Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering

Department of Physics
Nuclear and Particle Physics

DIPLOMA THESIS

Reconstruction of the muon production depth of extensive air showers

Bc. Antonín Kravka

Supervisor:	Dr. Eva Maria Martins dos Santos
Consultant:	Dr. Alexey Yushkov
Year:	2023

České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská

Katedra fyziky
Jaderná a částicová fyzika

DIPLOMOVÁ PRÁCE

Rekonstrukce mionové produkční hloubky v atmosférických sprškách kosmického záření

Bc. Antonín Kravka

Vedoucí práce: Dr. Eva Maria Martins dos Santos
Konzultant: Dr. Alexey Yushkov
Rok: 2023

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA JADERNÁ A FYZIKÁLNĚ INŽENÝRSKÁ
PRAHA 1 - STARÉ MĚSTO, BŘEHOVÁ 7 - PSČ 115 19

Katedra: fyziky *Akademický rok:* 2021/2022

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: Bc. Antonín Kravka

Studijní program: Jaderná a částicová fyzika

Název práce: (česky) Rekonstrukce mionové produkční hloubky v atmosférických sprškách kosmického záření

Název práce: (anglicky) Reconstruction of the muon production depth of extensive air showers

Pokyny pro vypracování:

1) Studium literatury [1, 5] vztahující se na rekonstrukci mionové produkční hloubky (MPD) ve sprškách kosmického záření a aplikace strojového učení

2) Prozkoumání vzorových algoritmů strojového učení projektu VISPA zkonstruovaných za použití různých typů hlubokých neuronových sítí

3) Identifikace vhodného algoritmu strojového učení k rekonstrukci MPD pomocí pozorovatelných veličin dostupných z typických polí mionových detektorů

4) Studium systematických chyb algoritmu strojového učení v závislosti na volbě vysoko-energetického modelu hadronických interakcí pomocí simulací spršek kosmického záření vytvořených programem CORSIKA

5) Zkonstruování druhé hluboké neuronové sítě propojující rekonstruovanou MPD a fyzikální parametry relevantní k produkci mionů ve sprškách kosmického záření

Doporučená literatura:

[1] L. Cazón, R. A. Vazquez, A. A. Watson and E. Zas: Time structure of muonic showers, Astropart. Phys. 21 (2004), 71–86

[2] L. Cazón, R. Conceição, M. Pimenta and E. Santos: A model for the transport of muons in extensive air showers, Astropart. Phys. 36 (2012), 211–223

[3] A. Aab et al.: Muons in air showers at the Pierre Auger Observatory:

Measurement of atmospheric production depth, Phys. Rev. D 90 (2014), 012012

[4] A. Geron: Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, O'Reilly, 2019

[5] M. Erdmann, J. Glombitza and D. Waltz: A deep learning-based reconstruction of cosmic ray-induced air showers, Astropart.Phys. 97 (2018), 46-53

Jméno a pracoviště vedoucího diplomové práce:

Dr. Eva Maria Martins dos Santos, Fyzikální ústav AV ČR, v. v. i., oddělení astročásticové fyziky

Jméno a pracoviště konzultanta:

Dr. Alexey Yushkov, Fyzikální ústav AV ČR, v. v. i., oddělení astročásticové fyziky

Datum zadání diplomové práce: 01.03.2022

Termín odevzdání diplomové práce: 05.01.2023

Doba platnosti zadání je dva roky od data zadání.

.....
garant studijního programu

.....
vedoucí katedry

.....
děkan

V Praze dne 01.03.2022

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta jaderná a fyzikálně inženýrská

Břehová 7
115 19 Praha 1

PROHLÁŠENÍ

Já, níže podepsaný

Jméno a příjmení studenta:	Antonín Kravka
Osobní číslo:	468030
Název studijního programu (oboru):	Jaderná a částicová fyzika

prohlašuji, že jsem diplomovou práci s názvem:

Rekonstrukce mionové produkční hloubky v atmosférických sprškách kosmického záření

vypracoval samostatně a uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze dne 30.4.2023

.....

podpis

Acknowledgements

My most sincere thanks go to Dr. Eva Maria Martins dos Santos, as this work would not be possible without her exceedingly helpful guidance. I am grateful for her constant moral support and kindness, which helped me immensely both in my studies and, with equal importance, in the outside world. I would also like to thank Dr. Alexey Yushkov for all the helpful discussions and assistance I received over the last year. Finally, I would also like to express gratitude to my friends and family, whose support is the main reason I have gotten this far in my studies.

