquina. I. Chinellato, Carola Dobrigkeit, 1952-. II. Universidade Estadual de Campinas. Instituto de Física Gleb Wataghin. III. Título.

Informações Complementares

Título em outro idioma: Investigação do desempenho de algoritmos de aprendizagem de máquinas para identificação de sinais muônicos no Observatório Pierre Auger
Palavras-chave em inglês:
Muons
Pierre Auger Observatory
Machine learning
Área de concentração: Física
Titulação: Mestre em Física
Banca examinadora:
Carola Dobrigkeit Chinellato [Orientador]
Pedro Cunha de Holanda
Rogerio Menezes de Almeida
Data de defesa: 16-05-2023
Programa de Pós-Graduação: Física

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-8179-9747
- Currículo Lattes do autor: http://lattes.cnpq.br/1454566343114537

INSTITUTO DE FÍSICA
GLEB WATAGHIN

MEMBROS DA COMISSÃO EXAMINADORA DA DISSERTAÇÃO DE MESTRADO DO ALUNO JOÃO VÍTOR REGINATTO AKIM - RA 199923 APRESENTADA E APROVADA AO INSTITUTO DE FÍSICA GLEB WATAGHIN, DA UNIVERSIDADE ESTADUAL DE CAMPINAS, EM 16/05/2023.

COMISSÃO JULGADORA:

- Profa. Dra. Carola Dobrigkeit Chinellato – Presidente e orientadora (IFGW/UNICAMP)

- Prof. Dr. Pedro Cunha de Holanda (IFGW/UNICAMP)

- Dr. Rogerio Menezes de Almeida (Universidade Federal Fluminense)

OBS.: Ata da defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

CAMPINAS

2023

Acknowledgements

I thank my supervisor, Professor Carola, for all the support and teaching throughout the years. I also thank my family and friends, who even though they did not understand my research, they believed and were with me in difficult moments. I also thank my girlfriend for her companionship and for helping me with the images and formatting of this work.

"This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001".

Resumo

O objetivo desta tese é estudar o desempenho de algoritmos de aprendizagem de máquina para estimar a componente muônica do sinal medido nas estações do detector de superfície do Observatório Pierre Auger. A componente muônica de um chuveiro atmosférico está altamente correlacionada com a massa da partícula primária que deu origem a este chuveiro. O conhecimento da massa dos raios cósmicos permite aos cientistas estudar melhor os mecanismos de propagação e aceleração destas partículas e formular melhor modelos de interações hadrónicas em energias que os aceleradores artificiais não são capazes de atingir. O tipo de algoritmo de aprendizagem de máquina utilizado foi o das redes neurais recorrentes. Estes algoritmos foram utilizados para prever a componente muônica em cada intervalo medido de tempo. Os resultados mostram que os algoritmos de aprendizagem de máquina podem estimar com precisão o sinal muônico.

Palavras-chave: Componente muônica, Observatório Pierre Auger, aprendizagem de máquina.

Abstract

The objective of this thesis is to study the performance of machine learning algorithms for estimating the muon component of the signal measured at the surface detector stations of the Pierre Auger Observatory. The muon component of an atmospheric shower is highly correlated to the primary particle mass. The knowledge of the cosmic ray mass enables scientists to better study the propagation and acceleration mechanisms of these particles and to better formulate hadronic interaction models at energies that human-made accelerators are not able to generate. The type of machine learning algorithm used was a recurrent neural network. It was used to predict the muon component at each time bin. The results show that machine learning algorithms can accurately estimate the muon signal.

Keywords: Muon component, Pierre Auger Observatory, machine learning.

List of Figures

Figure 1 – Changes in ionization with altitude. Left panel: Data from the final ascension of Hess (1912), which carried two ionization chambers. Right panel: Data from Kolhörster's ascension (1913, 1914). 18
Figure 2 – Cosmic-ray energy spectrum. 19
Figure 3 – Comparison between nuclear abundances in low-energy cosmic rays and in the Solar System. Normalized to C=100. 20
Figure 4 – Representation of iterations of a shower cascade according to the Heitler model. 22
Figure 5 – Representation of iterations of a shower cascade according to the Heitler-Matthews model. 22
Figure 6 – A schematic view of the Pierre Auger Observatory where each dot represents an SD station. The FD buildings are shown with their respective names, and the lines indicate each telescope's individual field of view. . 24
Figure 7 – A picture of an SD station, highlighting its principal components. . . . 25
Figure 8 – Charge histogram of the SD station for atmospheric muons (open histogram). Charge histogram of the SD station when just vertical muons are allowed. 26
Figure 9 – Picture of Los Leones building with the telescopes on display. 27
Figure 10 – Representation of the coordinate system used on the Pierre Auger Observatory. 28
Figure 11 – Diagram of the air-shower development considering the shower front as a plane. 29
Figure 12 – Lateral distribution function for an event recorded at the Pierre Auger Observatory. The signal at 1000 m from the shower core is highlighted. 30
Figure 13 – **a**: The equations utilized for carrying out the forward propagation in a neural network with two hidden layers and a single output layer. **b**: The procedures for computing the backward pass involve computing the error derivatives at each hidden layer with respect to the output of each neuron. The details of the procedures are explained in the text. . 34
Figure 14 – Graphic representation of the ReLU function for values between −10 and 10. 35
Figure 15 – An LSTM cell diagram showing the gates that determine how the inputs are combined to produce the output. 36
Figure 16 – Schematic representation of the modules used to prepare the static inputs to be fed for the encoder. 38

Figure 17 – Schematic representation of the first model used. It consists of two LSTM layers in the encoder and one LSTM layer on the decoder. . . . 39
Figure 18 – Schematic representation of the second model used. It consists of one LSTM layer in the encoder and one LSTM layer on the decoder. 40
Figure 19 – Schematic representation of the third model used. It consists of two LSTM layers in the encoder and one LSTM layer on the decoder. The last cell state and hidden state from one LSTM layer are passed through a projector, serving as inputs for the next LSTM layer. 41
Figure 20 – Schematic representation of the fourth model used. It consists of one LSTM layer in the encoder and one LSTM layer on the decoder. The last cell state and hidden state from one LSTM layer are passed through a projector, so they serve as inputs for the next LSTM layer. 42
Figure 21 – Distributions of stations in relation to the energy. 43
Figure 22 – Distributions of stations in relation to the arrival angle. 44
Figure 23 – Distributions of stations in relation to the distance to the core on the plane perpendicular to the shower plane. 44
Figure 24 – Loss values as a function of epochs of training and validation. 45
Figure 25 – Distribution of the simulated and predicted muon signals for every station in the dataset, separated by the primary cosmic-ray composition. 46
Figure 26 – Distribution of the difference between simulated and predicted muon signals for every station in the test dataset, separated by the primary cosmic-ray composition. 47
Figure 27 – Mean value (27a) and standard deviation (27b) of the difference between the predicted and simulated values as a function of the logarithm of the energy. Mean value (27c) and standard deviation (27d) of the difference between the predicted and simulated values divided by the total signal as a function of the logarithm of the energy. 48
Figure 28 – Mean value (28a) and standard deviation (28b) of the difference between the predicted and simulated values as a function of the secant of the arrival angle. Mean value (28c) and standard deviation (28d) of the difference between the predicted and simulated values divided by the total signal as a function of the secant of the arrival angle. 49
Figure 29 – Integral of the predicted muon signal as a function of the integral of the simulated muon signal. The black line corresponds to the linear fit of the points, and the red line corresponds to the ideal case. 50

Figure 30 – 30a: Mean and standard deviation of the difference between simulated and predicted muon signals, for every station with energies and zenith angles specified in the box. 30b: Standard deviation of the difference between simulated and predicted muon signals, for every station separated by the primary cosmic-ray composition. 51
Figure 31 – 31a: Distribution of $\widehat{S}^{\mu}$ and S^{μ} for all stations in the test dataset. 31b: Distribution for S^{μ} for all stations in the test dataset with an atmospheric shower initiated by a proton. 51
Figure 32 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an electromagnetic-dominated signal. The total simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line. 52
Figure 33 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an muon-dominated signal. The total simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line. 53
Figure 34 – Distribution of the simulated and predicted muon signals for every station in the dataset, separated by the primary cosmic-ray composition. 58
Figure 35 – Distribution of the difference between simulated and predicted muon signals for every station in the test dataset, separated by the primary cosmic-ray composition. 59
Figure 36 – Mean value (36a) and standard deviation (36b) of the difference between the predicted and simulated values as a function of the logarithm of the energy. Mean value (36c) and standard deviation (36d) of the difference between the predicted and simulated values divided by the total signal as a function of the logarithm of the energy. 60
Figure 37 – Mean value (37a) and standard deviation (37b) of the difference between the predicted and simulated values as a function of the secant of the arrival angle. Mean value (37c) and standard deviation (37d) of the difference between the predicted and simulated values divided by the total signal as a function of the secant of the arrival angle. 61
Figure 38 – Integral of the predicted muon signal as a function of the integral of the simulated muon signal. The black line corresponds to the linear fit of the points, and the red line corresponds to the ideal case. 62

Figure 39 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an eletromagnetic-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line. 62
Figure 40 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an muon-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line. . . 63
Figure 41 – Distribution of the simulated and predicted muon signals for every station in the dataset, separated by the primary cosmic-ray composition. 64
Figure 42 – Distribution of the difference between simulated and predicted muon signals for every station in the test dataset, separated by the primary cosmic-ray composition. 65
Figure 43 – Mean value (43a) and standard deviation (43b) of the difference between the predicted and simulated values as a function of the logarithm of the energy. Mean value (43c) and standard deviation (43d) of the difference between the predicted and simulated values divided by the total signal as a function of the logarithm of the energy. 66
Figure 44 – Mean value (44a) and standard deviation (44b) of the difference between the predicted and simulated values as a function of the secant of the arrival angle. Mean value (44c) and standard deviation (44d) of the difference between the predicted and simulated values divided by the total signal as a function of the secant of the arrival angle. 67
Figure 45 – Integral of the predicted muon signal as a function of the integral of the simulated muon signal. The black line corresponds to the linear fit of the points, and the red line corresponds to the ideal case. 68
Figure 46 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an eletromagnetic-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line. 68
Figure 47 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an muon-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line. . . 69
Figure 48 – Distribution of the simulated and predicted muon signals for every station in the dataset, separated by the primary cosmic-ray composition. 70

Figure 49 – Distribution of the difference between simulated and predicted muon signals for every station in the test dataset, separated by the primary cosmic-ray composition. 71
Figure 50 – Mean value (50a) and standard deviation (50b) of the difference between the predicted and simulated values as a function of the logarithm of the energy. Mean value (50c) and standard deviation (50d) of the difference between the predicted and simulated values divided by the total signal as a function of the logarithm of the energy. 72
Figure 51 – Mean value (51a) and standard deviation (51b) of the difference between the predicted and simulated values as a function of the secant of the arrival angle. Mean value (51c) and standard deviation (51d) of the difference between the predicted and simulated values divided by the total signal as a function of the secant of the arrival angle. 73
Figure 52 – Integral of the predicted muon signal as a function of the integral of the simulated muon signal. The black line corresponds to the linear fit of the points, and the red line corresponds to the ideal case. 74
Figure 53 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an eletromagnetic-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line. 74
Figure 54 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an muon-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line. . . 75

List of Tables

Table 1 – Number of initial particles in each dataset 43

List of abbreviations and acronyms

UNICAMP	Universidade Estadual de Campinas
IFGW	Instituto de Física Gleb Wataghin
UHECRs	ultra-high energy cosmic rays
EAS	extensive air shower
SD	surface detector
FD	fluorescence detector
PMT	photomultiplier
VEM	vertical equivalent muon
RNN	recurent neural network
LSTM	Long Short Term Memory
FC	Fully connected layer

Contents

1 Introduction . . . 17
1.1 Cosmic Rays . . . 17
1.1.1 Energy spectrum . . . 18
1.1.2 Composition . . . 20
1.1.2.1 Air Showers . . . 21
1.1.2.1.1 Electromagnetic component . . . 21
1.1.2.2 Hadronic component . . . 22
2 The Pierre Auger Observatory . . . 24
2.1 Surface detector . . . 24
2.2 Fluorescence detector . . . 26
2.3 Angular convention . . . 27
2.4 Event reconstruction . . . 28
2.4.1 *Herald* framework . . . 30
2.4.1.1 Lateral distribution function . . . 30
2.4.2 *Observer* reconstruction . . . 31
2.4.2.1 Lateral distribution function . . . 31
2.4.3 Energy calibration . . . 31
3 The method . . . 32
3.1 The input . . . 32
3.2 Neural network . . . 33
3.2.1 Fully connected layer . . . 34
3.2.2 Long short-term memory . . . 35
3.3 Models used . . . 38
3.3.1 First model . . . 38
3.3.2 Second model . . . 39
3.3.3 Third model . . . 40
3.3.4 Fourth model . . . 41
3.4 Dataset . . . 42
4 Results . . . 46
4.1 Differences in the traces . . . 46
4.2 Comparison . . . 50
4.3 Trace examples . . . 52
5 Conclusions . . . 54

BIBLIOGRAPHY . . . 55

Appendix **57**

APPENDIX A Models 1, 2, 4 figures **58**

A.1 First model 58

A.2 Second model 64

A.3 Fourth model 70

1 Introduction

The discovery of cosmic rays by Victor Hess in 1912 sparked great interest and investigation. Despite being researched for nearly a century, the true nature of these rays remains a mystery. This is particularly true for ultra-high energy cosmic rays (UHECRs), whose source and mechanism behind their acceleration have yet to be uncovered.

The Pierre Auger Collaboration was established to investigate UHECRs. Over the last decade, the Collaboration has made notable progress in its research. One such advancement was the discovery of a dipole pattern in the arrival direction distribution of UHECRs on Earth, which points to a direction 125° away from the center of our galaxy [1]. This anisotropy indicates that UHECRs may have an extragalactic origin. However, the exact source of these cosmic rays remains unknown despite these advancements.

A different study has found that cosmic rays with higher energy tend to have a larger mass than those with lower energy [2]. As cosmic rays are charged particles, their trajectory is affected by magnetic fields on their journey to Earth, causing their arrival direction to deviate from their source. The degree of deflection is determined by the rigidity of the particle, which is defined as the momentum of the particle divided by its charge. For example, a proton would experience less deflection than an iron nucleus of the same energy. By focusing on lighter nuclei, it may be possible to gain a deeper understanding of UHECRs and differentiate between heavy and light particles. This work aims to provide a way to make this differentiation.

1.1 Cosmic Rays

The study of cosmic rays dates back to the early 1900s and started with the observation that electroscopes would discharge even when positioned far from radioactive materials. At the time, researchers were uncertain whether the radioactivity originated from Earth or the sky. Subsequently, numerous experiments were conducted to measure atmospheric ionization at various altitudes to shed light on the issue.

The Austrian-American physicist Victor Hess conducted balloon flights in 1912 to measure atmospheric ionization at different altitudes, reaching as high as 5 km [3]. His experiments discovered that ionization levels increased with altitude, leading him to conclude that the ionization was caused by radiation originating from space.

In 1913, the German physicist Werner Kolhörster conducted similar balloon flights, reaching altitudes of around 9 km [4]. He obtained similar results as Hess, as demonstrated in Figure 1.

In their experiments, Hess and Kolhörster measured not only cosmic rays but also particles generated from air showers created by cosmic rays interacting with the atmosphere. An air shower refers to the chain reaction of particles and photons produced when a cosmic ray collides with a nucleus from an atmospheric molecule. The formation and evolution of air showers will be discussed in further detail in a later section, 1.1.2.1.

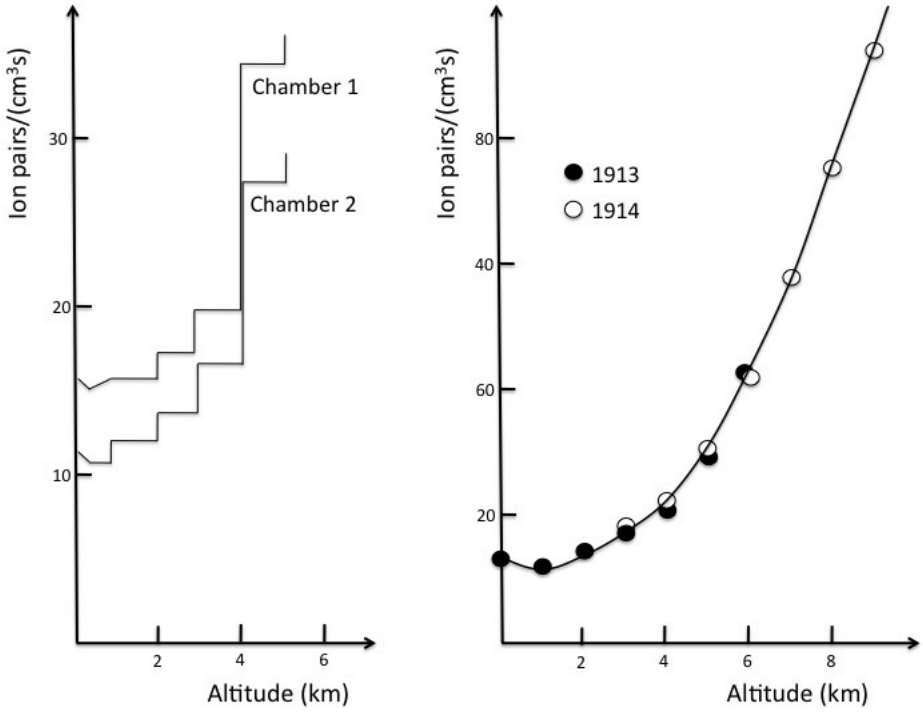


Figure 1 – Changes in ionization with altitude. Left panel: Data from the final ascension of Hess (1912), which carried two ionization chambers. Right panel: Data from Kolhörster's ascension (1913, 1914). Taken from [5]

1.1.1 Energy spectrum

The energy spectrum of cosmic rays spans a wide range, ranging from 10^9 eV to 10^{21} eV, and can be approximated by a power law. Two distinct parts of the spectrum stand out: the 'knee' at 10^{15} eV and the 'ankle' at 10^{18} eV. The complete spectrum is depicted in Figure 2.

Due to the broadness of the cosmic-ray energy spectrum, different measurement techniques are required to study each portion. It is common to divide the spectrum into two groups: before and after the 'knee.' Particles with lower energy than the 'knee' are more abundant and can be measured by detectors in space. However, particles after the 'knee' are less frequent, and it is more effective to measure the air showers they initiate when entering the atmosphere. An exciting aspect of the cosmic-ray spectrum is the particles with energies less than 10^9 eV. In Figure 2, it is evident that these particles do not conform to a power law. This phenomenon is called Solar Modulation and is caused

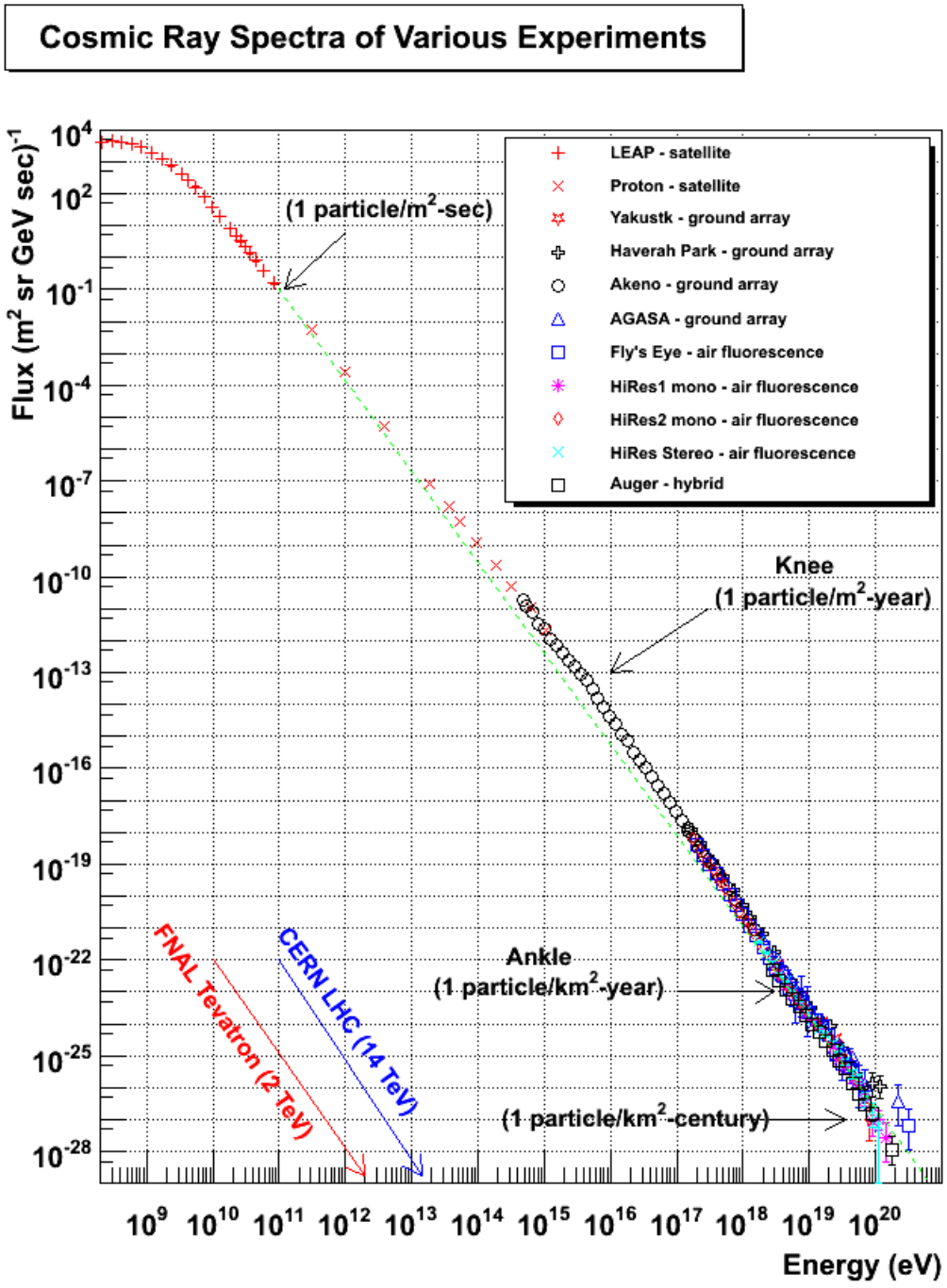


Figure 2 – Cosmic-ray energy spectrum. Taken from [6]

by the Solar Wind, which deflects cosmic rays with energies lower than 10^9 eV. This study will concentrate on UHECRs, which are cosmic rays with energies greater than 10^{18} eV.