Bc. Antonín Kravka

Název práce:
Rekonstrukce mionové produkční hloubky v atmosférických sprškách kosmického záření

Autor: Bc. Antonín Kravka

Studijní program: Jaderná a částicová fyzika

Druh práce: Diplomová Práce

Vedoucí práce: Dr. Eva Maria Martins dos Santos
Fyzikální ústav AV ČR, v. v. i.

Konzultant: Dr. Alexey Yushkov
Fyzikální ústav AV ČR, v. v. i.

Abstrakt:
Atmosférické spršky kosmického záření jsou komplexními kaskádami částic, které vznikají při srážkách kosmického záření s jádry prvků v atmosféře. Prostřednictvím mionů vytvořených v takových sprškách lze získat důležité informace o chemickém složení kosmického záření a hadronových interakcích, ke kterým dochází na počátku vývoje spršek. Tyto informace lze získat skrze rekonstrukci Mionové Produkční Hloubky (MPD). Současný model rekonstrukce MPD je přizpůsoben sprškám šířených pod vysokými zenitovými úhly a mionům detekovaným daleko od jádra spršek. V této práci je představena nová metoda rekonstrukce MPD, využívající miony detekované v zakopaných detektorech a algoritmy strojového učení. Rozsah rekonstrukce je zde rozšířen na spršky šířené pod nižšími zenitovými úhly a miony dopadající blíže k jádru spršky. Použitelnost modelu pro různé energie a částice kosmického záření a odlišné modely hadronických interakcí je prozkoumána. Nakonec je představen druhý model strojového učení, jehož cílem je rekonstruovat spektrum energie mionů ve sprškách kosmického záření.

Klíčová slova: Kosmické záření, atmosférická sprška, mionová produkční hloubka, strojové učení, CORSIKA

Title:
Reconstruction of the muon production depth of extensive air showers

Author: Bc. Antonín Kravka

Field of study: Nuclear and Particle Physics

Thesis type: Diploma Thesis

Supervisor: Dr. Eva Maria Martins dos Santos
Institute of Physics of the Czech Academy of Sciences

Consultant: Dr. Alexey Yushkov
Institute of Physics of the Czech Academy of Sciences

Abstract:
Extensive Air Showers (EAS) are complex cascades of particles, emerging from collisions of cosmic rays with atmospheric nuclei. Muons created in EAS convey relevant information about the mass composition of cosmic rays and hadronic interactions occurring early in the EAS development. We can extract this information by reconstructing the Muon Production Depth (MPD). The existing method of the MPD reconstruction is tailored to EAS with high zenith angles and muons arriving far from the shower core. In this thesis, a new method of reconstructing the MPD is proposed, utilizing muons detected underground and machine learning algorithms. The reconstruction range is extended to low-zenith EAS while considerably reducing the present radial cut. We explore the model's applicability to different energies and species of cosmic rays and various models of hadronic interactions. Lastly, a second machine learning model is introduced, with the aim of reconstructing the long-evading muon energy spectrum in EAS.

Keywords: Cosmic Rays, Extensive Air Shower, Muon Production Depth, Machine Learning, CORSIKA

Contents

Introduction

Cosmic rays are a fascinating and mysterious phenomenon that has puzzled generations of scientists for over a century. These highly energetic particles are considered to be one of the most energetic occurrences in the Universe, and despite being studied intensively with increasingly complex instruments, their origins and acceleration mechanisms still remain open questions. Starting with their discovery in 1912 by Victor Hess, the increasingly deeper understanding of cosmic rays has led scientists to construct ever-evolving cosmic-ray detectors operating in space, the Earth's atmosphere, and both on and under the Earth's surface. We now live in a time period where it is possible to detect cosmic rays of energies higher than 10^{20} eV (or ~ 16 J), energies typically associated with macroscopic objects. As an example, the most energetic cosmic ray ever detected, the so-called "Oh-my-god particle", was registered on the 15th of October, 1991, by the Fly's Eye camera [1] in Utah and had an energy of approximately $(3.2 \pm 0.9) \times 10^{20}$ eV. However, since such highly-energetic particles are extremely rare, we are only able to detect them indirectly through the so-called Extensive Air Showers (EAS), cascades of secondary particles arising from collisions of cosmic rays with air molecules in the Earth's atmosphere. Due to the steeply falling spectrum of cosmic rays, where for $E \simeq 10^{18}$ eV, we only expect to detect less than one particle per square kilometer per year, giant detector arrays comprising a few thousand square kilometers are required to accumulate relevant statistics at the highest energies, considering an operation period of a few decades. Such experiments have been built on the Earth's surface, the most prominent being the Pierre Auger Observatory [2] and the Telescope Array [3]. Both experiments detect EAS using a hybrid detection technique, utilizing extensive ground-based detector arrays and fluorescence telescopes, resulting in the most precise cosmic-ray measurements currently available.

Extensive efforts are currently engaged in finding answers to the three major open questions of ultra-high energy cosmic ray physics, namely: what is the origin of the observed features of the cosmic ray spectrum, what is their nuclear mass composition, and, lately, what are the sources and acceleration mechanisms of such extremely energetic particles. Investigating the mass composition of cosmic rays is particularly important, as it can provide key information about their sources and propagation mechanisms, as well as for the improvement of our knowledge of hadronic interactions at energy scales above the LHC, which govern our most advanced simulations of EAS and are our best means of interpreting cosmic-ray data.

As of now, one of the key methods of inferring the cosmic-ray mass composition is the determination of a mass-sensitive observable called the depth of shower maximum

$X_{\max}$, which is defined as the atmospheric slant depth at which the number of particles in an EAS reaches its maximum. This observable can be directly measured by fluorescence telescopes, and it is the least model-dependent variable, i.e., the one with the lowest systematic uncertainties stemming from hadronic interaction models in mass composition studies. However, it suffers from low statistics (particularly at the highest cosmic-ray energies), as the duty cycles of fluorescence telescopes are typically about $\sim 15\%$. While other methods of inferring the mass composition or even $X_{\max}$ via ground detector arrays exist, utilizing their almost 100% duty cycle, the resulting observables need to be calibrated, since the signals from the ground detectors and fluorescence telescopes contain different composition of particles. See, for instance, [4, 5].

As an alternative to the classical $X_{\max}$ analyses, a reconstruction of the longitudinal profiles of muons in EAS was proposed in [6, 7]. Muons are the primary decay products of most hadrons in EAS and propagate through the atmosphere almost unattenuated, meaning that they carry information about their point of origin. The muonic longitudinal profile is defined as the distribution of the production points of muons in units of atmospheric slant depth, called the Muon Production Depth (MPD) distribution. The corresponding mass-sensitive observable, analogous to the $X_{\max}$ analyses, is the maximum in the MPD distribution $X^{\mu}_{\max}$.

The current method of the MPD reconstruction in EAS utilizes the relative arrival times of muons at the detectors, which are recorded with respect to the fastest EAS particles impacting the air shower ground epicenter and are therefore registered by the surface detector arrays. These time delays are predominantly caused by the geometric path traveled by muons in the atmosphere, and also by their energy, which is typically not measured by the ground detectors. As a consequence, the estimation of the energy-dependent part of the muon arrival time, called the kinematic delay, represents the method's largest source of systematic uncertainties, as it is currently parametrized from Monte Carlo simulations and only applicable to certain subsets of muons in EAS.

The objective of this thesis is to propose a new model of the MPD reconstruction based on machine learning algorithms, which are particularly suited for data-driven analyses and finding complex, high-dimensional relationships between related observables, which would otherwise be hard to find using traditional statistical methods. The proposed model setup is aimed at a future implementation at the AMIGA (Auger Muons and Infill for the Ground Array) upgrade, where muon detectors are currently being deployed at the Infill region of the Pierre Auger Observatory at a depth of 2.3 m. This depth provides a vertical overburden of $\sim 540\,\mathrm{g/cm^2}$, allowing for a pure measurement of the muonic signal. We aim to simultaneously improve and extend the MPD reconstruction to phase-space regions where the current reconstruction performs poorly or outright fails. To develop the model, we use Monte Carlo simulations of EAS, in which we study the performance of several machine learning algorithms on the reconstruction of the MPD. This allows us to select and investigate the best-performing model for the MPD analysis, a crucial step before applying the reconstruction to real-world data. As a second goal, we aim to use the reconstructed MPD for predictions of muon energies. If successful, this will allow us

to infer relevant information about the long-evading muon energy spectrum in EAS, and it might also shed more light on the hadronic interactions governing the EAS evolution.

The structure of this work proceeds as follows:

Chapter 1: As the introductory part of this work, an overview of basic cosmic ray and EAS features is outlined, covering the topics of the cosmic ray energy spectrum, mass composition and the behavior of various EAS components. Additionally, various detection techniques of EAS are mentioned, along with a brief summary of the current simulation methods.

Chapter 2: Focusing on the muonic component of EAS, the second chapter describes the features of the muonic longitudinal profiles and the underlying structure of the current MPD reconstruction method, alongside its applications and limitations.