1.1.2 Composition

The composition of cosmic rays holds vital information about their origin and how they were accelerated. To gather this information, John Alexander Simpson, an American physicist, collected data on the composition of cosmic rays and compared it to the abundance of elements in the Solar System [7]. The comparison can be seen in Figure 3.

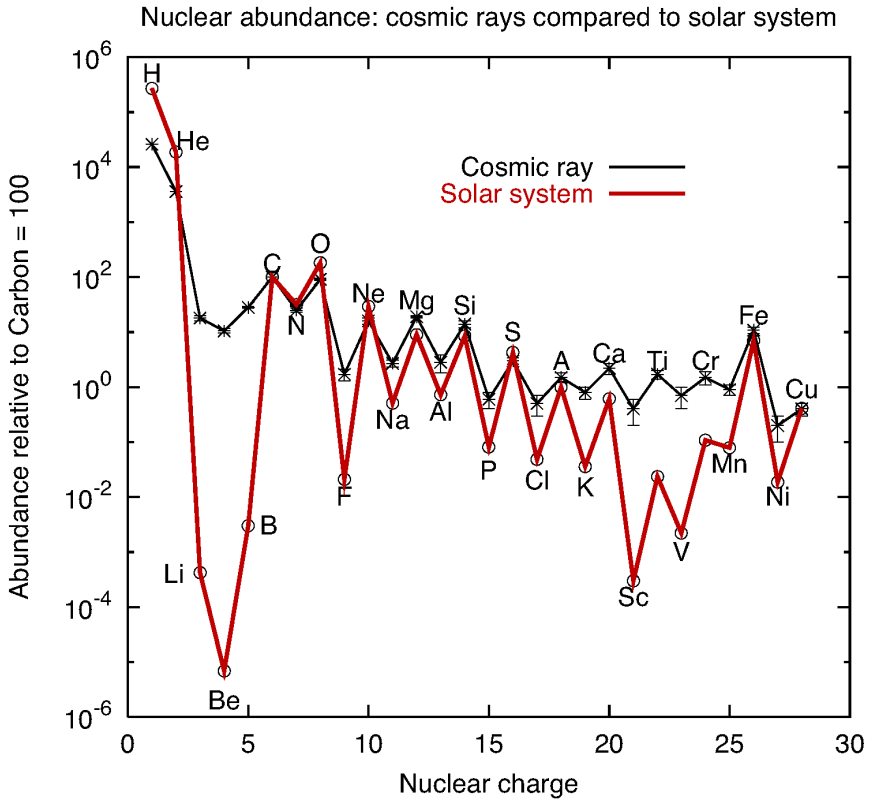


Figure 3 – Comparison between nuclear abundances in low-energy cosmic rays and in the Solar System. Normalized to C=100. Taken from [8]

The similarities between both compositions are noticeable, but there are significant divergencies in two cases:

- For lighter elements, such as lithium, beryllium, and boron, their abundance in cosmic rays is higher than in the Solar System.
- For heavier elements, such as scandium, titanium, vanadium, chromium, and manganese, their abundance is also higher in cosmic rays than in the Solar System.

These differences could be explained through the process of spallation, which involves the breaking apart of nuclei like carbon and iron. Simpson's comparison was based on data from cosmic rays with low energy (less than 10^{15} eV) [7]. The composition of higher-energy cosmic rays, such as UHECRs, is different. Using various techniques, the Pierre Auger Collaboration has found that UHECRs are mainly composed of protons at energies of 10^{18} eV and show an increasing mass at higher energies[2].

1.1.2.1 Air Showers

The interaction of a cosmic ray with air molecules can result in a cascade of particles known as an air shower. If the cosmic-ray energy exceeds 10^{14} eV, the air shower created can reach the ground level and is referred to as an extensive air shower (EAS). The size of an EAS increases rapidly with the energy of the initial cosmic ray, for instance, at 10^{15} eV, an air shower would consist of 10^6 particles and cover an area of 10^4 m^2 on the ground, while at 10^{20} eV, it would consist of approximately 10^{11} particles and cover an area of 10 km^2 [9]. Pierre Auger and his team established the existence of EAS in 1939 [10].

Due to the limited number of UHECRs, direct measurement of these high-energy cosmic rays is not feasible. The Pierre Auger Collaboration demonstrated this in their experiments, as UHECRs above $10^{18.5}$ eV are found only once per square kilometer per year. To study these rare events, researchers must instead observe the EAS they produce. Air showers have two major components, an electromagnetic component and a hadronic component. The electromagnetic component produces photons, electrons, and positrons, and the hadronic component produces mostly pions, which in turn decay and produce electrons, positrons, muons, and neutrinos.

1.1.2.1.1 Electromagnetic component

A simple model used to understand the development of the electromagnetic component is the Heitler model [11]. This model assumes that a particle (electron, positron, or photon) with energy E_0 travels a fixed length X_0 until it interacts, creating a pair of particles or a particle and a photon by the bremsstrahlung process, each with half of the available energy. After n repetitions, there are $N = 2^n$ particles, each with energy $E = E_0/N$. This process continues until the energy E is too low to create new pairs. At that point, the air shower reaches its peak, and the distance traveled up until then is referred to as the atmospheric depth of the shower maximum, $X_{\mathrm{MAX}} = n_{\mathrm{MAX}} X_0$. An illustration of this process can be seen in Figure 4.

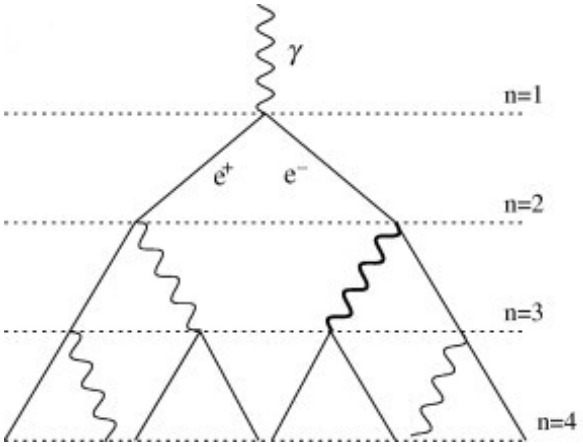


Figure 4 – Representation of iterations of a shower cascade according to the Heitler model. Taken from [12]

1.1.2.2 Hadronic component

Matthews, using a similar approach to Heitler's model, modeled a shower initiated by a proton [12]. In his model, a proton travels a fixed distance until it interacts, producing N_{ch} charged pions and $1/2N_{ch}$ neutral pions. Each pion created has one-third of the available energy. It is important to note that this process creates more than just pions. However, the kaons created rapidly decay into pions, muons, or electrons and will not be considered in this model. An illustration to this model is shown in Figure 5.

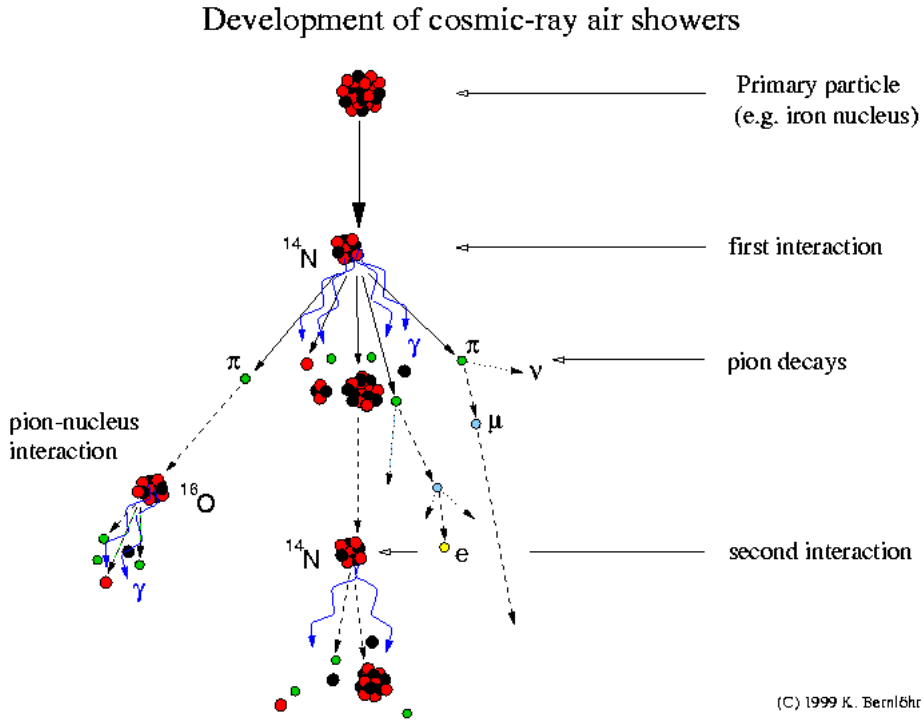


Figure 5 – Representation of iterations of a shower cascade according to the Heitler-Matthews model. Taken from [13]

Similarly, the neutral pions rapidly decay into muons and electrons, which start new electromagnetic sub-showers. After n repetitions, there are $N_\pi = N_{ch}^n$ particles, each with energy $E = \frac{E_0}{(2/3N_{ch})^n}$.

In the later stage of the particle shower, charged pions that can no longer produce new pions begin to decay into muons and neutrinos. The amount of muons created is calculated through $N_\mu = N_{ch}^{n_{max}}$, with n_{max} denoting the maximum number of interactions before the energy of the charged pions is no longer enough to generate new particles. When this occurs, the only remaining particles that can reach the ground are electrons, positrons, neutrinos, and high-energy muons. Meanwhile, muons with low energy will decay into electrons and neutrinos.

The number of muons in a shower created by a cosmic ray with mass number A can be estimated using the following equation:

$$N_\mu^A = N_\mu^p A^{0.15}. \tag{1.1}$$

This calculation assumes that the shower is equivalent to multiple independent proton showers, each with an energy of E_0/A.

This basic model demonstrates that the quantity of muons depends solely on the mass of the primary cosmic particle. As a result, it is feasible to identify the particle's mass just by analyzing the number of muons observed in the air shower.

The purpose of this study is to develop a technique for separating the signal of the muon component in the total signal registered by the surface detector stations of the Pierre Auger Observatory. Doing so allows differentiating between air showers initiated by lighter elements versus those initiated by heavier particles.

2 The Pierre Auger Observatory

The Pierre Auger Observatory is a product of a unique partnership of researchers from 18 countries. This partnership is known as the Pierre Auger Collaboration. The observatory construction in Malargüe, Argentina, began in 2001 and finished in 2008. The Collaboration projected the observatory to study cosmic rays with energy superior to 10^{18} eV, and to that goal, the observatory covers approximately 3000 km^2. Additionally, it uses two types of detectors: a large surface detector (SD) and a fluorescence detector (FD).This hybrid scheme allows for part of the events (an event is a detected air shower) to be detected by both detector types simultaneously, which permits energy calibration of the SD.

2.1 Surface detector

The Pierre Auger Observatory has over 1600 water-Cherenkov detectors arranged as an array on a triangular grid with 1500 m spacing. Additional 60 detectors separated by 750 m form an infilled array. A scheme of the entire observatory is shown in Figure 6.

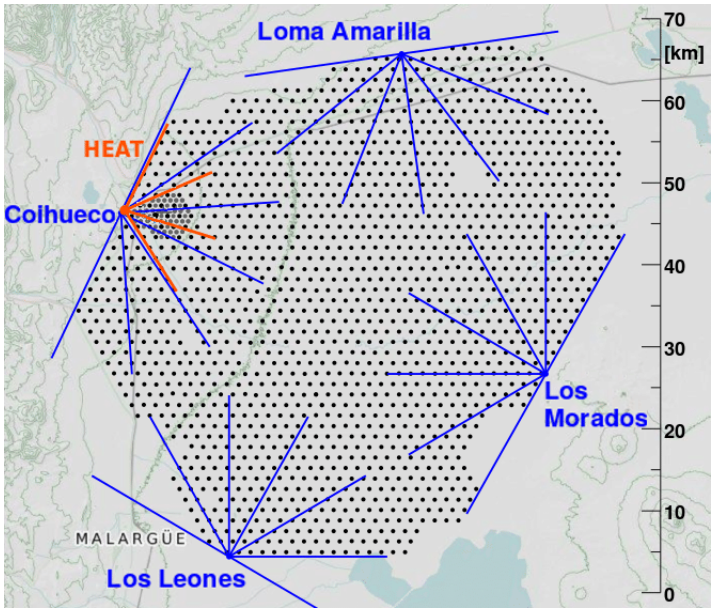


Figure 6 – A schematic view of the Pierre Auger Observatory where each dot represents an SD station. The FD buildings are shown with their respective names, and the lines indicate each telescope's individual field of view. Taken from [14]

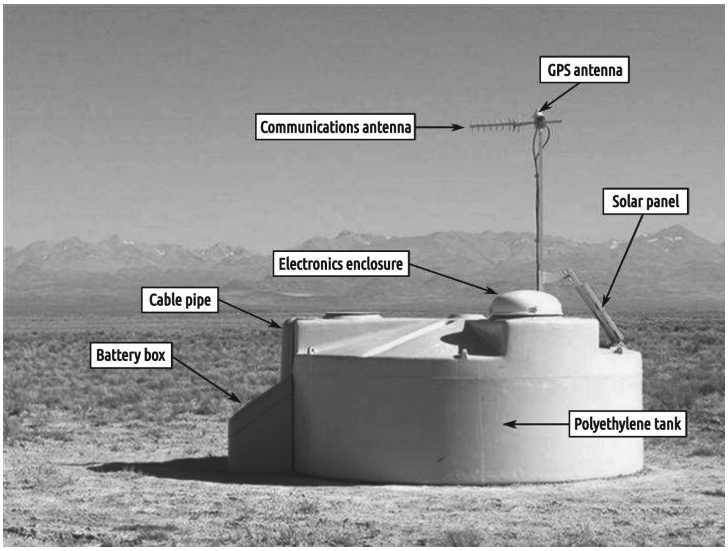


Figure 7 – A picture of an SD station, highlighting its principal components. Taken from [15]

Each station in the Surface Detector (SD) comprises a tank with a diameter of 3.6 m and reflective walls. The tank holds 12,000 l of ultra-pure water, along with three photomultipliers (PMT), a GPS receiver, a radio transceiver, and a solar power system with batteries to power the electronics. A single station is depicted in Figure 7. The PMT counts the number of photons produced by the Cherenkov process. This process occurs when shower particles cross the water at supra-luminal speeds. The recorded signal is then transformed from a count into vertical equivalent muons (VEM). The unit VEM is defined as the signal registered by a station when a muon traverses it vertically through its center. This conversion requires calibration to be performed accurately. The calibration is performed with the measurement of atmospheric muons [16]. They provide an excellent method for measuring 1 VEM in terms of the PMT charge signal because they pass through stations with a frequency of approximately 2500 Hz. However, an SD station cannot select from these muons only those that pass vertically, but the charge distribution they produce has a peak that equals 1.09 VEM. Figure 8 shows an example of a charge histogram. The passage of vertical muons creates the second peak of the open histogram. The leftmost peak of the open histogram exists due to low-energy and clipping muons. The hatched histogram is the charge distribution when only the vertical muons enter the station.

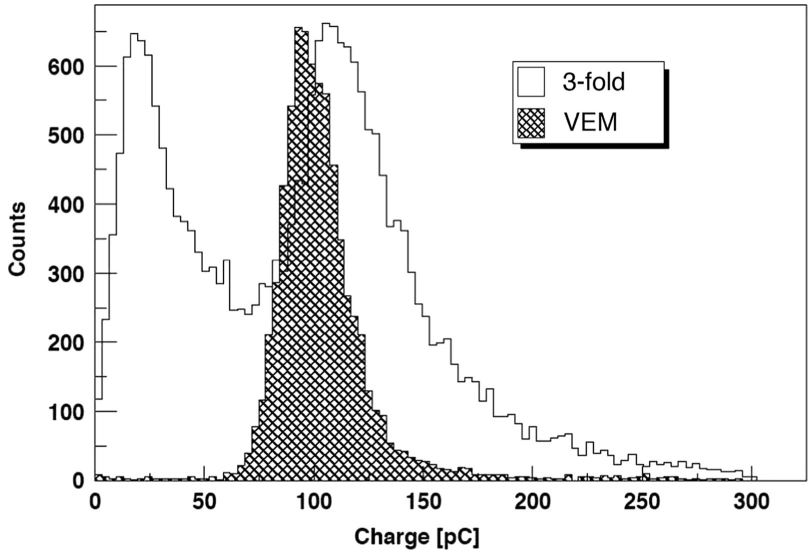


Figure 8 – Charge histogram of the SD station for atmospheric muons (open histogram). Charge histogram of the SD station when just vertical muons are allowed. Taken from [15].

In this work, the prediction of the muon signal is based only on SD measurements.

2.2 Fluorescence detector

The Fluorescence Detector (FD) consists of 27 telescopes located at five different sites: Loma Amarilla, Los Morados, Los Leones, Coihueco, and HEAT (High-Elevation Auger Telescopes). These locations are indicated in Figure 6. The first four locations have six telescopes each, while HEAT has only three. Each telescope has a field of view of $30^\circ \times 30^\circ$, allowing for 180° coverage in azimuthal angle by combining six telescopes. Figure 9 shows the Los Leones facility with the telescopes on display.

Figure 9 – Picture of Los Leones building with the telescopes on display. Taken from [15].

The telescopes at HEAT can observe up to 58° as the facility can incline. This higher viewing angle allows the detection of cosmic rays with lower energy down to 10^{17} eV. All the telescopes monitor the atmosphere above the SD, enabling simultaneous measurement of events by both detectors. The FD measures the intensity of fluorescence light emitted by the nitrogen molecules excited by the air shower particles. Since the light emitted by the shower is proportional to the collisional energy, this approach provides a near-calorimetric measurement of the cosmic-ray energy. However, due to the sensitivity of the telescopes, they can only be operated on dark, clear nights, which represents approximately 15% of the time.

2.3 Angular convention

The axes of the coordinate system are defined as follows: z is normal to the observatory pointing to the zenith, x is tangent to the parallel that crosses through the observatory and points to the East, and y is tangent to the meridian that crosses through the observatory and points to the North.The coordinate system is shown in figure 10.The angles used are the local azimuthal ϕ and zenithal θ. The angle ϕ is the angle formed with x, and the angle θ is the angle formed with z. It is important to note that the time of measurement is also necessary because the coordinate system is local and rotates with Earth.

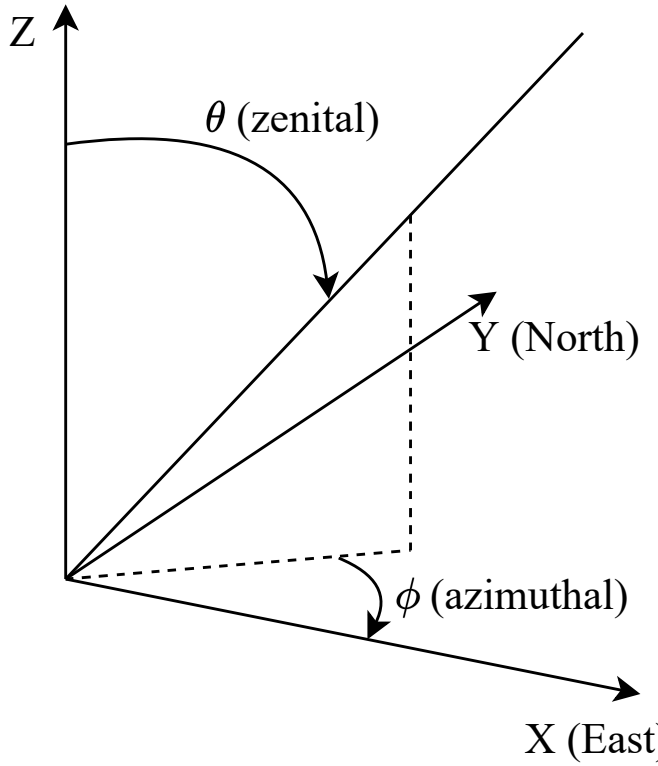


Figure 10 – Representation of the coordinate system used on the Pierre Auger Observatory.

2.4 Event reconstruction

Since this work will focus on analyzing SD station signals, only the reconstruction of SD events will be explained. The *seed* reconstruction is the first step in reconstructing an SD event. This first step helps to identify and exclude accidental stations and provides the initial estimates for the proper event reconstruction. For the *seed* reconstruction, the first step is finding the point $\vec{x}_b$, which is the signal-weighted center-of-mass of stations in an event. This point is also used for the first impact position of the shower core on the ground $\vec{x}_{gr}$. Now, assuming that the shower front moves as a plane perpendicular to the shower axis at the speed of light, as shown in figure 11, it is possible to obtain the time $t(\vec{x})$ when the shower front passes through the point $\vec{x}$ with the formula:

$$ct(\vec{x}) = ct_b - \hat{a} \cdot (\vec{x} - \vec{x}_b), \tag{2.1}$$

where t_b is the time when the shower plane passes through the point $\vec{x}_b$ and $-\hat{a}$ is the direction of the shower propagation. This equation gives the first estimate of the cosmic-ray direction.With this first approximation, there are two ways to continue: the *Herald* framework and the *Observer* reconstruction, each of which uses different assumptions and different lateral distribution functions (LDF), $S(r)$.