Chapter 3: This chapter is dedicated to explain the machinery behind machine learning, its core concepts and the application possibilities of such algorithms. A significant portion of the chapter is aimed towards elucidating the Gradient-Boosted Decision Trees algorithm [8], the basis of the proposed MPD reconstruction model.

Chapter 4: In this chapter, we present the details of the proposed MPD reconstruction model and its performance on various samples of simulated EAS. We compare the model's reconstruction quality to the current MPD model and discuss its prediction capabilities in detail.

Chapter 5: The final chapter concerns the application of the proposed MPD model on predicting individual muon energies, implemented via construction of a second machine learning model. The results are briefly discussed and the prospect of such approach is evaluated.

Conclusion: In the summary part of this work, we discuss the overall performance of the proposed reconstruction models, along with their advantages and weaknesses, and outline the possible outlooks regarding their implementation to an array of buried muon detectors, as it is the case of AMIGA.

Chapter 1

Cosmic Rays & Extensive Air Showers

Cosmic rays are ultra-relativistic particles that propagate through outer space, reaching the Earth from outside the Solar System. For the majority of cosmic rays, their energies are predominantly in the GeV-TeV range, but, what scientists are mostly interested about, they can reach values exceeding 10^{20} eV, making cosmic rays one of the most energetic phenomena in the Universe. Nowadays, cosmic rays are classified to be predominantly atomic nuclei and, in much less numerous cases, particles as neutrons, electrons, and various antiparticles. The most abundant cosmic rays are protons ($\sim$ 86%) followed by helium nuclei ($\sim$ 12%), with heavier nuclei and the remaining particles accounting for only about 1%. Cosmic rays, together with gamma rays, neutrinos and gravitational waves, form the emerging field of Multi-messenger Astronomy, which is becoming one of the most exciting topics in modern astrophysics. In order to understand the fundamental origins of cosmic rays, astroparticle physicists have been studying their energy spectrum and mass composition for decades, delving ever deeper into their respective features. At the highest energies, both the energy spectrum and mass composition must be studied through cascades of particles, known as Extensive Air Showers (EAS), produced by interactions of cosmic rays with air molecules in the Earth's atmosphere. In this chapter, we summarize the main features of both the energy spectrum and mass composition subfields, and explore the EAS mechanisms, along with the detection and simulation techniques used in EAS measurements.

1.1 Cosmic-Ray Energy Spectrum

The energy spectrum of cosmic rays is one of the best-known characteristics in cosmic-ray physics, thanks to the combination of numerous experiments using complementary and independent detection techniques used for its measurement. The energy spectrum ranges from GeV-like energies to over 10^{20} eV or about 16 J, i.e., approximately the same energy carried by a baseball thrown at about 100 km/h. The cosmic-ray spectrum, comprised of various measurements from many cosmic-

ray experiments, is shown in Fig. 1.1. Even though it extends for about 12 orders of magnitude in energy and 10 orders of magnitude in flux, it can be approximately described by a power-law as

$$N(E)\mathrm{d}E \approx E^{-\gamma}\mathrm{d}E, \tag{1.1}$$

where the spectral index γ hardly deviates from the value of 2.7 over a wide energy range. Specifically, there are five energy regions where the spectral index changes its value, namely:

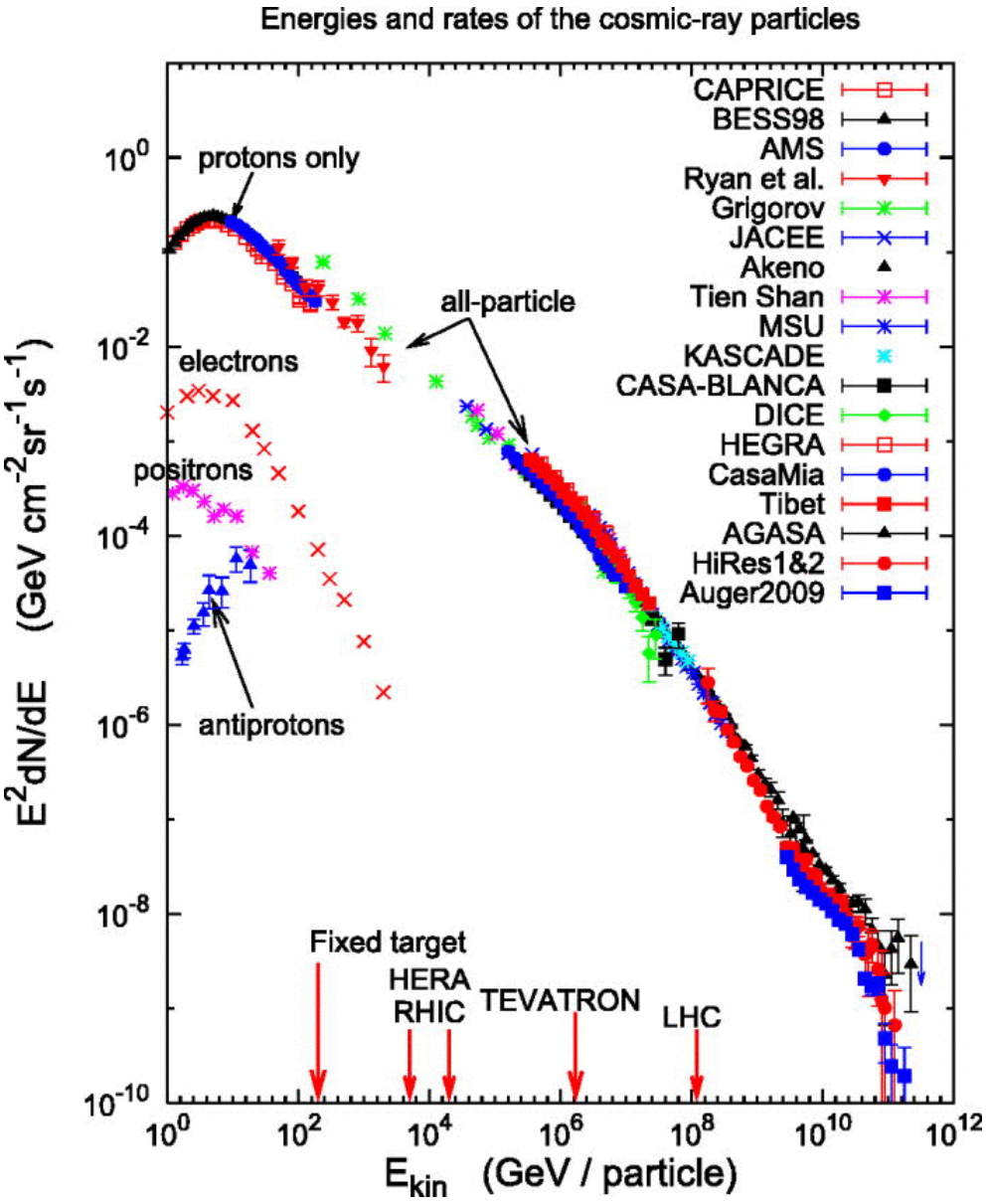

Figure 1.1: The energy spectrum of cosmic rays, composed of data from several experiments, taken from [9]. For comparison, nominal energies of major particle accelerators are highlighted.

- The **knee**: Between 10^{15} and 10^{16} eV, where the energy spectrum steepens $\implies \gamma \sim 2.7 \rightarrow 3.1$
- The **second knee**: At $E \approx 10^{17}$ eV, a further steepening of the spectrum is observed $\implies \gamma \sim 3.1 \rightarrow 3.3$
- The **ankle**: Located at $E \approx 10^{18.5}$ eV, the spectrum flattens, recovering its initial value $\implies \gamma \sim 3.3 \rightarrow 2.7$
- The **instep region**: An area between $10^{19.1}$ and $10^{19.5}$ eV, representing the beginning of the cosmic ray flux suppression at the highest energies $\implies \gamma \sim 2.7 \rightarrow 3.3$
- The **toe**: Located at $10^{19.5}$ eV, where a strong suppression of the cosmic ray flux is observed $\implies \gamma \sim 3.3 \rightarrow 5.1$

Presently, the reasons for the observed features in the cosmic ray spectrum are still largely unknown. The most accepted theories assume that these breaks arise due to changes in the cosmic ray nuclear mass composition, cosmic ray sources, acceleration and propagation mechanisms, or a combination of several factors. It is believed that the origin of the knee and second knee is linked to the acceleration limits of cosmic ray sources in our Galaxy. The most widely accepted explanation for this phenomenon is the Peters cycle [10], which attributes the knee steepening to the maximum energy at which cosmic ray protons can be accelerated by an astrophysical source, such as a supernova remnant shock. Similarly, the second knee indicates the maximum energy for heavier nuclei (such as iron) can reach through these mechanisms. Furthermore, cosmic rays with energies at this level should begin to leak out from our galaxy, causing a decrease in the cosmic-ray flux [11]. On the other hand, the ankle in the cosmic ray spectrum is typically attributed to the shift from a galaxy-dominated to an extragalactic-dominated population of cosmic rays. The reason for the suppression of the cosmic ray flux at the highest energies is still an unresolved issue, with two potential explanations. One proposes that the Greisen-Zatsepin-Kuzmin (GZK) cutoff [12, 13], caused by the interaction of cosmic ray protons with photons from the Cosmic Microwave Background (CMB) radiation, is responsible for the observed suppression. Protons with energies above 5×10^{19} eV lose energy when interacting with CMB photons over a mean free path of around 13 Mpc [12], meaning that the highest-energy particles would have to originate from nearby sources. The other proposition, supported by the Pierre Auger Observatory data [14], explains the suppression as a lack of nearby cosmic-ray sources with the ability to accelerate cosmic rays, with a mixed nuclear mass composition, up to the highest energies [15].