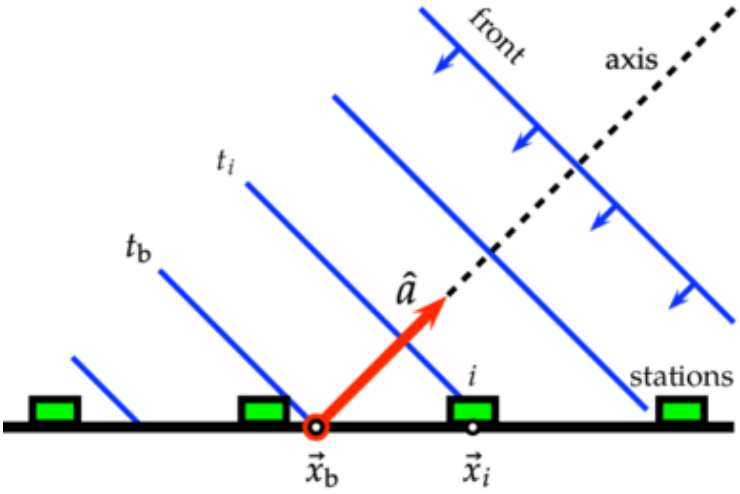


Figure 11 – Diagram of the air-shower development considering the shower front as a plane. Taken from [17].

The process of reconstructing SD events involves determining the LDF at the position of each station. The LDF is a representation of a station's signal as a function of its perpendicular distance from the shower axis, represented by r. Due to statistical fluctuations, two air showers created by cosmic rays with similar mass, energy, and direction can trigger different stations. To accurately estimate the shape of the LDF, a sufficient number and distribution of stations must be activated by the shower. To overcome this challenge, the signals from the stations are adjusted to a function $f_{LDF}(r)$ such that:

$$S(r) = S(r_{opt}) f_{LDF}(r), \tag{2.2}$$

where $S(r_{opt})$ is the shower size estimator and r_{opt} is the optimal value distance that minimizes the differences due to statistical fluctuations. This distance is chosen based on the array configuration. At the Pierre Auger Observatory, this value is $r_{opt} \approx 1000$ m. One important constraint of $f_{LDF}(r)$ is that $f_{LDF}(r_{opt}) = 1$. Figure 12 shows the LDF of an event.

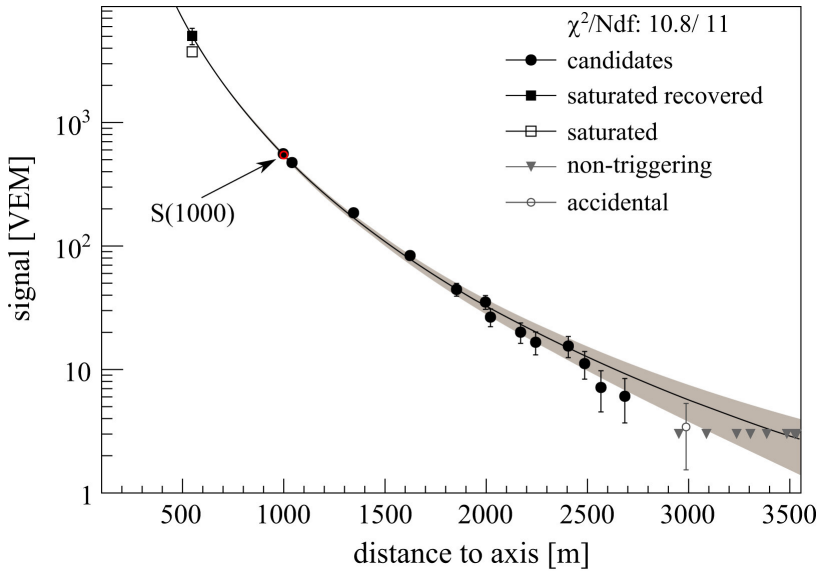


Figure 12 – Lateral distribution function for an event recorded at the Pierre Auger Observatory. The signal at 1000 m from the shower core is highlighted. Taken from [15].

2.4.1 *Herald* framework

The assumption made for this framework is: The air-shower front is curved, and the curvature is considered constant. With this assumption, the particles are delayed proportionally to

$$R_0 - \sqrt{R_0^2 - [r_{\hat{a}}(\vec{x} - \vec{x}_{gr})]^2}, \tag{2.3}$$

where R_0 is the constant radius, and

$$r_{\hat{a}}(\vec{x}) = |\hat{a} \times \vec{x}|, \tag{2.4}$$

is the perpendicular distance of $\vec{x}$ to $\hat{a}$. Expanding equation 2.3 up to the second order in r/R_0, we obtain a paraboloidal extension of eq. 2.4 as:

$$ct(\vec{x}) = ct_{gr} - \hat{a} \cdot (\vec{x} - \vec{x}_{gr}) + k_0[r_{\hat{a}}(\vec{x} - \vec{x}_{gr})]^2, \tag{2.5}$$

where $k_0 = 1/2R_0$ is the curvature parameter, and t_{gr} is the time when the shower hits the ground.

2.4.1.1 Lateral distribution function

A log-log parabola is the $f_{LDF}(r)$ used in the *Herald* framework. The form of the parabola is given by:

$$\ln f_{LDF}(r) = \beta\rho + \gamma\rho^2, \tag{2.6}$$

where $\rho = \ln(r/r_{opt})$, β and γ are the adjusted parameters.

2.4.2 *Observer* reconstruction

The assumption made for this framework is: The air shower development is considered a sphere propagating toward the ground from the point $\vec{x}_0$ with a start time of t_0.With this assumption, the arrival time is given by:

$$ct(\vec{x}) = ct_0 + |\vec{x} - \vec{x}_0|, \tag{2.7}$$

with the equation 2.7, it is possible to obtain the starting point and time. The arrival direction is obtained through the equation:

$$\hat{a} = \frac{\vec{x}_0 - \vec{x}_{gr}}{|\vec{x}_0 - \vec{x}_{gr}|}. \tag{2.8}$$

Again, it is possible to use the lateral distribution function to obtain a better approximation in the the value of $\vec{x}_{gr}$.

2.4.2.1 Lateral distribution function

A modified NKG function is the $f_{LDF}(r)$ used in the *Observer* framework. The form of the parabola is given by:

$$f_{LDF}(r) = (\frac{r}{r_{opt}})^{\beta}(\frac{r + r_s}{r_{opt} + r_s})^{\beta+\gamma}, \tag{2.9}$$

where $r_s = 700$ m, β and γ are the adjusted parameters.

2.4.3 Energy calibration

For the SD to be able to estimate the energy of a cosmic ray more accurately, it needs to be calibrated with hybrid events. Hybrid events are detected by both the SD and FD simultaneously. This calibration step is crucial to ensure that energy estimations are precise. The signal of an event is inversely proportional to its zenith angle, so an attenuation curve is necessary to allow comparing different events. Using the assumption of an isotropic flux of cosmic ray particles and the Constant Intensity Cut (CIC) method, the attenuation curve is adjusted to the function $f_{CIC}(\theta) = 1 + ax + bx^2 + cx^3$, where $x = \cos^2\theta - \cos^2\overline{\theta}$, $\overline{\theta} = 38^\circ$, $a = 0.980 \pm 0.004$, $b = -1.68 \pm 0.01$, and $c = -1.30 \pm 0.45$. The $\overline{\theta} = 38^\circ$ is the median angle. With $f_{CIC}(\theta)$ defined, it is possible to estimate the signal of an event at $r = 1000$ m as if it arrived at a zenithal angle of $\theta = 38^\circ$. This attained signal is given by $S_{38} = S(1000)/f_{CIC}(\theta)$. In figure 12, the $S(1000)$ is marked by the red circle in the curve. Since hybrid events are detected individually at both detectors, a power law is used to correlate the energy measured at the FD, E_{FD}, with the S_{38} calculated for the event measured at the SD. This power law is given by: $E_{FD} = A(S_{38}/\text{VEM})^B$, where $A = (1.90 \pm 0.05) \times 10^{17}$ eV, $B = 1.025 \pm 0.007$, and VEM is the signal unit used in SD.

3 The method

The method employed to separate the muon signal component from the total signal registered by the SD station is a recurrent neural network (RNN), which belongs to the category of machine learning algorithms. Machine learning algorithms leverage vast amounts of data to uncover patterns that are then utilized to carry out an assigned task.

A recurrent neural network (RNN) was selected due to its proven effectiveness in various tasks, including natural language processing and machine translation. Its success in these applications is because it utilizes the output from the previous time step as input for the next. Among the different types of RNNs, this project will utilize the Long Short Term Memory (LSTM) network, one of the most widely used and particularly well-suited for the intended task due to its extended memory capability.

3.1 The input

The inputs fed into the neural network include the total trace, which is the signal recorded at the SD station, the secant of the reconstructed zenith angle ($\sec\theta$), and the distance between the station and the shower core on the shower plane (r). The signal recorded by the station is expressed in VEM and stored in bins of 25 nanoseconds each, where each bin represents the average signal measured by the station's three PMTs.

The total trace length is 768 bins, but the last 568 bins are not required as most of the muon signal is concentrated within the first 200 bins. In the utilized dataset, the first 200 bins contain the complete muon signal in 90% of the stations for cosmic rays with energy $E < 10^{19}$ eV and 70% of the stations for cosmic rays with energy $E > 10^{19}$ eV [18]. The remaining stations have more than 99% of their muon signal within the first 200 bins for cosmic rays with energy $E < 10^{19}$ eV, and for cosmic rays with energy $E > 10^{19}$ eV, approximately 99% of the muon signal is in the first 200 bins [18]. Using more than 200 bins does not significantly enhance the prediction capability of the neural network and would consume a larger amount of memory. Henceforth, the first 200 bins of the total trace will be referred to as the "trace".

The distance traveled by the particles in the atmosphere affects the air-shower composition. This happens because of the different cross-sections of each particle. Electrons and positrons have higher cross-sections than muons, so they are more likely to interact with atmospheric particles. Particles in inclined showers transverse more atmosphere, so the relative concentration of muons at the ground is higher when compared to vertical showers that do not transverse as much atmosphere. So it is essential to give the neural

network a way to know the amount of atmosphere traversed, which is the role of the inputs $\sec\theta$ and r. The distance traveled is proportional to $\sec\theta$ and r. As explained earlier, both parameters can be obtained by event reconstruction.

3.2 Neural network

The type of machine learning used in this work is a supervised learning algorithm. Therefore, the neural network must go through a process of training with labeled examples to learn the patterns in the data. To explain the training process, the following example will be used: Imagine that a neural network must be trained to label an image according to its content. For simplicity, the images can only be of a cat, a dog, or a bird.

The initial step in training the neural network is to present it with images of each animal. The neural network will then generate a label for the image, which is typically incorrect. This output process is referred to as forward propagation and a pictorial representation is shown in figure 13 **a**. The generated label is then compared to the correct label for the image using a loss function, which measures the prediction error. Forward propagation involves calculating the total input, z, to each neuron at each layer. This is done by taking a weighted sum of the outputs from the units in the previous layer. Subsequently, a non-linear function, $f(.)$, is applied to z to obtain the output of the neuron. The bias terms have been omitted for simplicity. Common non-linear functions used in neural networks include the rectified linear unit function (ReLU), the hyperbolic tangent, and the logistic function.

To improve its predictions, the neural network changes its internal parameters. These internal parameters are weights that are used to multiply the input values. To modify the weights in the neural network, the algorithm employs the backpropagation process. It begins by calculating the gradient for each weight, which indicates how the error would increase if the weight were slightly increased. The weights are then adjusted in the opposite direction of the gradient, effectively minimizing the error. This process is illustrated in Figure 13 **b**.

During the backward pass, the computation involves determining the error derivatives at each hidden layer concerning the output of each neuron. This is done by considering the weighted sum of the error derivatives with respect to the total inputs to the units in the layer above. To convert the error derivative from output to input, it is multiplied by the gradient of the non-linear function used in the neural network, such as ReLU or sigmoids. At the output layer, the error derivative is computed by differentiating the cost function, such as the mean square error. If the cost function is $0.5(y_l - t_l)^2$, where t_l is the target value, the error derivative is given by $y_l - t_l$. By knowing the $\partial E/\partial z_k$ (where z_k is the input to neuron k), we can calculate the error derivative for the weight w_{jk} that

connects the neuron j in the layer below, using the equation $y_j * \partial E/\partial z_k$.

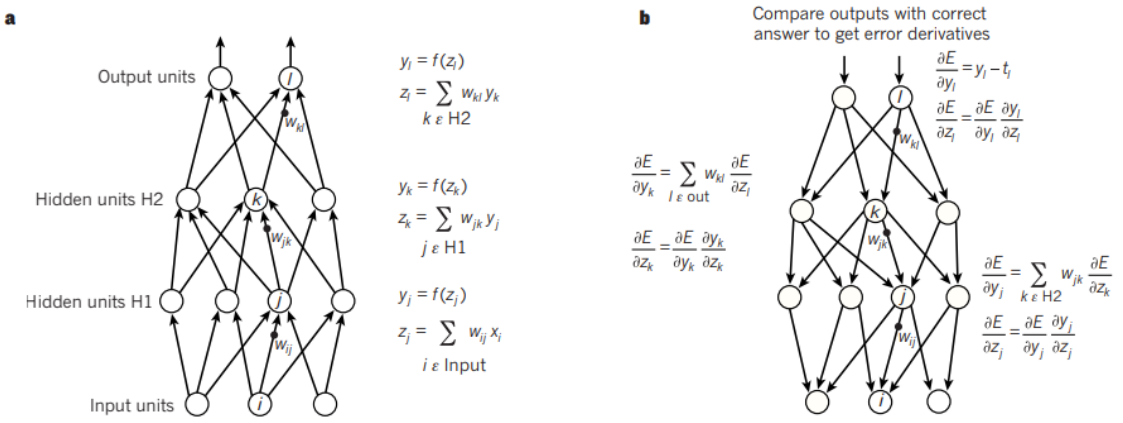


Figure 13 – **a**: The equations utilized for carrying out the forward propagation in a neural network with two hidden layers and a single output layer. **b**: The procedures for computing the backward pass involve computing the error derivatives at each hidden layer with respect to the output of each neuron. The details of the procedures are explained in the text. Taken from [19]

The architecture

The architecture of the models used in this work combines two basic blocks: fully connected layers (FC) and long short-term memory (LSTM) cells.

3.2.1 Fully connected layer

FC stands for fully connected and refers to the type of layer commonly used in feed-forward neural networks, which are a traditional type of machine learning algorithm.

In a fully-connected layer, each neuron in a layer is connected to every neuron in the next layer through a set of weights. The input layer receives the input data, while the output layer generates the final prediction. The hidden layers, located between the input and output layers, perform intermediate computations to arrive at the final prediction. The strength of the connection between neurons is represented by the weights, which are adjustable through the training process to minimize the error.

The values of each neuron are calculated as the weighted sum of the values from the previous layer, and an activation function is applied to the values, except for the input layer, to allow for non-linear data fitting.

Figure 13 shows an example of an FC layer, where x_i are the inputs given to the network and w_{ij} is the weight between the neuron i and j.

This work will use FC layers with and without activation functions. FC layers without activation functions will be called "projectors," while those with activation functions will be referred to as "FC layers."

If an FC layer's activation function is not given, a ReLU activation function is used. Its equation and plot can be seen in Figure 14.

In PyTorch, the input of an FC layer must be of dimensions $[*, H_{in}]$, and the output must be of dimension $[*, H_{out}]$, where H_{in} and H_{out} are defined on the FC layer initialization, and $*$ is any dimension.

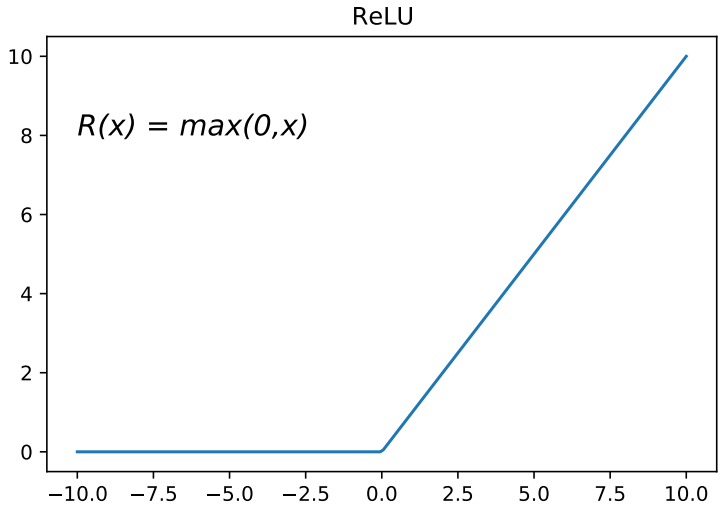


Figure 14 – Graphic representation of the ReLU function for values between −10 and 10.

3.2.2 Long short-term memory

The LSTM architecture was first introduced in 1997 [20]. It aimed to address the issue of vanishing gradients that was present in traditional RNNs, allowing LSTMs to store information over an extended period of time. To achieve this goal, the LSTM cell requires three key elements at each time step: the *cell state*, the *hidden state*, and the *input*.

The vector responsible for preserving long-term memory is referred to as the *cell state*. The *hidden state* acts as the output from the previous time step and is in charge of the short-term memory. Finally, the *input* is the value provided at the current time step.

The LSTM cell utilizes three distinct gates to regulate the *cell state* vector: the forget gate, the input gate, and the output gate. A graphical representation of an LSTM cell can be seen in Figure 15. In this illustration, the cell state from the previous time step is represented as c_{t-1}, the previous hidden state is h_{t-1}, and the current input is x_t.

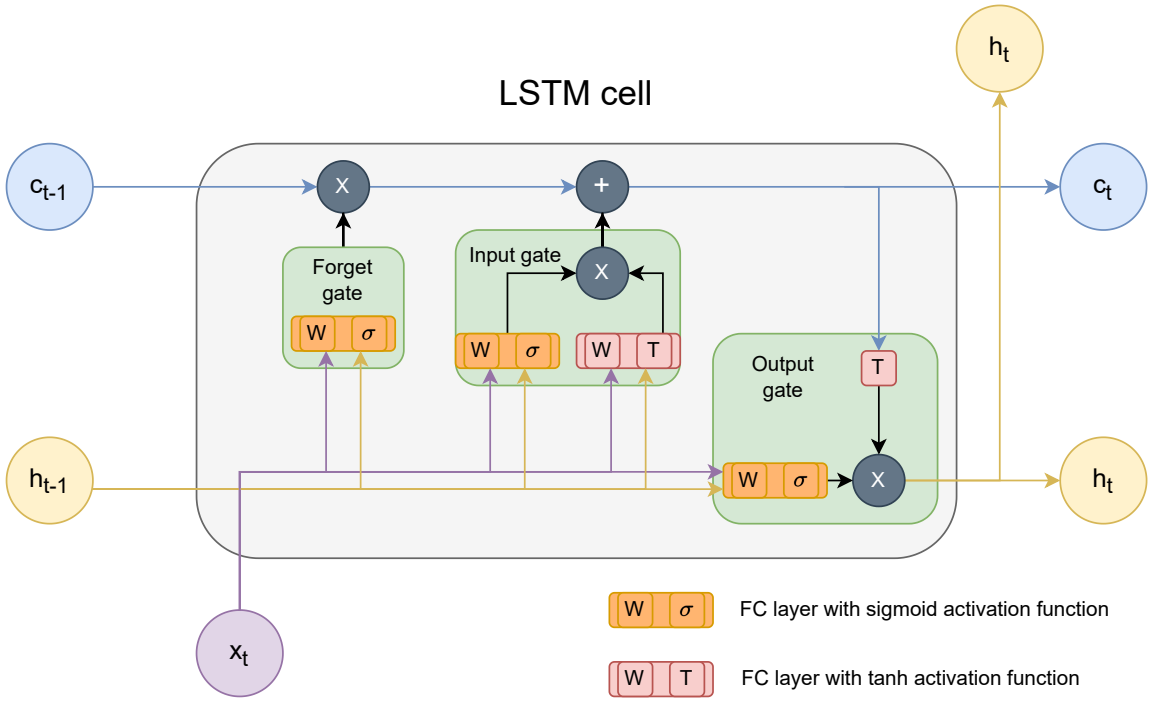


Figure 15 – An LSTM cell diagram showing the gates that determine how the inputs are combined to produce the output. Based on a figure from [20].

In PyTorch, the input of an LSTM cell must have dimensions $[L, H_{in}]$, where L is the sequence length and H_{in} is the number of sequences given to the LSTM. The *hidden state* and *cell state* provided to the LSTM cell must have dimensions $[1, H_{out}]$, where H_{out} is the number of produced sequences. The output, the last *hidden state*, and the last *cell state* of the LSTM cell are of dimensions $[L, H_{out}]$, $[1, H_{out}]$, and $[1, H_{out}]$, respectively.

Forget gate

The forget gate decides which information to discard from the *cell state* c_{t-1}. This is achieved by combining the current input, x_t, and the previous *hidden state*, h_{t-1}. The resulting concatenated vector is then fed through a fully connected layer (FC) that is equipped with a sigmoid activation function. The output of this FC layer, f_t, is then elementwise multiplied by the *cell state*. This process can be mathematically represented as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{3.1}$$

$$c_f = c_{t-1} \times f_t, \tag{3.2}$$

where the f subscript indicates that it belongs to the forget gate.

Input gate

The input gate is responsible for determining which information to add to the *cell state*: c_f. It consists of two components, one that decides what information to add and another that determines the magnitude of the added content. The initial step for both components is to concatenate the current input, x_t, and the previous *hidden state*, h_{t-1}. The concatenated vector is then passed through two parallel FC layers: one with a sigmoid activation function and the other with a hyperbolic tangent activation function. The output of the FC layer with the tanh activation, $\tilde{c}_t$, is used to decide what to add to the *cell state*, while the output of the FC layer with the sigmoid activation, m_t, modulates the magnitude of the added content. Both outputs are then multiplied and combined with the *cell state*, c_f. This process can be mathematically represented as follows:

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \tag{3.3}$$

$$m_t = \sigma(W_m \cdot [h_{t-1}, x_t] + b_m), \tag{3.4}$$

$$c_t = c_f + \tilde{c}_t * m_t. \tag{3.5}$$

Output gate

The output gate determines the final output of the LSTM cell, h_t, by combining the *cell state*, c_t, the *hidden state*, h_{t-1}, and the input, x_t. This is done by first concatenating the input, x_t, and the *hidden state*, h_{t-1}. The resulting concatenated vector is then passed through an FC layer with a sigmoid activation function. The output of this FC layer, o_t, is then multiplied by the hyperbolic tangent of the *cell state*. This process can be represented mathematically as follows:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{3.6}$$

$$h_t = o_t * \tanh(c_t). \tag{3.7}$$

It is important to note that the cell output is used as the next cell's *hidden state*.