The observed steeply falling flux in the cosmic ray energy spectrum results in a need for various experimental methods to detect cosmic rays within certain energy ranges. Values of the cosmic ray flux at several energies are shown in Table 1.1. Below 10^{14} eV, the flux is large enough to allow for a direct detection of cosmic rays (realised by various complex detectors placed on either a high-altitude balloon or a spacecraft). The knee of the cosmic ray spectrum marks the transition between

direct and indirect cosmic ray detection (illustrated by the gap in measurements in Fig. 1.1), with the indirect detection realised by measuring Extensive Air Showers.

E_{CR}	Flux
10^9 eV	10000 particles/m^2/s
10^{12} eV	1 particle/m^2/s
10^{16} eV	1 particle/m^2/yr
10^{19} eV	1 particle/km^2/yr
10^{20} eV	2 particles/km^2/millennium

Table 1.1: Approximate values of the cosmic ray flux for different cosmic ray energies.

1.2 Extensive Air Showers

From the interaction of a cosmic ray in the Earth's atmosphere, typically with a Nitrogen or an Oxygen molecule, many millions of secondary particles are generated in a cascading process, called an Extensive Air Shower. In the first interaction (a nucleon-nucleus or a nucleus-nucleus collision), a handful of hadrons is generated, initiating a hadronic cascade, which, in its turn, will ignite the whole development of the EAS. Depending on the particle's type and energy, different processes (mostly re-interaction, decay, and bremsstrahlung) will contribute to the production of more secondary particles. Each secondary particle carries a fraction of the initial energy of the incident cosmic ray and, as the shower develops, the number of secondary particles increases, while the average energy per particle decreases. The shower maximum occurs when the EAS reaches the maximum number of particles. After the shower maximum, the number of secondary particles attenuates, primarily due to absorption in the atmosphere and ground. This effect is magnified if an EAS propagates under a non-zero zenith angle θ[1], as shower particles lose more of their energy by traversing a larger distance through the atmosphere.

We can divide an EAS into four main components: The electromagnetic (EM), the hadronic, the muonic and the neutrino (with the neutrino component not being measured in classical EAS experiments and thus not taken into account in the following discussions). The individual components differ both in their development (i.e. are governed by different interaction processes) and their sizes (i.e. number of particles, fraction of the total EAS energy). Typically, following the interaction of a cosmic ray with an air molecule, the newly formed hadrons either interact with different air molecules, producing more hadrons (constituting the hadronic component), or decay, producing muons, electrons, or photons. This way, all the other components are fed from the hadronic component, as depicted in Fig. 1.2.

The development of an EAS in the atmosphere can be reasonably well described by the Heitler-Matthews model [17]. This model assumes that the only hadrons

[1]The angle between the direction of the cosmic ray and a direction perpendicular to the Earth's surface, at which the EAS propagates through the atmosphere.

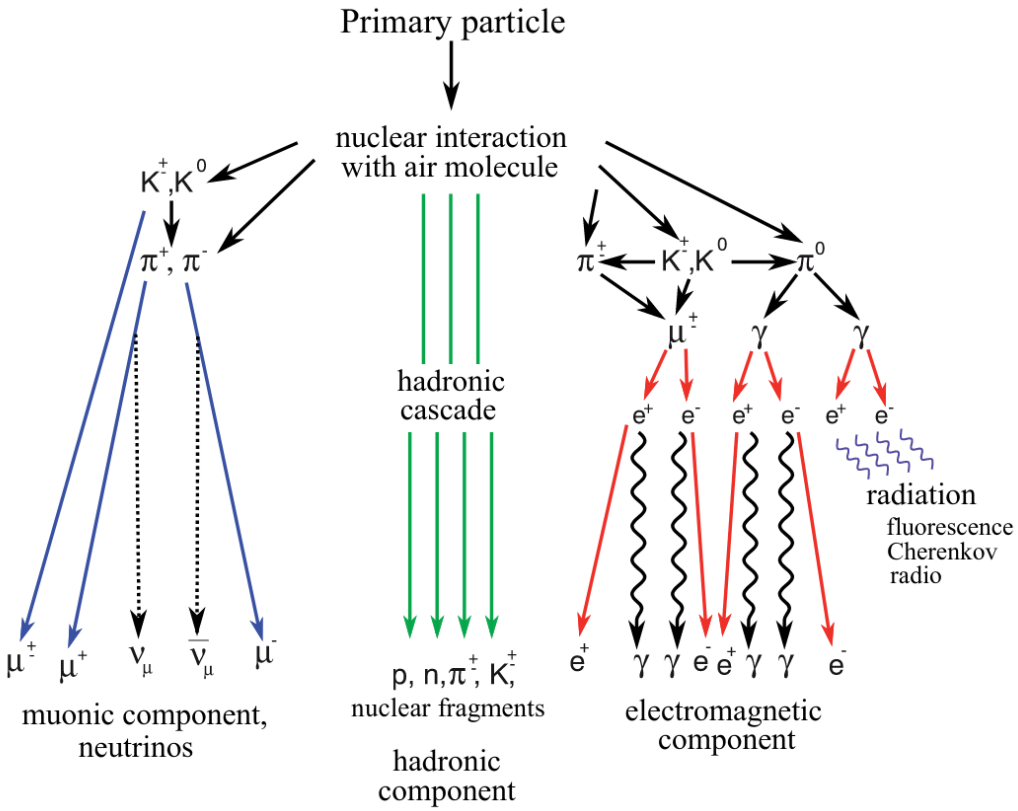

Figure 1.2: Illustration of an EAS development, taken from [16].

created are pions, which draws from the fact that pions are the dominant hadron species in an EAS, with roughly 10 times less-frequent kaons [11] being the next most abundant hadrons. The model further assumes that positive, negative, and neutral pions are created in equal numbers, the ratio of charged to neutral pions thus being 2 to 1, respectively. In the following sections, we will describe the individual EAS components and underlying mechanisms, complemented by the predictions of the Heitler-Matthews model.

The development of particle showers is typically expressed in units of traversed matter density per unit area, called the slant atmospheric depth X. In cosmic ray physics, the slant atmospheric depth yields the total matter traversed by the shower particles, and it is defined as follows:

$$X = \int_h^\infty \rho(z)\mathrm{d}z, \quad [X] = \mathrm{g\ cm^{-2}}, \tag{1.2}$$

where ρ is the atmospheric density measured from an altitude h along the distance z traversed by the shower particles. The slant depth is the natural unit used in measurements of the longitudinal profiles of air showers and expresses quantities such as the number of particles, or the energy deposited by the shower particles in the atmosphere as a function of X. The longitudinal profiles are heavily utilized in mass composition studies, since different cosmic rays give rise to different longitudinal profiles, as can be seen from Fig. 1.3.