3.3 Models used

In this work, four models were used to predict the muon signal of an event station. All of them are encoder-decoder-based architectures. The varying elements of the models are the number of LSTM cells on the encoder and if the *hidden* and *cell* states are passed or not from one LSTM to the other.

In all models, the distance, r, and the secant of the reconstructed zenithal angle, $\sec\theta$, are passed through two modules. Each module is composed of two consecutive FC layers.

The output of one module is used as the initial *hidden state* of the encoder, and the output of the other module is used as the initial *cell state* of the encoder. These modules are shown in figure 16. The inputs to the modules are of dimensions [1, 2]. The shape of the outputs depends on the model.

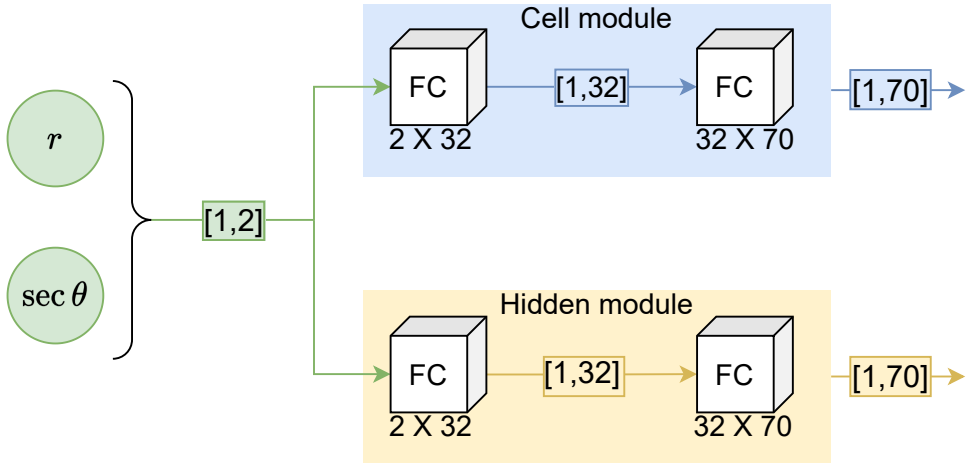


Figure 16 – Schematic representation of the modules used to prepare the static inputs to be fed for the encoder.

3.3.1 First model

This model comprises an encoder with two LSTM cells, a decoder with one LSTM cell, and an FC layer at the end. The last *hidden* and *cell* states from the LSTM cell do not pass to the next LSTM cell. This model is the same as used by the Pierre Auger Collaboration in the paper [18].

The encoder receives three inputs: the initial *hidden* and *cell* states, both of shape [1, 70], and the station trace, of shape [200, 1]. The first LSTM cell receives 1 trace from the encoder and outputs 70 sequences, each containing 200 values.

These sequences are then fed to the next LSTM cell, which produces a 32 series of 200 values. Based on these 32 series, the decoder generates a trace. This final trace is

passed through an FC layer that generates the prediction of the muon signal. Figure 17 shows the model schematics.

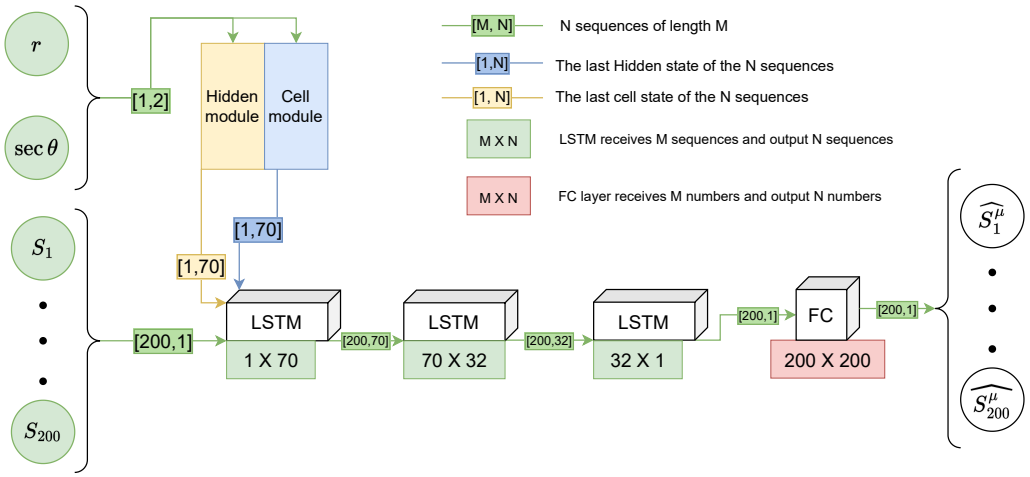


Figure 17 – Schematic representation of the first model used. It consists of two LSTM layers in the encoder and one LSTM layer on the decoder.

3.3.2 Second model

This model comprises an encoder and a decoder, both with just one LSTM cell. The last *hidden* and *cell* states from the LSTM cell do not pass to the next LSTM cell.

The encoder receives three inputs: the initial *hidden* and *cell* states, both of shape [1, 70] and the station trace of shape [200, 1]. The LSTM cell encoder receives 1 trace in the encoder and outputs 70 traces.

Based on these 70 traces, the decoder generates a trace. This final trace is passed through an FC layer that generates the prediction of the muon signal. Figure 18 shows the schematics of the model.

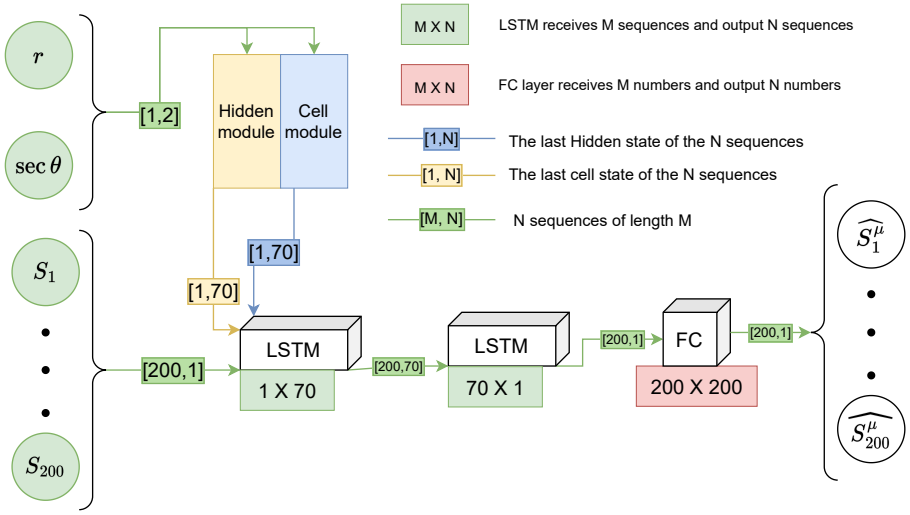


Figure 18 – Schematic representation of the second model used. It consists of one LSTM layer in the encoder and one LSTM layer on the decoder.

3.3.3 Third model

This model comprises an encoder with two LSTM cells and a decoder with one LSTM cell. The last *hidden* and *cell* states from the LSTM cell are inputs to the next LSTM cell.

The encoder receives three inputs: the initial *hidden* and *cell* states, both of shape [1, 70] and the station trace of shape [200, 1]. The first LSTM cell receives 1 trace from the encoder and outputs 70 sequences, each containing 200 values.

Because the number of sequences generated by LSTM cells is different, the output *hidden* and *cell* states must be projected to be used as inputs for the next LSTM cell.

Between the first and last encoder LSTM cells, a pair of projectors transforms the *hidden* and *cell* states dimension from [∗,70] to [∗, 32].

Together with the 70 sequences, the projected *hidden* and *cell* states are fed to the next LSTM cell that produces 32 sequences, each containing 200 values.

Between the last encoder and first decoder LSTM cell, another pair of projectors transforms the *hidden* and *cell* states dimensions from [∗,32] to [∗, 1].

The decoder generates a trace based on the 32 sequences and the projected *hidden* and *cell* states. This final trace is passed through an FC layer that generates the prediction of the muon signal. Figure 19 shows the schematics of the model.

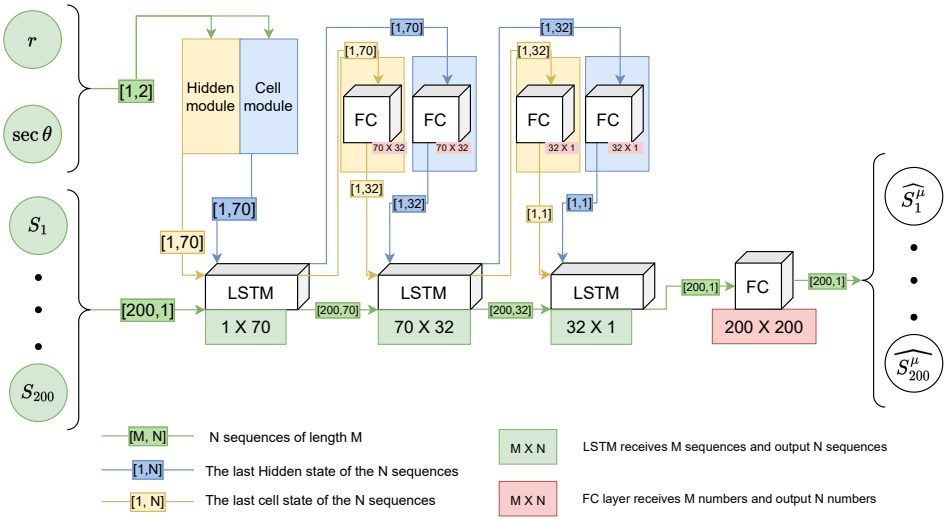


Figure 19 – Schematic representation of the third model used. It consists of two LSTM layers in the encoder and one LSTM layer on the decoder. The last cell state and hidden state from one LSTM layer are passed through a projector, serving as inputs for the next LSTM layer.

3.3.4 Fourth model

This model comprises an encoder with one LSTM cell and a decoder with one LSTM cell. The last *hidden* and *cell* states from the LSTM cell are inputs to the next LSTM cell.

The encoder receives three inputs: the initial *hidden* and *cell* states, both of shape [1, 70] and the station trace of shape [200, 1]. The first LSTM cell receives 1 trace from the encoder and outputs 70 sequences, each containing 200 values.

Between the first encoder and the first LSTM cell decoder, a pair of projectors transforms the *hidden* and *cell* states dimensions from [∗,70] to [∗, 1].

The decoder generates a trace based on the 70 sequences and the projected *hidden* and *cell* states. This final trace is passed through an FC layer that generates the prediction of the muon signal. Figure 20 shows the schematics of the model.

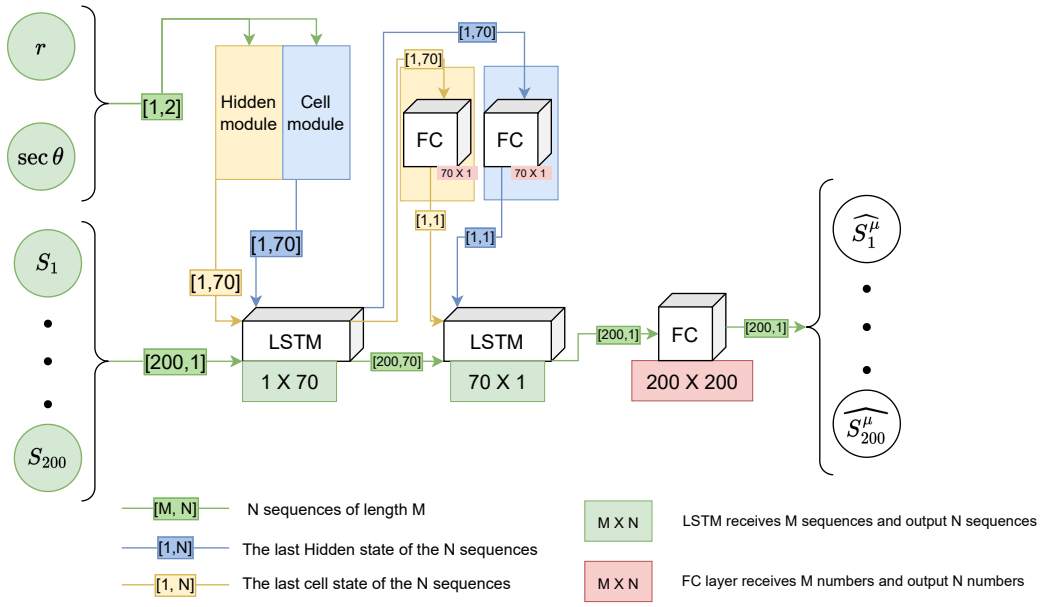


Figure 20 – Schematic representation of the fourth model used. It consists of one LSTM layer in the encoder and one LSTM layer on the decoder. The last cell state and hidden state from one LSTM layer are passed through a projector, so they serve as inputs for the next LSTM layer.

3.4 Dataset

Providing the muon signal of each station is a crucial aspect of the gradient descent procedure. However, this poses a challenge as the SD stations cannot differentiate between Cherenkov radiation produced by electrons, positrons, or muons. To overcome this obstacle, the neural network will be trained using simulations.

The air shower simulation in this study was carried out using the CORSIKA software with EPOS-LHC as the hadronic interaction model. In particular, the showers used were simulated with CORSIKA 7.6400, 7.7100, and 7.7400. The offline software of the Pierre Auger Collaboration was utilized to reconstruct each simulated air shower. The versions used for the reconstructions were v3r3p4 and v3r99p2a. The simulated events used in this work were compiled by the Pierre Auger Collaboration and downloaded via the Internet from the Naples Shower Library.

The study selected only simulated events that met certain conditions for further analysis. These conditions included the requirement that there were six operating stations around the station with the highest signal, which was done to exclude events that took place at the periphery of the array. Additionally, the signal in the stations had to be at least 5 VEM and not saturated, as saturation occurs when the Cherenkov radiation produced is too intense, causing the PMT to display nonlinear behavior.

After these cuts, there were 8.866.662 signals available for use. Due to the

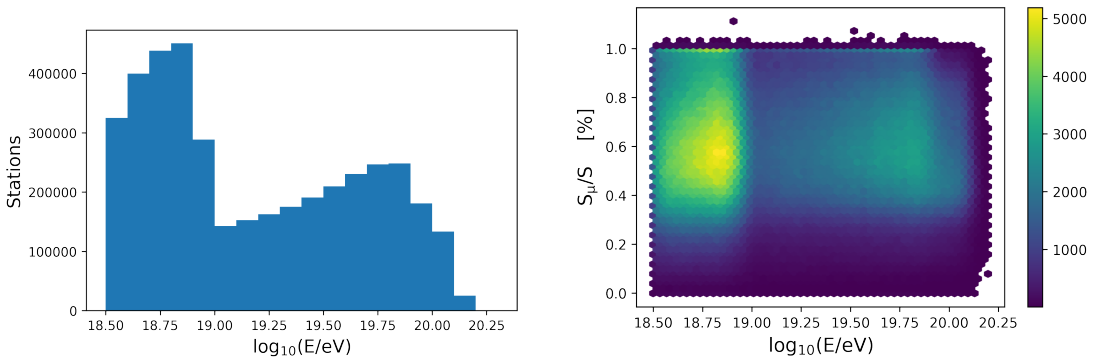
available memory, 50.26% of the available simulations were used to compose the dataset for this work. The signals available were randomly selected. To ensure reproducibility, a file containing only the selected events was created.

The selected simulations were divided into three sub-datasets: the training dataset, the validation dataset, and the test dataset. This division was made randomly, but the validation and training datasets were sampled using a uniform distribution in $\sec\theta$ and the logarithm of energy. Each of them has 4.000.000, 108.640, and 348.000 station simulations, respectively. The particle composition of each dataset is displayed in table 1. Approximately, there are 25% more events initiated by heavy particles (iron nuclei) than by light particles (hydrogen nuclei).

Table 1 – Number of initial particles in each dataset

	Training dataset	**Validation dataset**	**Test dataset**
Proton	924141	25158	79948
Helium	961188	26029	84121
Oxygen	1029537	27971	89469
Iron	1085133	29482	94462

The energy distribution of the training dataset is shown in Figure 21a. The distribution of the simulated energies has two peaks. Figure 21b shows the muon signal as a percentage of the total signal as a function of the logarithm of the cosmic-ray energy, where the color represents the concentration of points. The absence of an uptrend or a downtrend shows that cosmic-ray energy has little influence on the muon-to-total signal ratio.



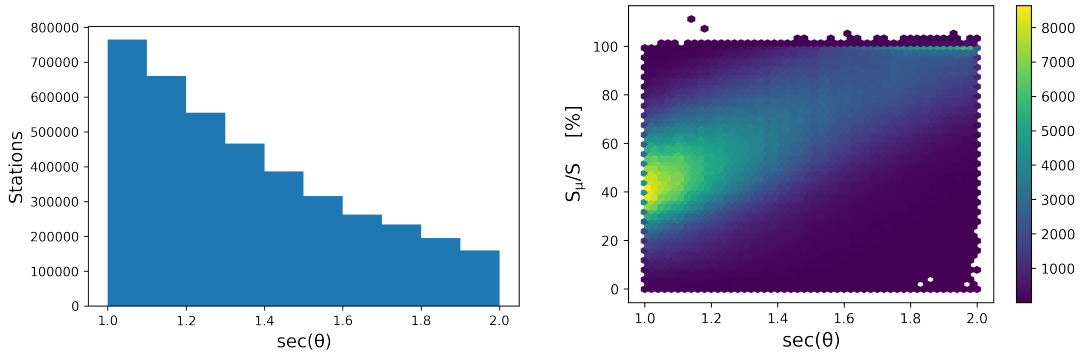
(a) Distribution of stations.

(b) Percent of muonic signal in relation to the total signal.

Figure 21 – Distributions of stations in relation to the energy.

The arrival angle distribution of the training dataset is shown in Figure 22a. The distribution of the simulated arrival angles shows a decreasing trend, which is expected

since the simulations were done with a uniform distribution with respect to $\cos^2\theta$. Figure 22b shows the percentage of muon signal in relation to the total signal as a function of the cosmic-ray arrival angle; it clearly has an upward trend, which shows that the muon-to-total signal ratio increases with higher arrival angles.

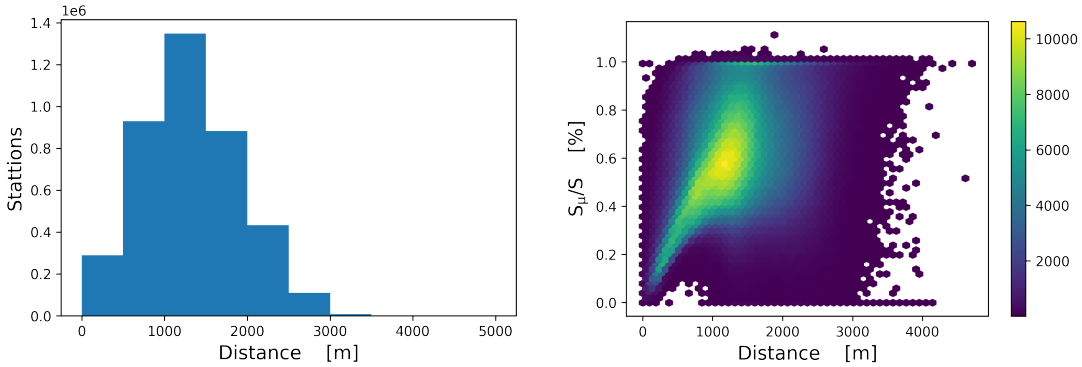


(a) Distribution of stations in relation to the secant of the arrival angle.

(b) Percent of muonic signal in relation to the total signal.

Figure 22 – Distributions of stations in relation to the arrival angle.

The distance between the station and the shower core distribution, as expected, is not isotropic. Figure 23 (a) shows a clear peak around 1500 m. Figure 23 (b) shows the percentage of muon signal in relation to the total signal as a function of this distance. It also shows an upward trend, which indicates that the muon signal depends on the distance.



(a) Distribution of stations.

(b) Percent of muonic signal in relation to the total signal.

Figure 23 – Distributions of stations in relation to the distance to the core on the plane perpendicular to the shower plane.

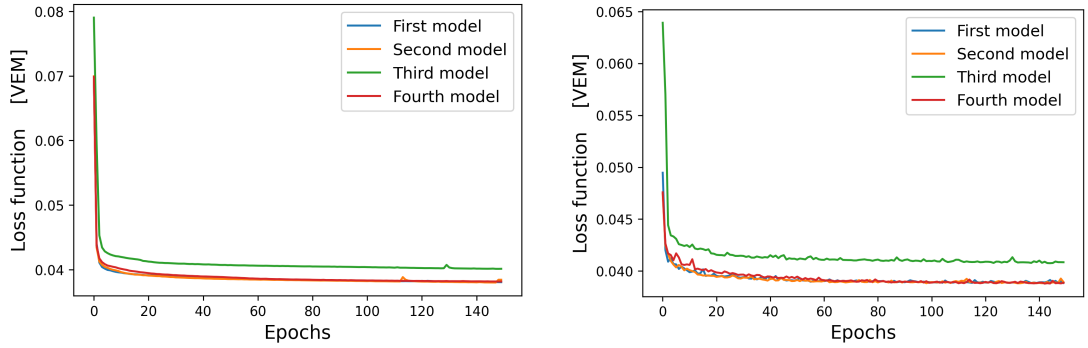
Before the training process begins, the data in the datasets are adjusted by scaling each trace to a range between 1 and 0. The largest value of the trace is used as the scale factor, and the muon signal is also scaled using this factor. The distance (r) and the

secant of the zenithal angle ($\sec\theta$) are also scaled by the largest value. The scaled datasets are then divided into batches of 512 samples. The mean squared root error (RMSE) is employed as the loss function and is defined as:

$$L = \sqrt{\frac{1}{200}\sum_{i=1}^{200}(\widehat{S}_i^{\mu} - S_i^{\mu})^2}, \tag{3.8}$$

where the $\widehat{S}_i^{\mu}$ represents the predicted muon signal at time step i, and S_i^{μ} represents the simulated muon signal at the same time step. This function is calculated for each batch trace and averaged over all examples.

The training process used the ADAM [21] optimization algorithm with a learning rate of 10^{-4}. The training was performed for a total of 150 epochs on a Nvidia Tesla T4 GPU using the Google Colab platform. The training time for each of the four models varied, but it took approximately eight hours to complete each one.



(a) Loss values as a function of epochs during training of all models.

(b) Loss values as a function of epochs during validation of all models.

Figure 24 – Loss values as a function of epochs of training and validation.

The training loss as a function of epochs is shown in Figure 24 (a), the validation is shown in Figure 24 (b). Since both have similar values, it indicates that the models learned the patterns of the data without overfitting.

4 Results

In this chapter, only the results of the third model will be presented. This choice is made because the models have similar results, so there is no point in discussing them repeatedly. The graphs of the remaining models are presented in Appendix A.

Beyond that, a comparison between the results of this work and that done by the Collaboration [18], and some examples of traces, will also be done in this chapter.

4.1 Differences in the traces

One of the ways to compare the simulated muon signal, S^{μ}, to the one predicted by the neural network, $\widehat{S^{\mu}}$, is by comparing the integral of the signals, $S^{\mu} = \sum_{i=1}^{200} S_i^{\mu}$ and $\widehat{S}^{\mu} = \sum_{i=1}^{200} \widehat{S}_i^{\mu}$. The integral of the muon trace is proportional to the number of muons that reach the ground and therefore is related to the primary cosmic-ray mass.

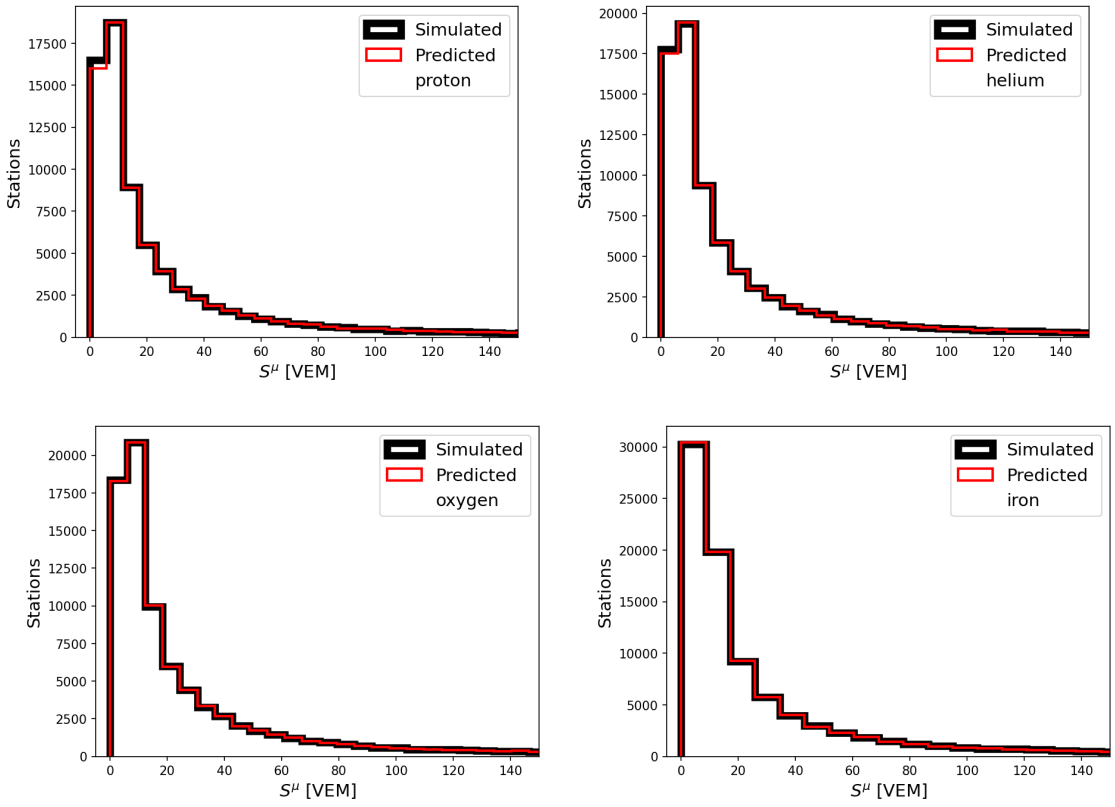


Figure 25 – Distribution of the simulated and predicted muon signals for every station in the dataset, separated by the primary cosmic-ray composition.

Figure 25 hows the distribution of the muon signal across all stations in the test dataset alongside the distribution predicted by the third model. Each figure contains the distribution of showers initiated by a proton, an iron nucleus, a helium nucleus, and an oxygen nucleus. All the models have successfully reproduced the distribution shape.

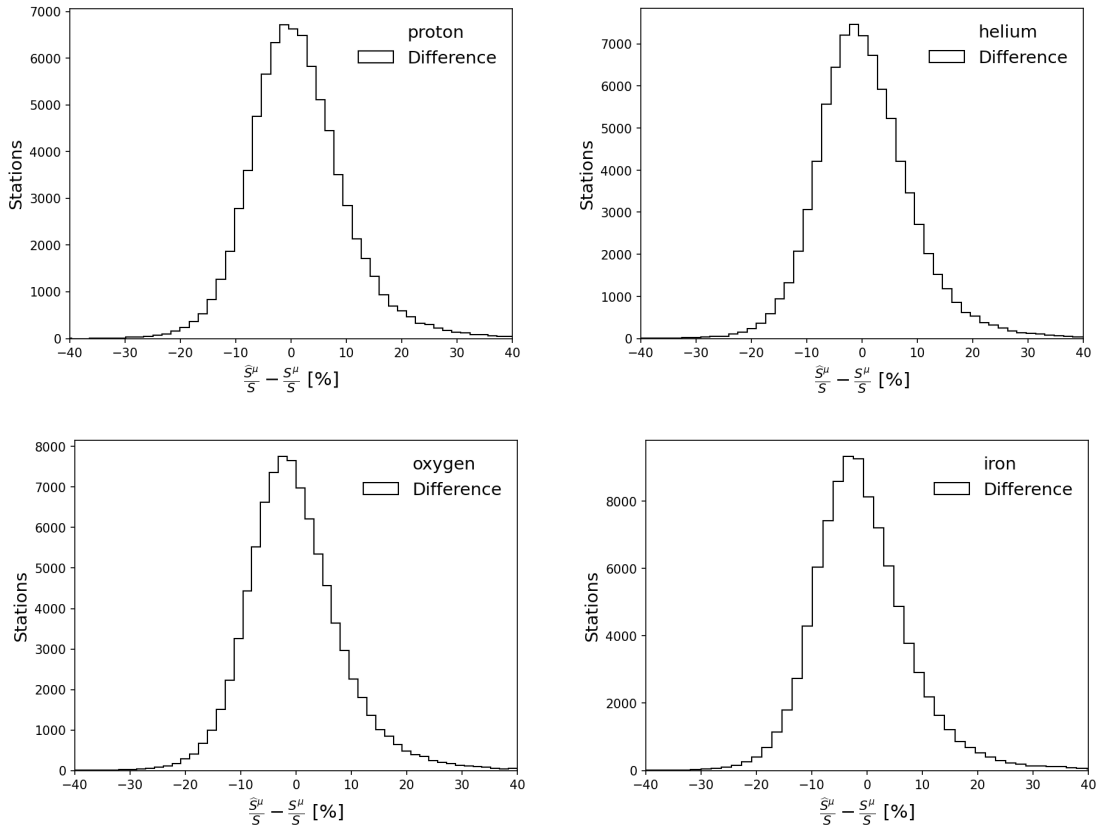


Figure 26 – Distribution of the difference between simulated and predicted muon signals for every station in the test dataset, separated by the primary cosmic-ray composition.

The difference between the predicted integrated signal and the simulated integrated signal as a percentage of the total signal, S, for each station is shown in Figure 26. The distribution is centered near 0 for every primary cosmic ray, indicating that the third model does not have a significant tendency to under or overestimate the muon signal. The other models follow the same tendency. The standard deviation of the same distribution shows that around 99.7% of the predictions agree within 30%.

Figures 27 and 28 show the mean value of the difference as a function of the logarithm of the energy and $\sec\theta$, respectively.

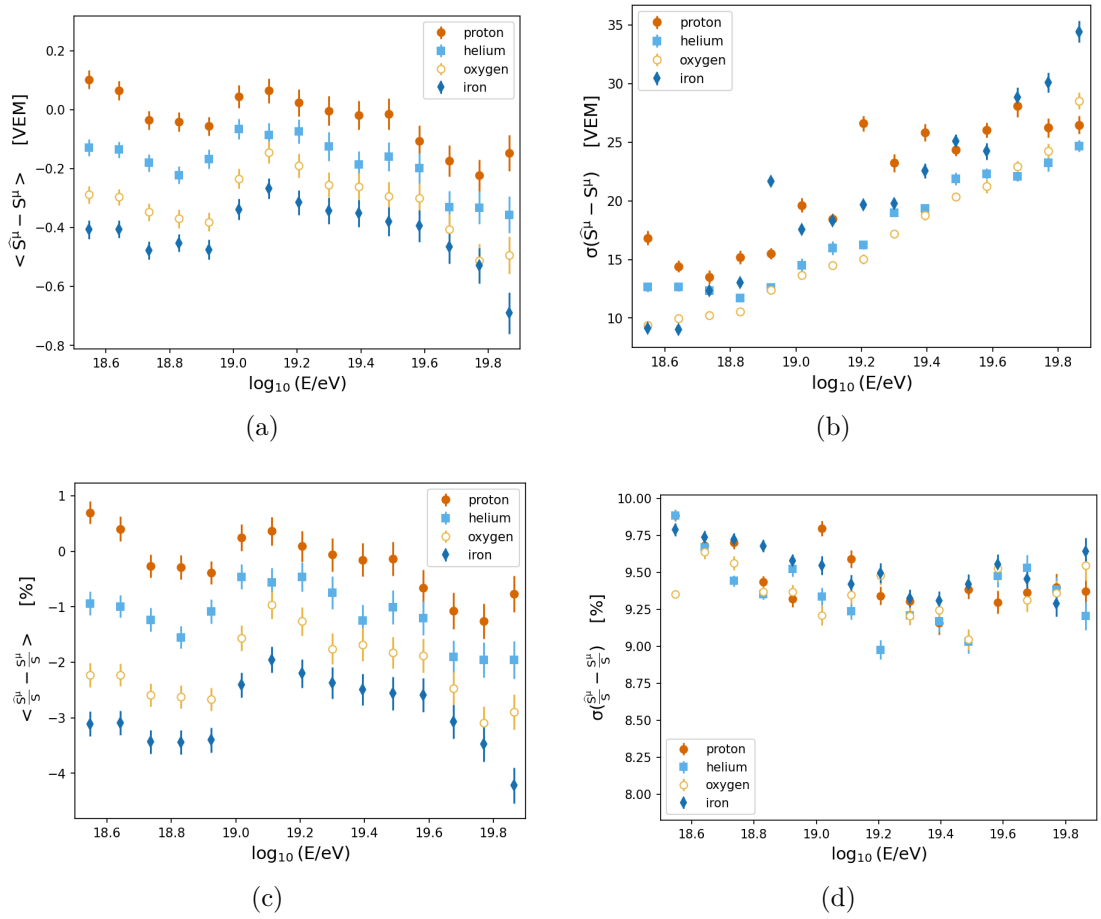


Figure 27 – Mean value (27a) and standard deviation (27b) of the difference between the predicted and simulated values as a function of the logarithm of the energy. Mean value (27c) and standard deviation (27d) of the difference between the predicted and simulated values divided by the total signal as a function of the logarithm of the energy.

Figure 27a shows that for all bins, the mean value is close to zero. Figure 27b shows an increase in the standard deviation for higher energies; this uptrend happens because the signal increases with the energy, and consequently, the difference between the predicted signal and the simulated signal increases. But this does not mean that the accuracy of the third model decreases with energy, as is shown in figure 27d.

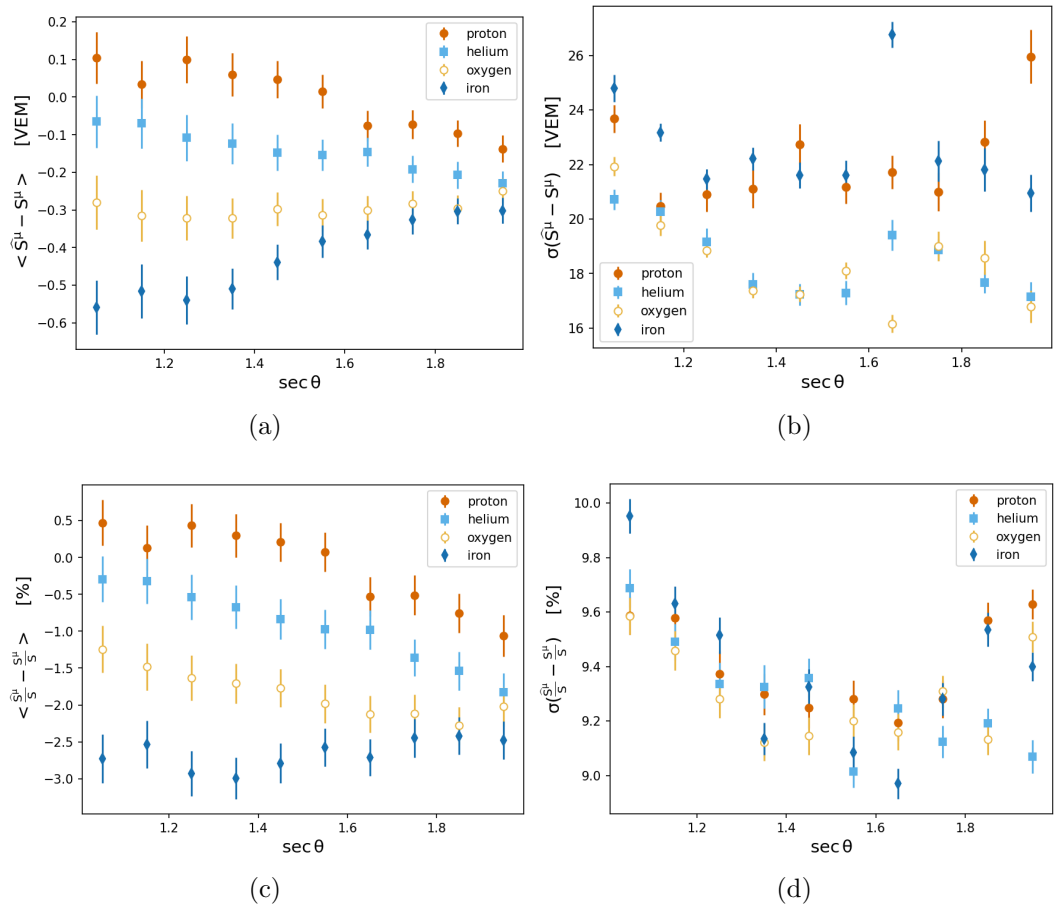


Figure 28 – Mean value (28a) and standard deviation (28b) of the difference between the predicted and simulated values as a function of the secant of the arrival angle. Mean value (28c) and standard deviation (28d) of the difference between the predicted and simulated values divided by the total signal as a function of the secant of the arrival angle.

Figure 28a shows the mean value of the difference for values inside a 0.1 $\sec\theta$ bin. This figure has an interesting information for the more inclined events with $\sec\theta \approx 2$: the prediction is less dependent on the primary cosmic ray composition, and the predicted muon signal is slightly underestimated. Figure 28b shows the standard deviation of the difference between the predicted and simulated muon signals as a function of $\sec\theta$. The accuracy of the predictions improves with higher arrival angles up to $\sec\theta \approx 1.6$ and after this point, the accuracy worsens, as shown in Figure 28d.

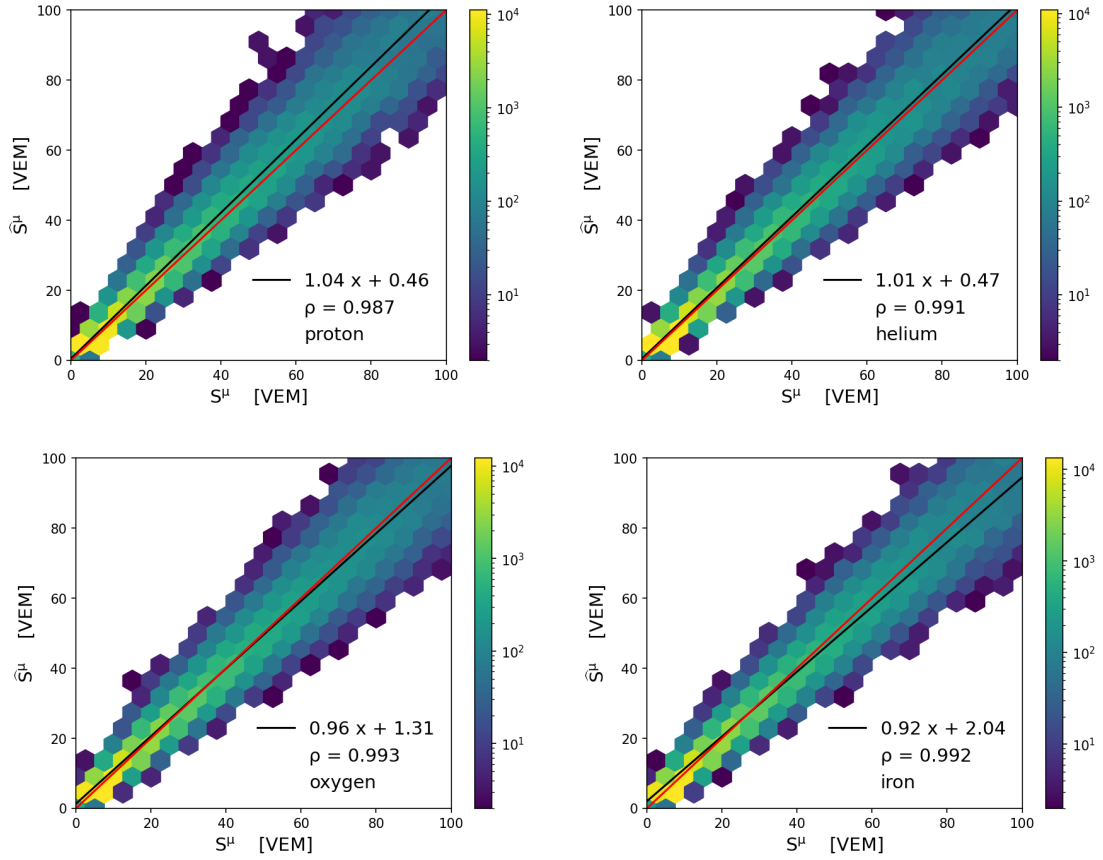
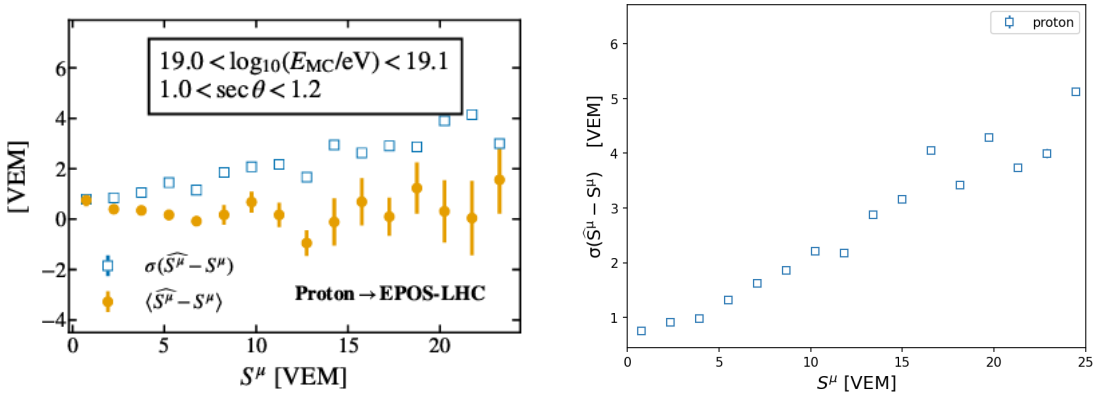


Figure 29 – Integral of the predicted muon signal as a function of the integral of the simulated muon signal. The black line corresponds to the linear fit of the points, and the red line corresponds to the ideal case.

Figure 29 shows the predicted muon signal as a function of the simulated muon signal. The color of each hexagon corresponds to the number of points inside of it. This figure provides another good way to evaluate the model performance. The Pearson correlation coefficient between the predicted and simulated muon signals is higher than 0.98 for all primary, suggesting a strong positive linear correlation between the predicted and simulated muon signals. To further investigate this linear correlation, a linear regression was performed. By comparing the linear regression, the black line in Figure 29, and the perfect case, the red line in Figure 29, it is clear that for lighter primaries the model has a tendency to overestimate the muon signal, while for heavier primary cosmic rays the muon signal is underestimated.

4.2 Comparison

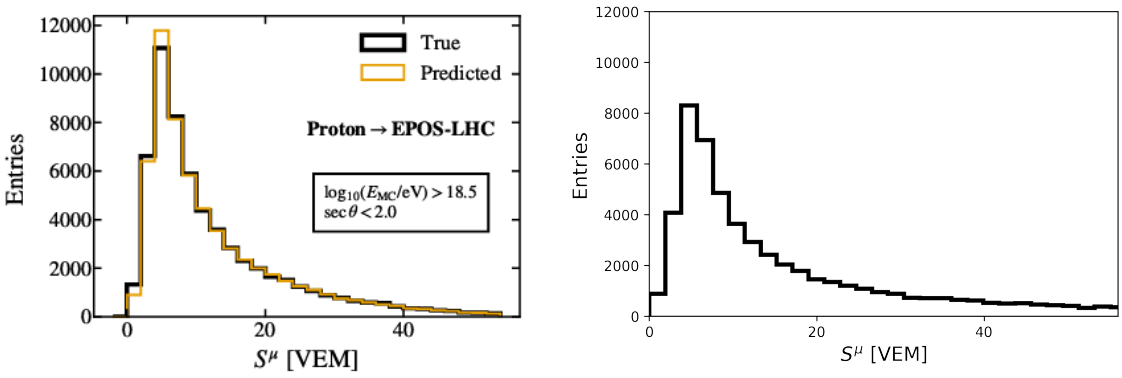
The findings of this study align well with those of the Pierre Auger Collaboration study[18]. However, there is a notable difference between the two studies regarding their standard deviation comparison.



(a) (b)

Figure 30 – 30a: Mean and standard deviation of the difference between simulated and predicted muon signals, for every station with energies and zenith angles specified in the box. Taken from [18]. 30b: Standard deviation of the difference between simulated and predicted muon signals, for every station separated by the primary cosmic-ray composition.

This variation in results is due to the different datasets used. The Pierre Auger Collaboration study used a dataset with fewer stations that recorded signals greater than 40 VEM compared to the dataset used in this study. Therefore, the difference between the predicted and simulated signals is smaller in their study. Figure 30a and 30b depicts the relationship between the standard deviation and the muon signal.



(a) (b)

Figure 31 – 31a: Distribution of $\widehat{S}^{\mu}$ and S^{μ} for all stations in the test dataset. Taken from [18]. 31b: Distribution for S^{μ} for all stations in the test dataset with an atmospheric shower initiated by a proton.

To support this claim, some findings from the collaboration work will be presented. Figure 31a illustrates the distribution of muon signals for the test dataset used in the Pierre Auger study, comprising approximately 65720 stations (calculated using

ImageJ software). As we can see, there are not many stations with muon signals greater than 40 VEM in this figure. In contrast, the dataset used in this study features numerous events above 40 VEM, as shown in figure 31b.

4.3 Trace examples

Figures 32 and 33 present two examples of randomly selected muon traces obtained with the third model, comparing them with the initially simulated ones; The predictions of the other models are shown in Appendix A.

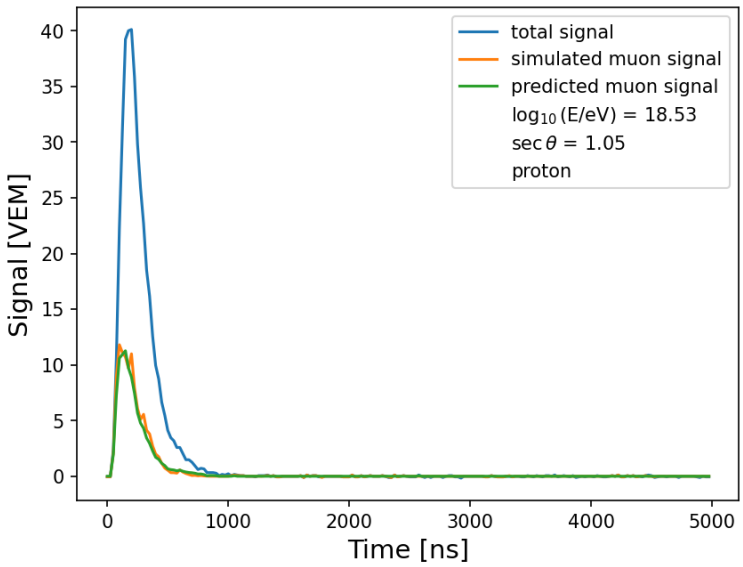


Figure 32 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an electromagnetic-dominated signal. The total simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line.

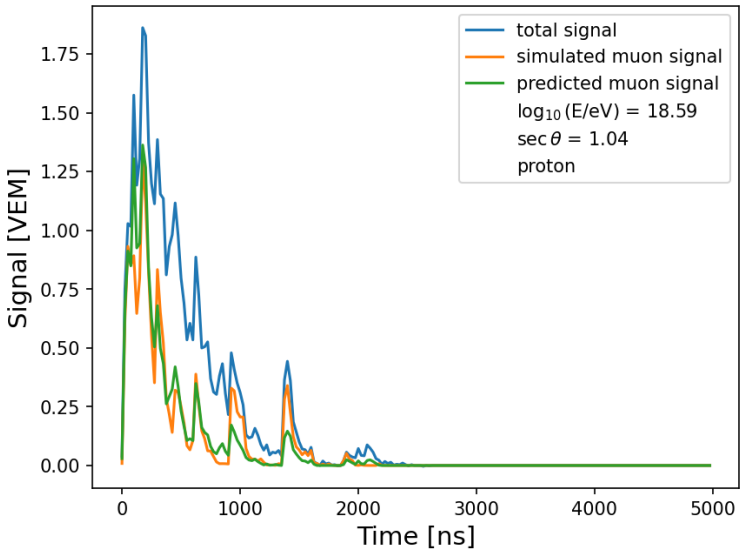


Figure 33 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an muon-dominated signal. The total simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line.

In both examples shown in Figures 32 and 33, the model could satisfactorily reproduce the muon signal of the simulated event.

5 Conclusions

In this work, the performance of machine learning algorithms was examined to predict the muonic component on the signal recorded with a surface detector station of the Pierre Auger Observatory.

The analysis was made using recurrent neural networks, or RNNs, as the main component of the algorithm. Four models were developed in total. They were all based on the model developed by the Pierre Auger Collaboration [18]. The models received the same inputs, which were the recorded signals in the first 200 time bins at the surface detector station, the logarithm of the energy, and the arrival direction of the cosmic ray.

The models were trained using simulations because the current surface detector stations are unable to differentiate the particles that cause the Cherenkov effect. The simulations were done with the CORSIKA program with EPOS-LHC as the hadronic model of each simulated atmospheric shower was reconstructed using the Offline software.

The results found show that the resolution varies from 10% to 9% depending on the arrival direction. Another interesting result is found by comparing my results with those of [18]. The used dataset has a strong influence on the standard deviation of the predictions but does not have a strong influence on the resolution. To ensure this is not just statistical fluctuations, a bigger variety of datasets should be used.

Bibliography

[1] Aab, A. *et al.* Observation of a large-scale anisotropy in the arrival directions of cosmic rays above 8×10^{18} eV. *Science* **357**, 1266–1270 (2017).

[2] Aab, A. *et al.* Depth of maximum of air-shower profiles at the Pierre Auger Observatory. i. measurements at energies above $10^{17.8}$ eV. *Phys. Rev. D* **90**, 122005 (2014).

[3] Hess, V. F. Über Beobachtungen der durchdringenden Strahlung bei sieben Freiballonfahrten. *Phys. Z.* **13**, 1084–1091 (1912).

[4] Kolhörster, W. Messungen der durchdringenden Strahlung im Freiballon in grösseren Höhen. *Phys. Z.* **14**, 1153–1155 (1913).

[5] Carlson, P. & De Angelis, A. Nationalism and internationalism in science: The case of the discovery of cosmic rays. *Eur. Phys. J. H* **35**, 309–329 (2010).

[6] Hanlon, W. F. Cosmic ray spectra of various experiments. https://web.physics.utah.edu/~whanlon/spectrum.html (2011). Online; accessed 29 December 2022.

[7] Simpson, J. A. Elemental and isotopic composition of the galactic cosmic rays. *Annual Review of Nuclear and Particle Science* **33**, 323–382 (1983).

[8] Aguilar, J. A. Particle astrophysics lecture 3 cosmic rays. https://w3.iihe.ac.be/~aguilar/PHYS-467/PA3.html (2015). Online; accessed 29 December 2022.

[9] Kampert, K.-H. & Watson, A. A. Extensive air showers and ultra high-energy cosmic rays: A historical review. *The European Physical Journal H* **37**, 359–412 (2012).

[10] Auger, P., Ehrenfest, P., Maze, R., Daudin, J. & Fréon, R. A. Extensive cosmic-ray showers. *Reviews of Modern Physics* **11**, 288–291 (1939).

[11] Heitler, W. *The Quantum Theory of Radiation*. Dover Books on Physics Series (Dover Publications, 1984).

[12] Matthews, J. A Heitler model of extensive air showers. *Astroparticle Physics* **22**, 387–397 (2005).

[13] Bernlöhr, K. Cosmic-ray air showers. https://www.mpi-hd.mpg.de/hfm/CosmicRay/Showers.html (2000). Online; accessed 07 July 2023.

[14] Aab, A. *et al.* Spectral calibration of the fluorescence telescopes of the Pierre Auger Observatory. *Astroparticle Physics* **95**, 44–56 (2017).

[15] Aab, A. *et al.* The Pierre Auger Cosmic Ray Observatory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **798**, 172–213 (2015).

[16] Bertou, X. *et al.* Calibration of the surface array of the Pierre Auger Observatory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **568**, 839–846 (2006).

[17] Aab, A. *et al.* Reconstruction of events recorded with the surface detector of the Pierre Auger Observatory. *Journal of Instrumentation* **15**, P10021 (2020).

[18] Aab, A. *et al.* Extraction of the muon signals recorded with the surface detector of the Pierre Auger Observatory using recurrent neural networks. *Journal of Instrumentation* **16**, P07016 (2021).

[19] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

[20] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).

[21] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. & LeCun, Y. (eds.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015). URL http://arxiv.org/abs/1412.6980.

Appendix

APPENDIX A – Models 1, 2, 4 figures

Here we present the results for the first, second and fourth models, and as explained in Chapter 4, there will be no discussion since there is no significant difference between the models.

A.1 First model

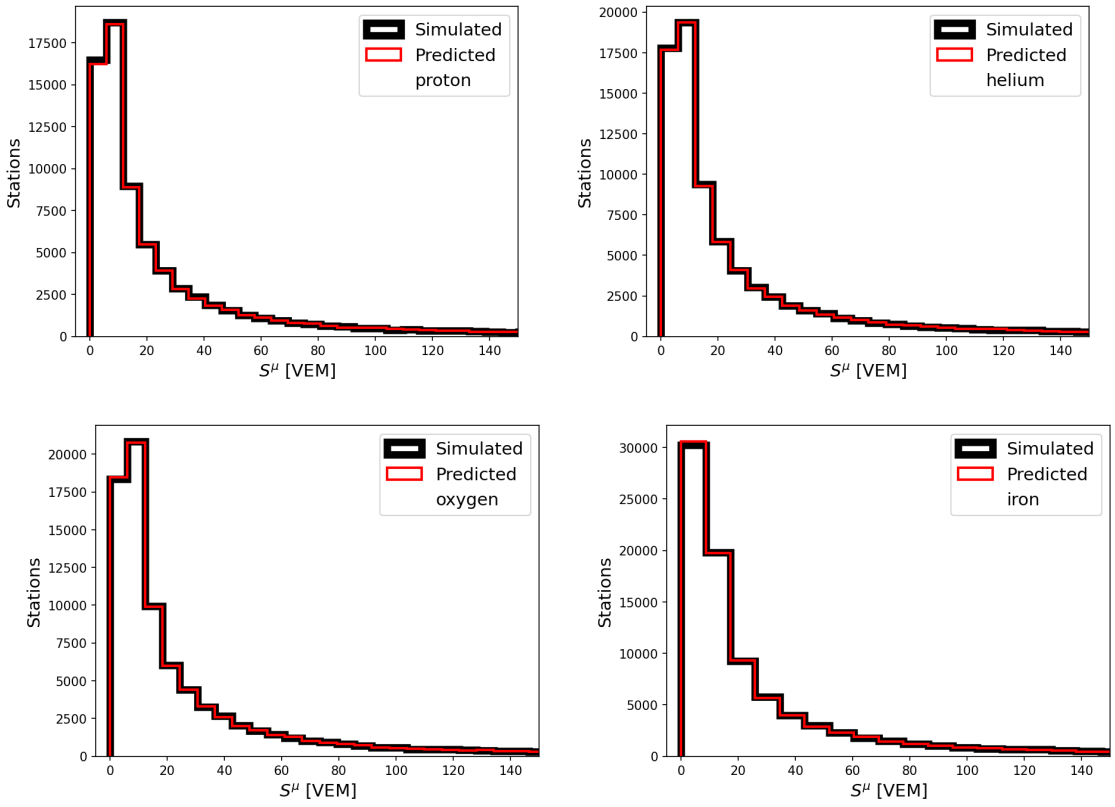


Figure 34 – Distribution of the simulated and predicted muon signals for every station in the dataset, separated by the primary cosmic-ray composition.

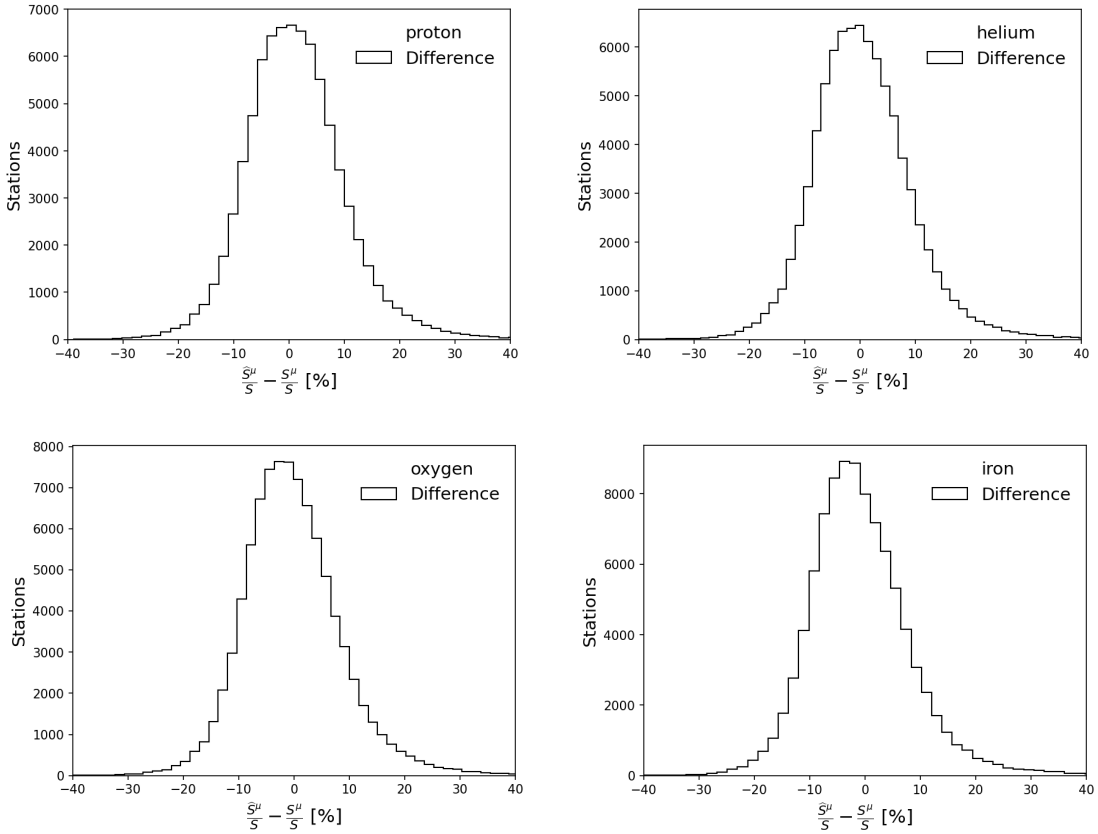


Figure 35 – Distribution of the difference between simulated and predicted muon signals for every station in the test dataset, separated by the primary cosmic-ray composition.

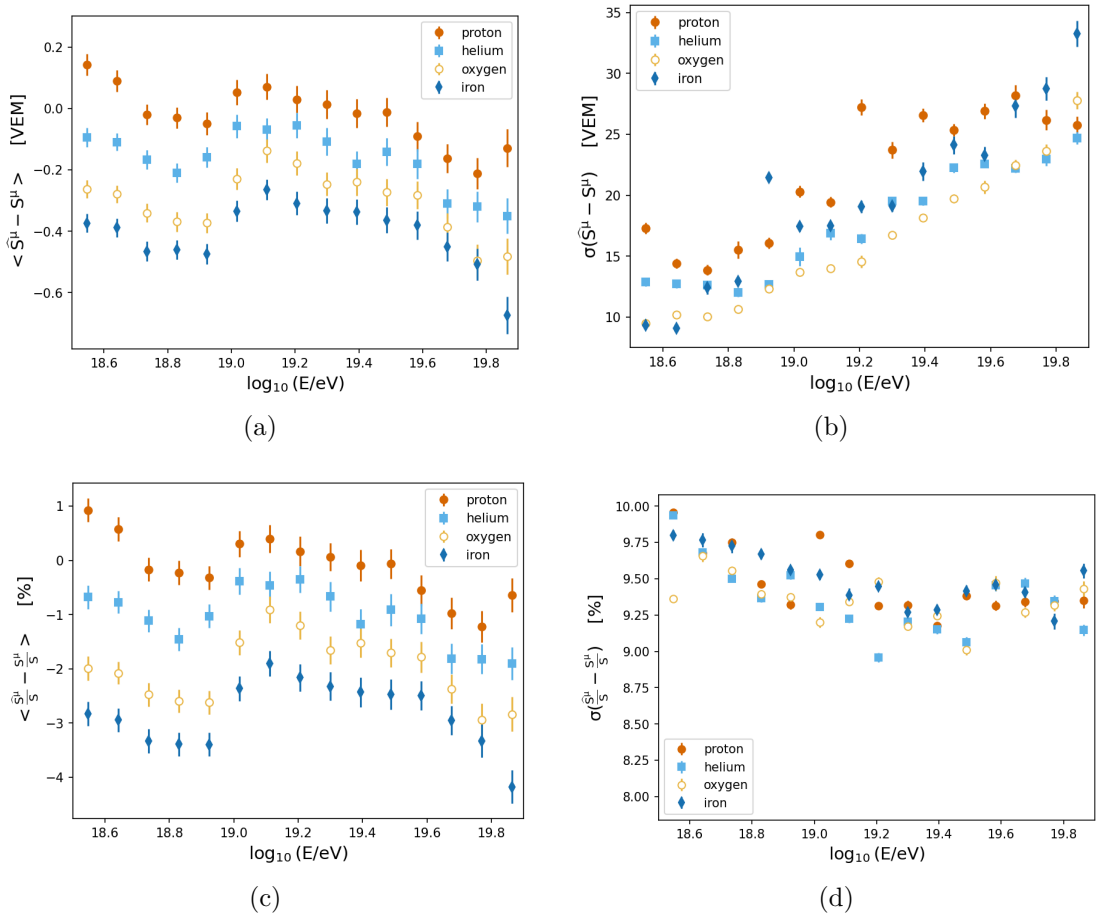


Figure 36 – Mean value (36a) and standard deviation (36b) of the difference between the predicted and simulated values as a function of the logarithm of the energy. Mean value (36c) and standard deviation (36d) of the difference between the predicted and simulated values divided by the total signal as a function of the logarithm of the energy.

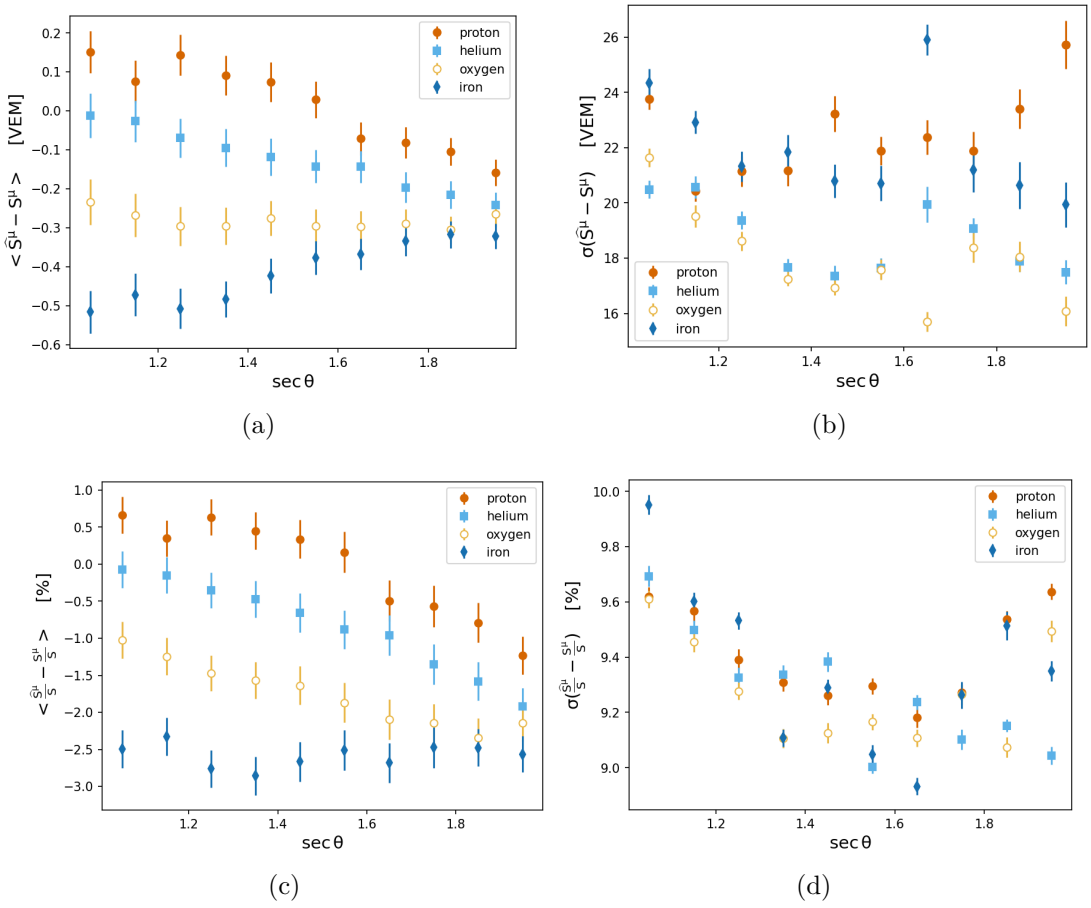


Figure 37 – Mean value (37a) and standard deviation (37b) of the difference between the predicted and simulated values as a function of the secant of the arrival angle. Mean value (37c) and standard deviation (37d) of the difference between the predicted and simulated values divided by the total signal as a function of the secant of the arrival angle.

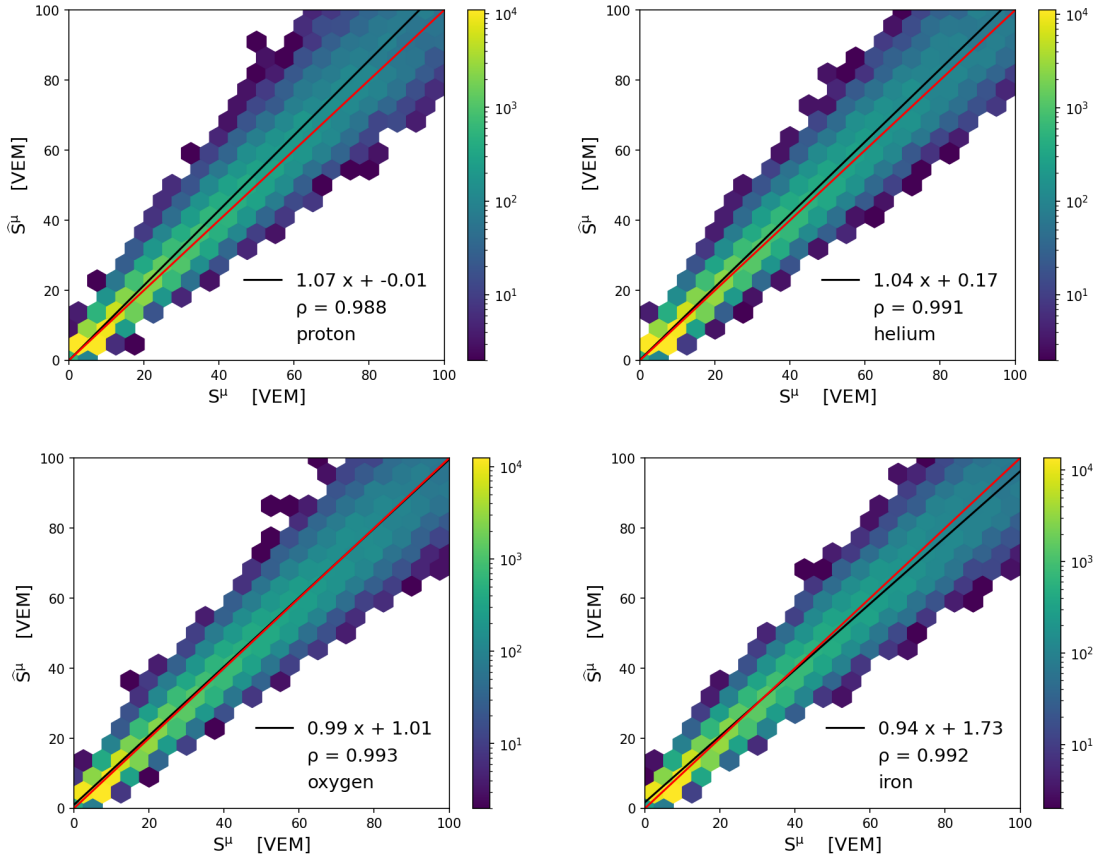


Figure 38 – Integral of the predicted muon signal as a function of the integral of the simulated muon signal. The black line corresponds to the linear fit of the points, and the red line corresponds to the ideal case.

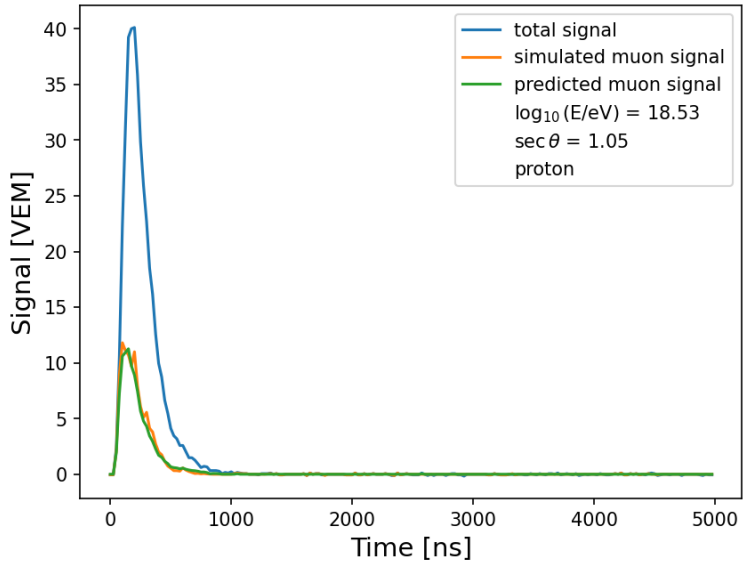


Figure 39 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an eletromagnetic-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line.

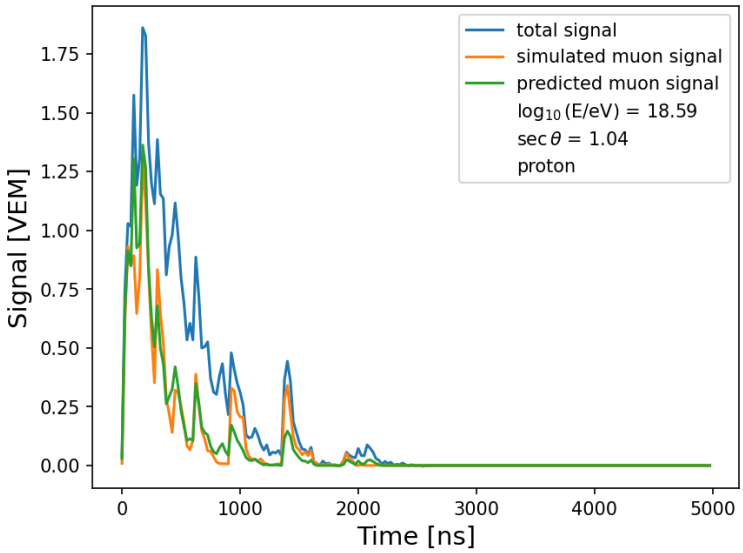


Figure 40 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an muon-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line.

A.2 Second model

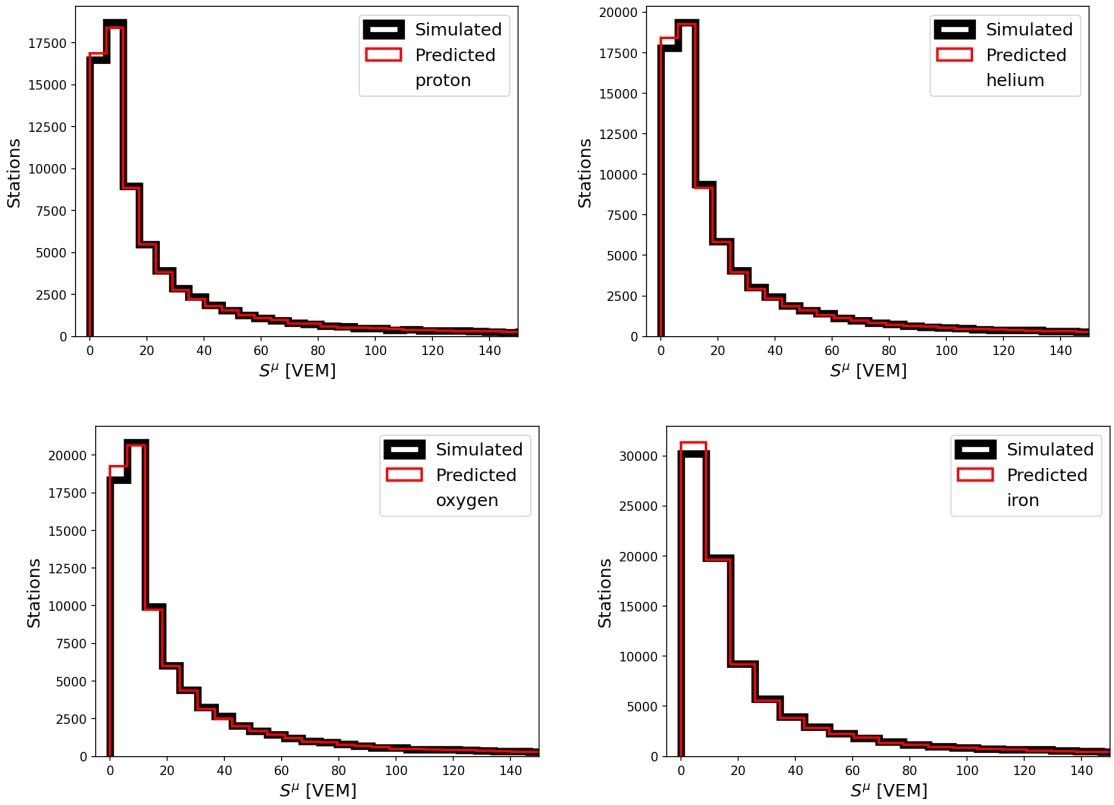


Figure 41 – Distribution of the simulated and predicted muon signals for every station in the dataset, separated by the primary cosmic-ray composition.

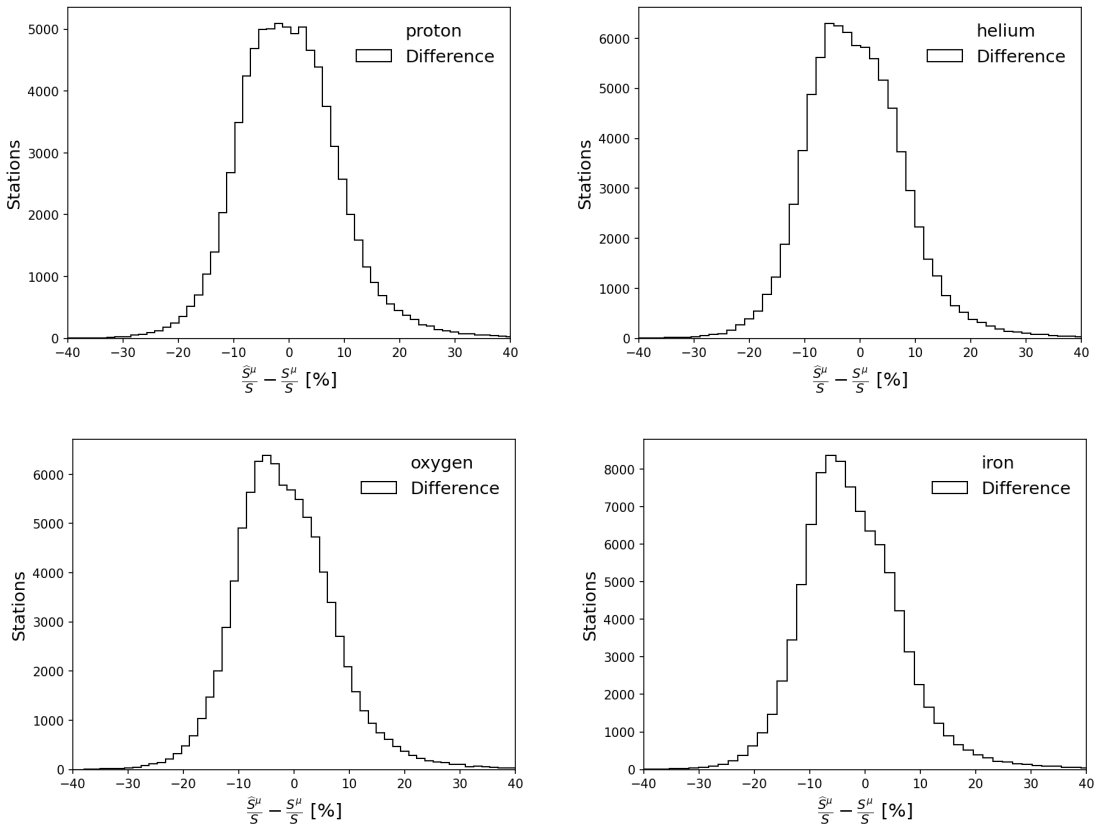


Figure 42 – Distribution of the difference between simulated and predicted muon signals for every station in the test dataset, separated by the primary cosmic-ray composition.

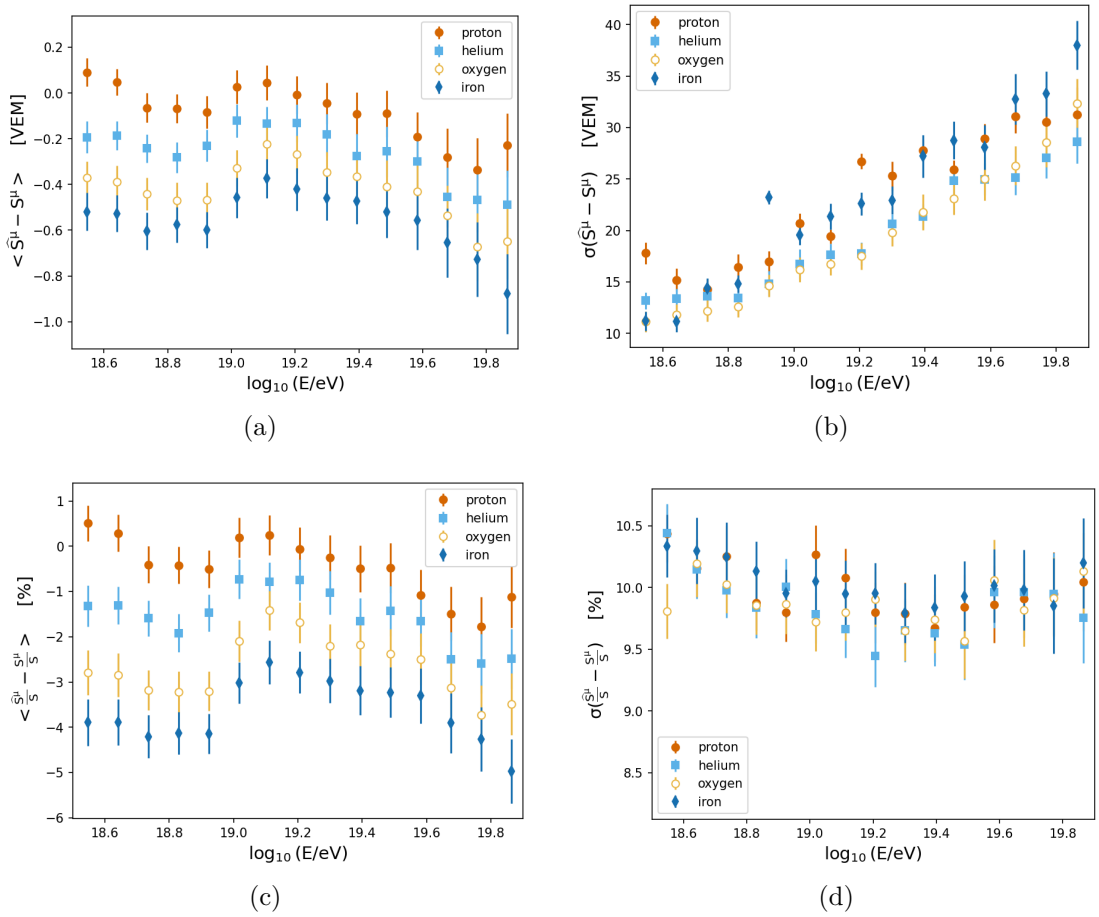


Figure 43 – Mean value (43a) and standard deviation (43b) of the difference between the predicted and simulated values as a function of the logarithm of the energy. Mean value (43c) and standard deviation (43d) of the difference between the predicted and simulated values divided by the total signal as a function of the logarithm of the energy.

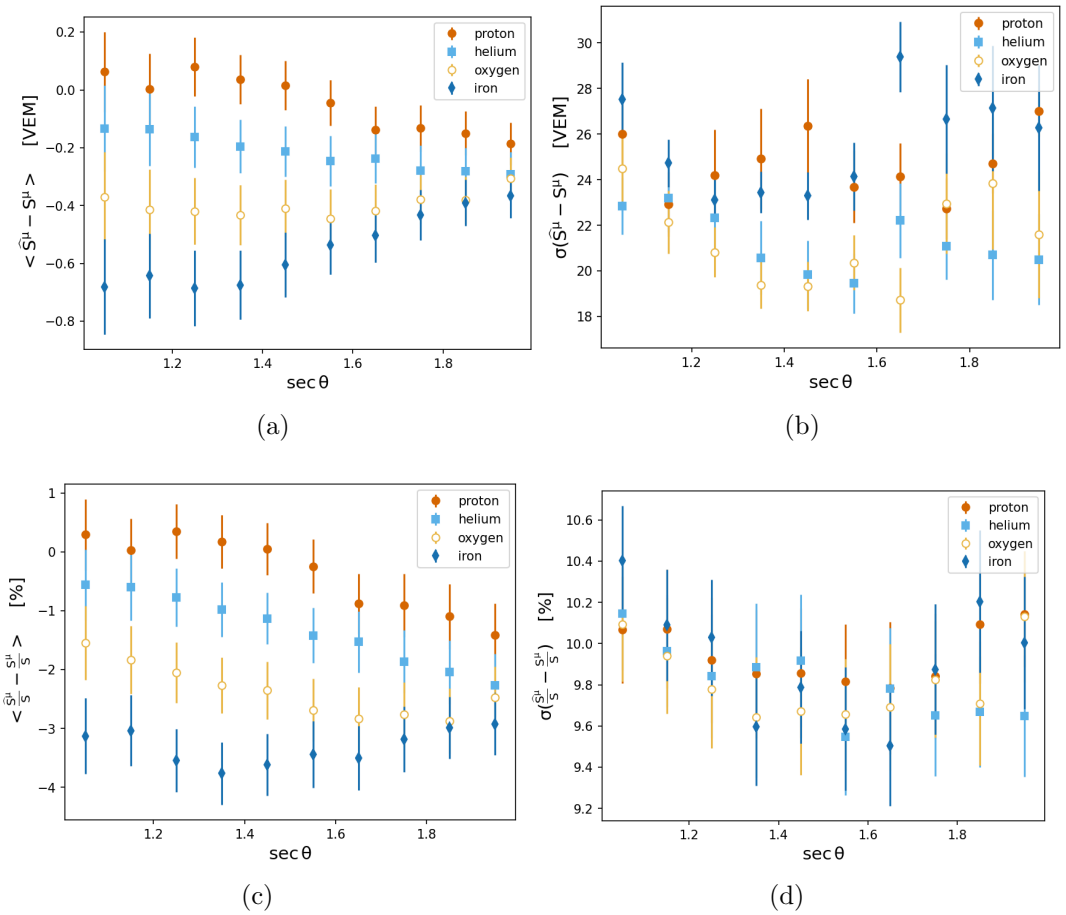


Figure 44 – Mean value (44a) and standard deviation (44b) of the difference between the predicted and simulated values as a function of the secant of the arrival angle. Mean value (44c) and standard deviation (44d) of the difference between the predicted and simulated values divided by the total signal as a function of the secant of the arrival angle.

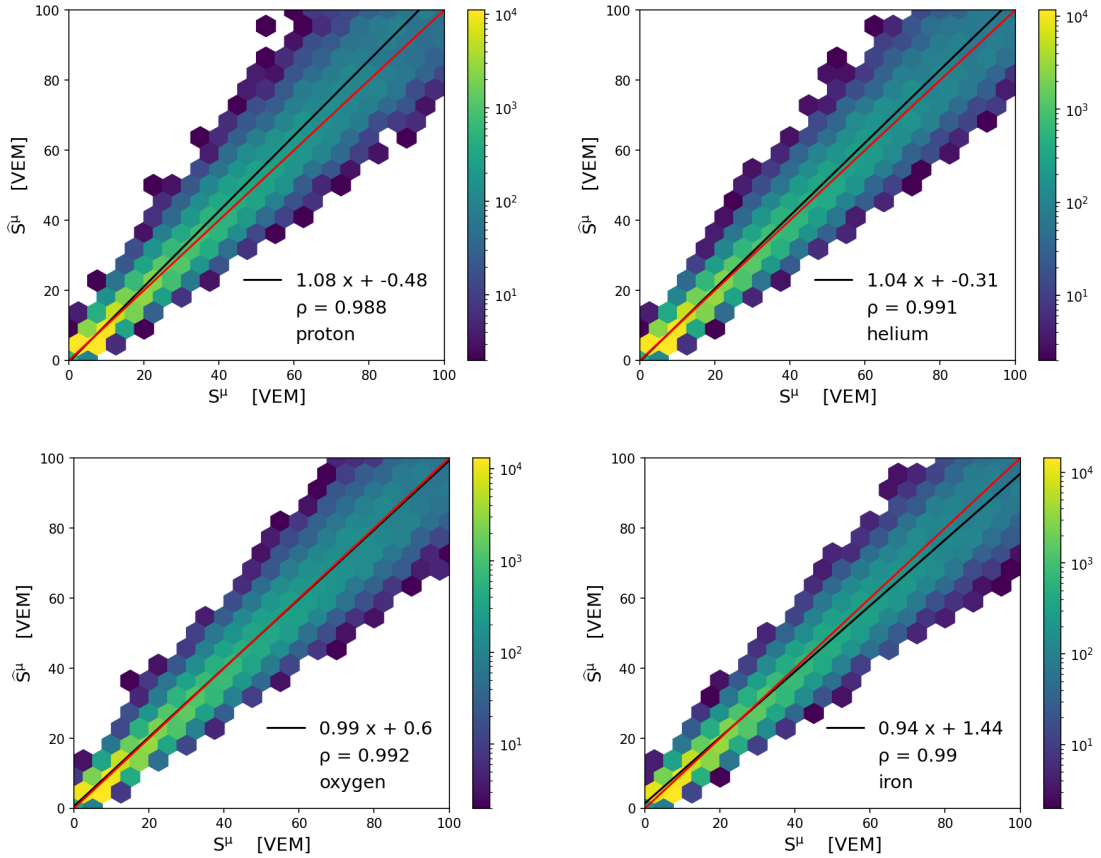


Figure 45 – Integral of the predicted muon signal as a function of the integral of the simulated muon signal. The black line corresponds to the linear fit of the points, and the red line corresponds to the ideal case.

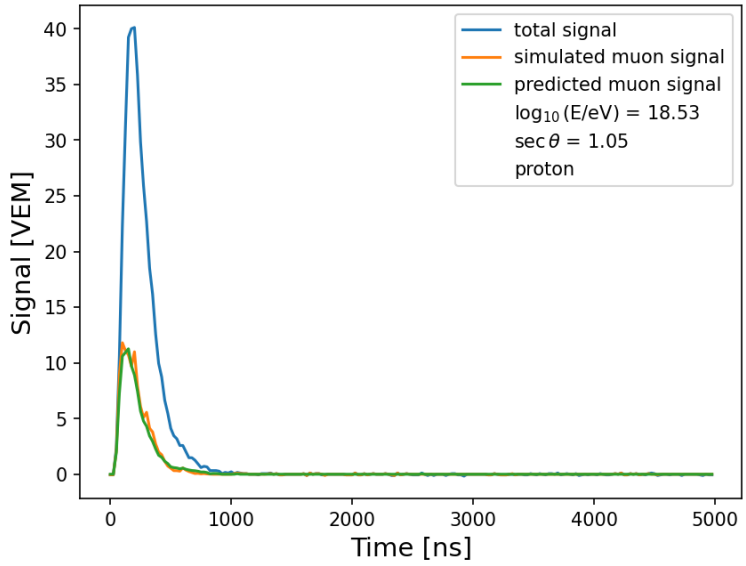


Figure 46 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an eletromagnetic-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line.

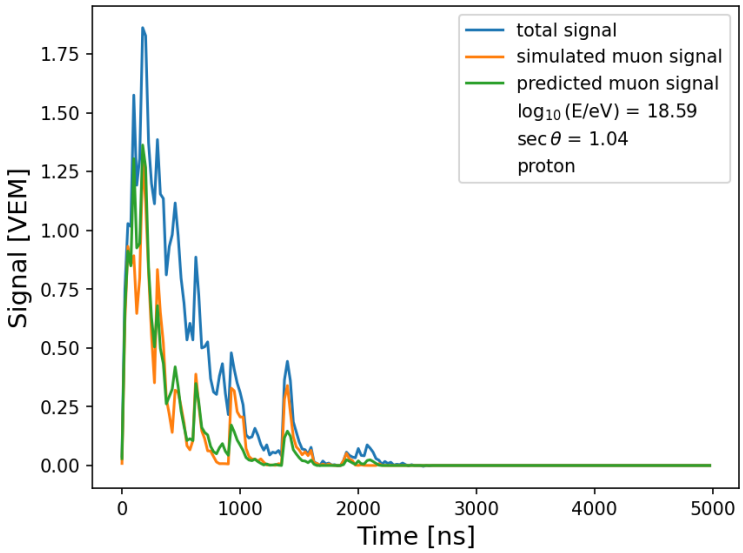


Figure 47 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an muon-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line.

A.3 Fourth model

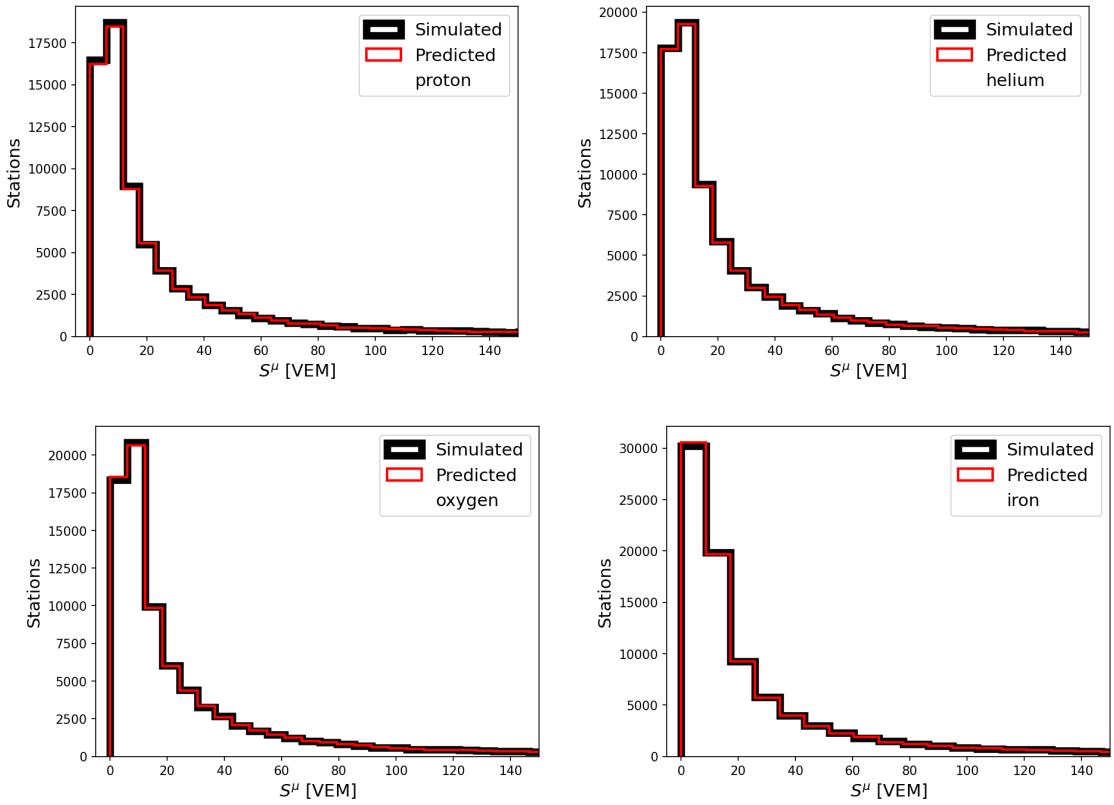


Figure 48 – Distribution of the simulated and predicted muon signals for every station in the dataset, separated by the primary cosmic-ray composition.

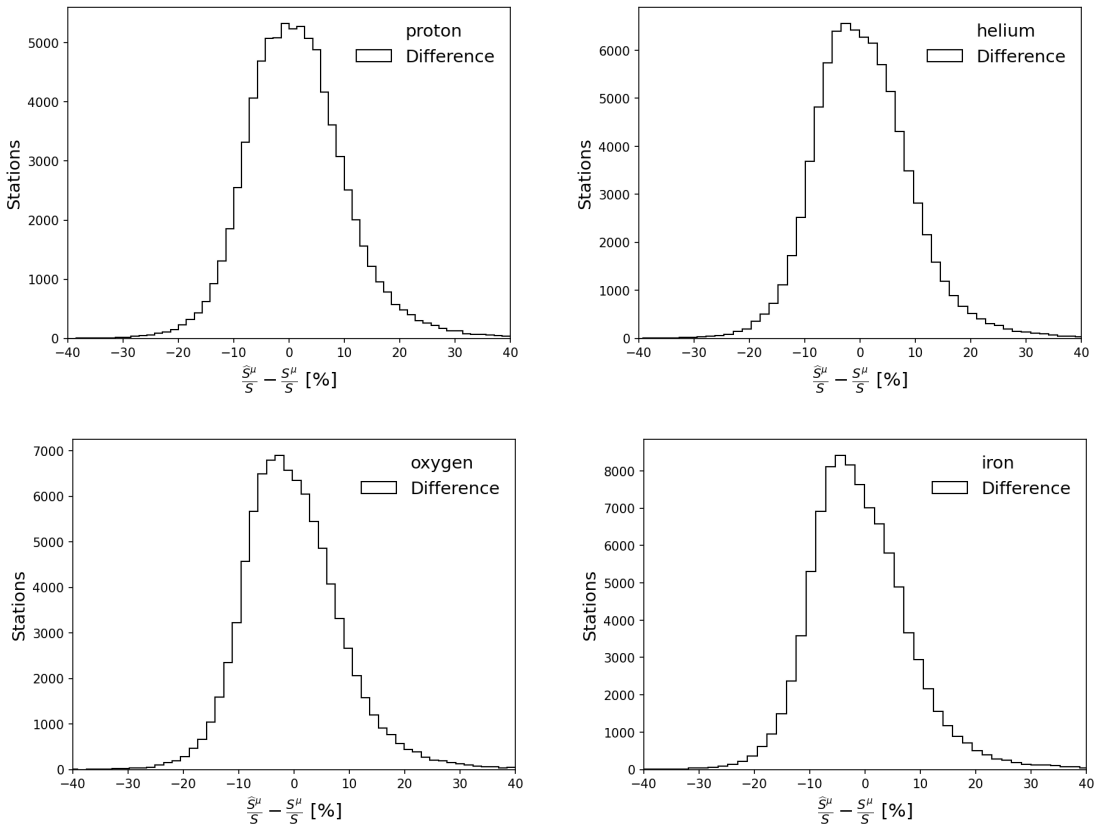


Figure 49 – Distribution of the difference between simulated and predicted muon signals for every station in the test dataset, separated by the primary cosmic-ray composition.

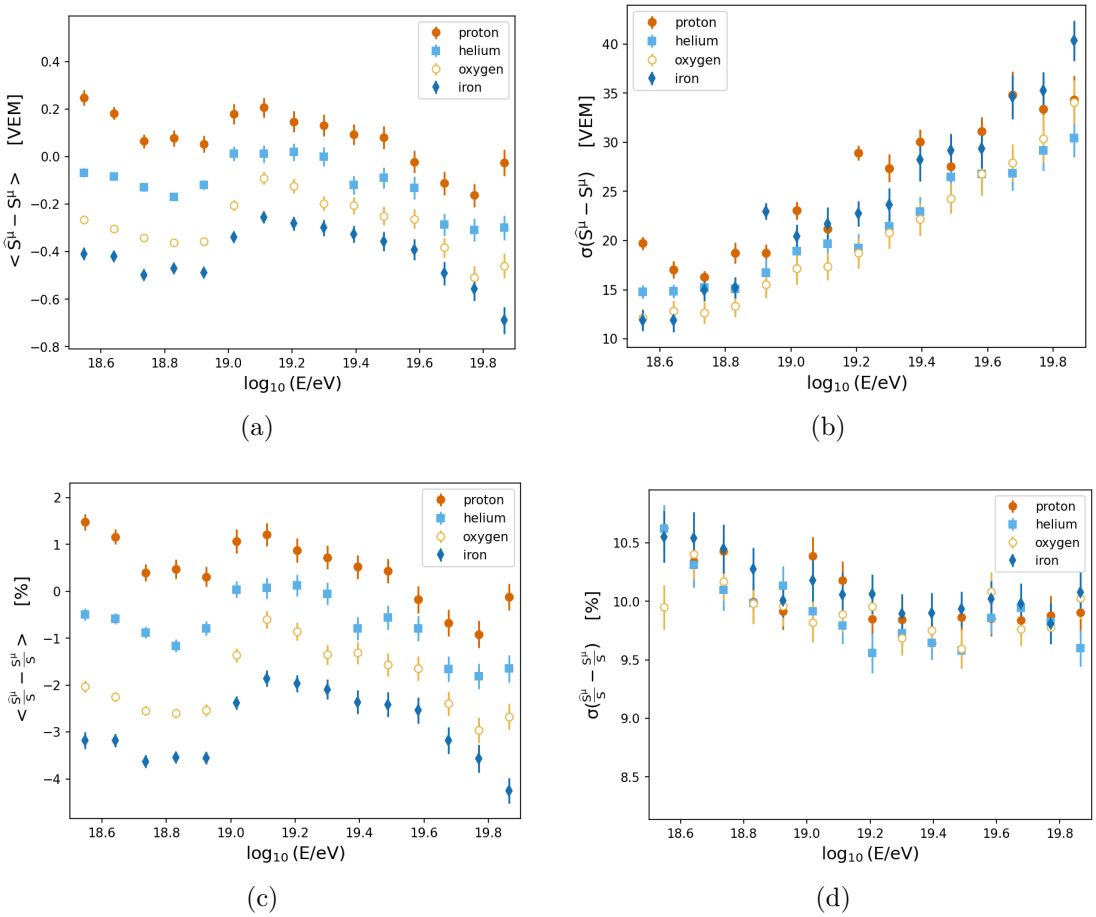
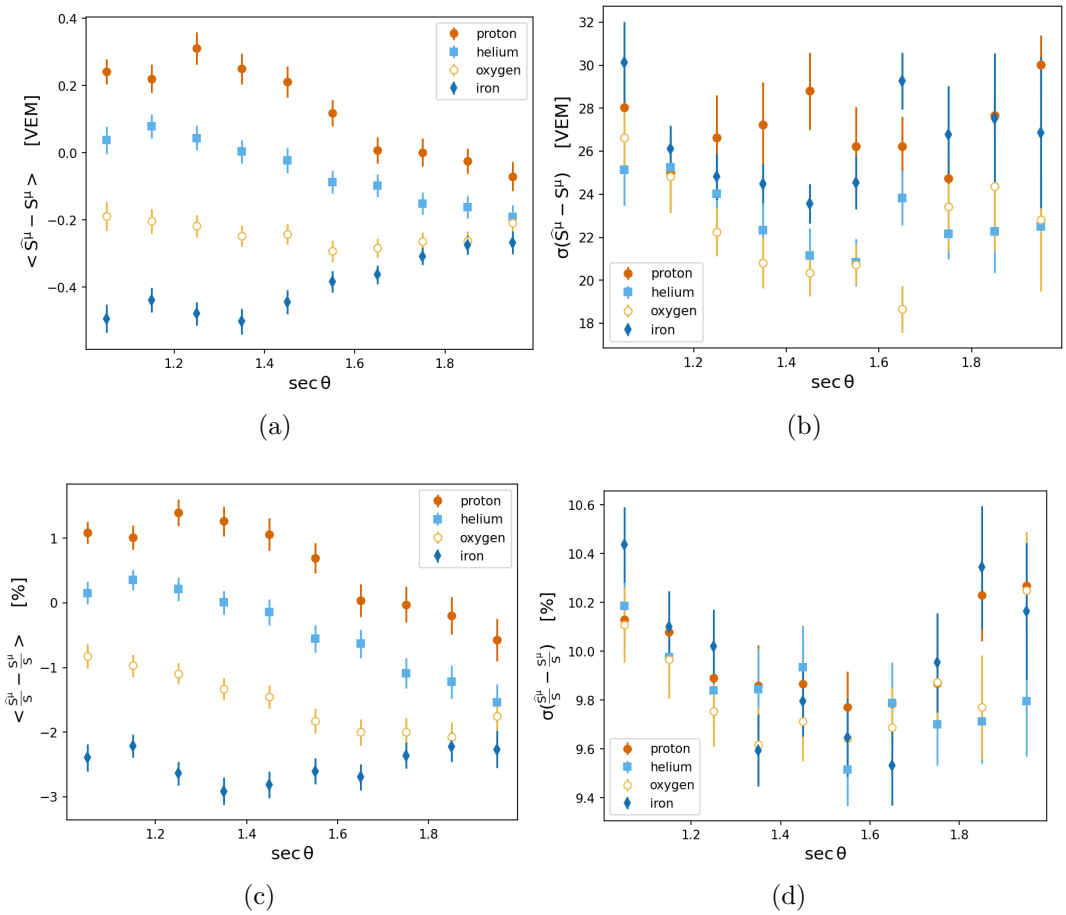


Figure 50 – Mean value (50a) and standard deviation (50b) of the difference between the predicted and simulated values as a function of the logarithm of the energy. Mean value (50c) and standard deviation (50d) of the difference between the predicted and simulated values divided by the total signal as a function of the logarithm of the energy.



(a) (b)

(c) (d)

Figure 51 – Mean value (51a) and standard deviation (51b) of the difference between the predicted and simulated values as a function of the secant of the arrival angle. Mean value (51c) and standard deviation (51d) of the difference between the predicted and simulated values divided by the total signal as a function of the secant of the arrival angle.

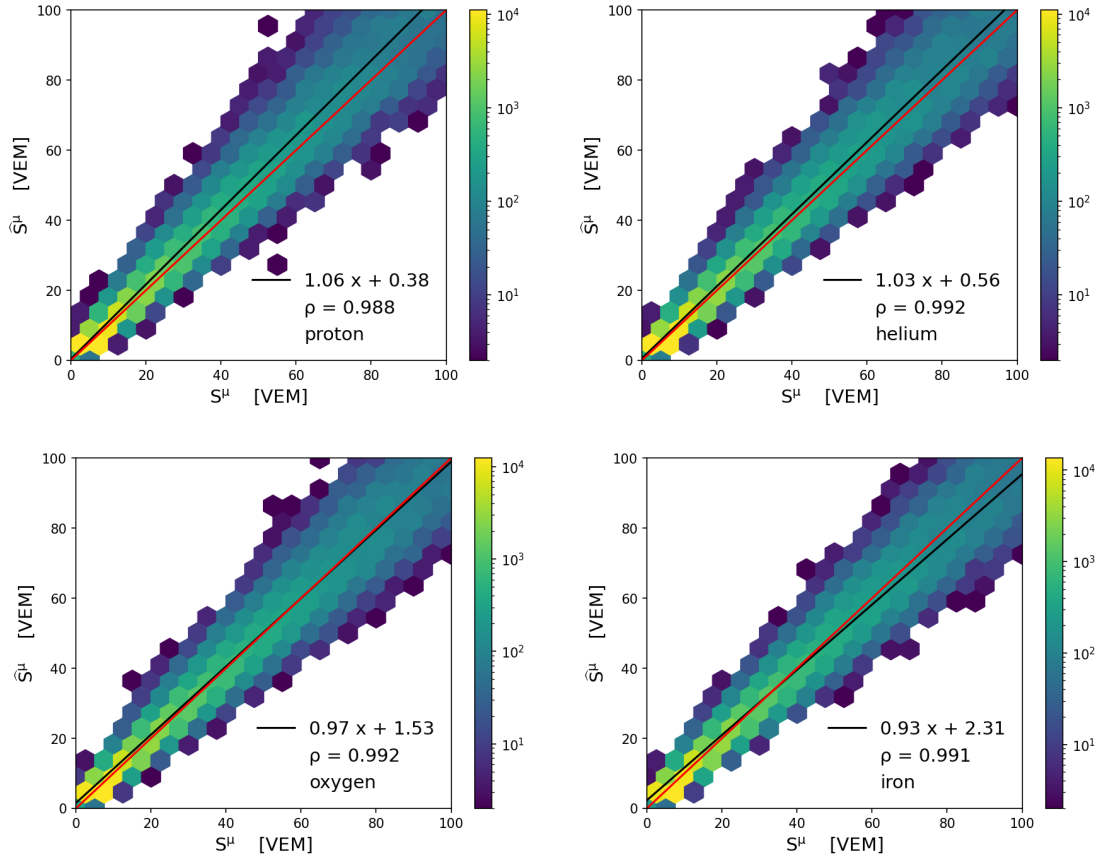


Figure 52 – Integral of the predicted muon signal as a function of the integral of the simulated muon signal. The black line corresponds to the linear fit of the points, and the red line corresponds to the ideal case.

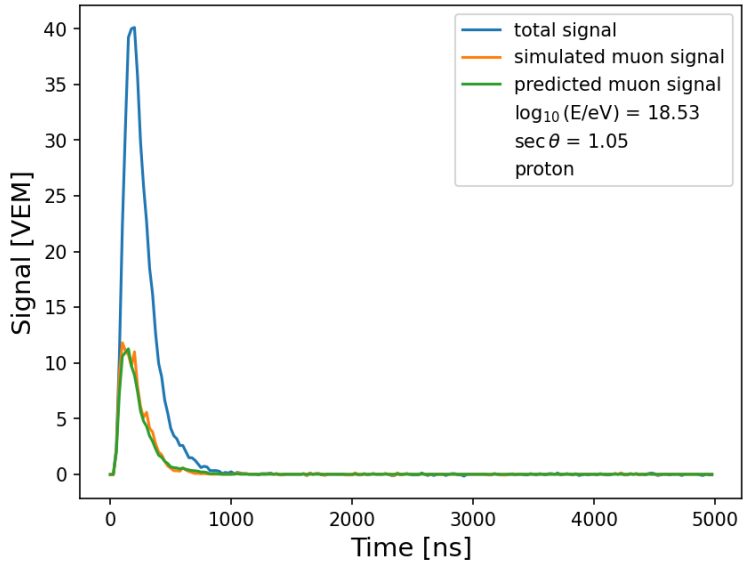


Figure 53 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an eletromagnetic-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line.

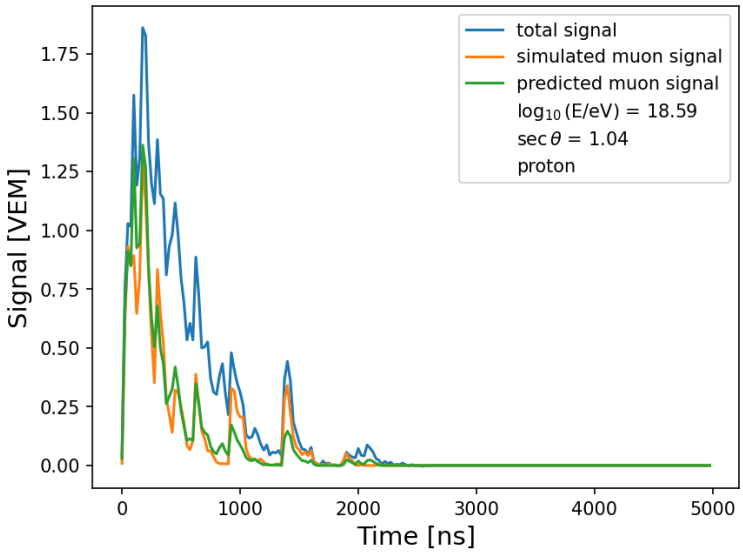


Figure 54 – Example of predicted muon trace for one simulated event with EPOS-LHC, for an muon-dominated signal. The simulated signal is represented by the blue line, the simulated muon signal is represented by the orange line, and the predicted muon signal is represented by the green line.