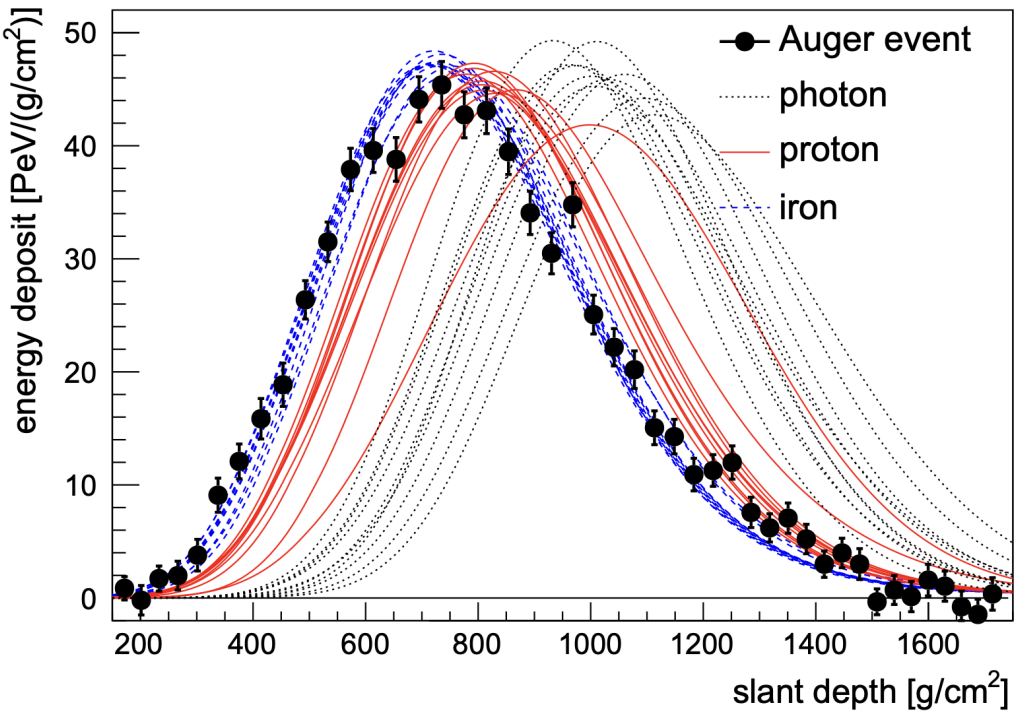

Figure 1.3: Example of longitudinal profiles of air showers initiated by various primary particle types. The data points represent the measurements from the fluorescence detectors of the Pierre Auger Observatory and the lines stand for simulated showers (see [18] for data citations).

1.2.1 Electromagnetic Cascades

The EM component consists of electrons, positrons, and photons. It is mostly fed by the decay of neutral pions, which, having a mean lifetime of around 8×10^{-17} s, almost immediately decay into two photons. In the vicinity of air molecules (or rather their respective nuclei), photons with enough energy can create electron-positron pairs, which, in turn, create additional photons by bremsstrahlung. In far less-occurring cases, the EM particles can feed the hadronic and the muonic component by photoproduction and muon pair production, respectively.

During the EAS development, around 90% of the cosmic ray energy is transferred to the electromagnetic cascade and is subsequently dissipated in the atmosphere. The EM component is also, by far, the most numerous of the three components, representing up to 99% of all particles produced in an EAS (with a ratio of roughly 9 photons to one electron/positron) [19]. By the time an EAS reaches the ground level, this number can severely change, depending on the incident zenith angle: For very large zenith angles ($\theta > 70°$), the EM component gets exponentially attenuated[2] in the atmosphere, becoming almost absent at the ground level.

The main features of the EM cascades can be relatively well described by the Heitler model [20]. According to this model, the EM particles undergo n two-body splittings,

[2] Assuming the Earth's atmosphere is flat, X grows with $\sec\theta$.

each after traversing a particle's radiation length λ_r (see Fig. 1.4). Each particle inherits half of the parent particle's energy and after n splittings, the number of particles reaches 2^n. The particle creation ceases when the particles' energies drop to a critical energy ξ_c^{EM}. For EM cascades, this is a threshold where radiation losses of the EM particles become less important then collisional losses. In air, $\xi_c^{EM} \approx 85$ MeV.

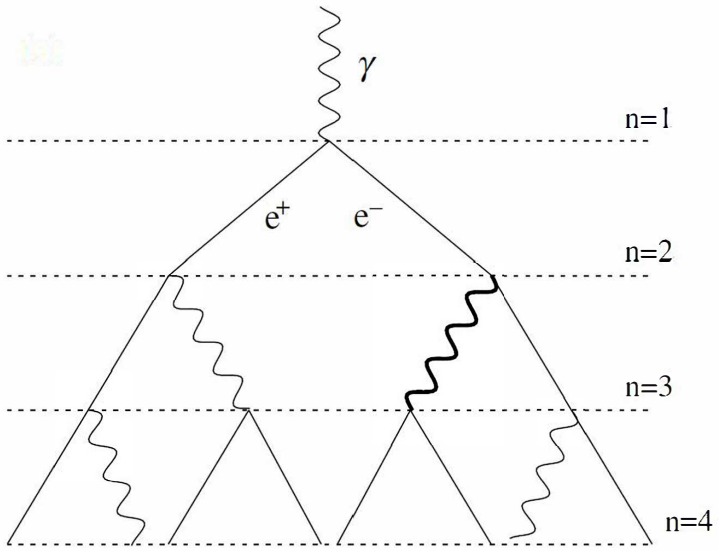

Figure 1.4: Schematic illustration of the Heitler model of the EM cascade development, taken from [17].

The model does not treat particles after reaching the critical energy, but, nevertheless, it still uncovers two important features of EM cascades:

- The maximum number of particles in an EM cascade N_{max} is proportional to the energy of the cosmic ray $E_0 \implies N_{max} = \dfrac{E_0}{\xi_c^{EM}}$
- The depth of the EM cascade maximum X_{max}^{EM} is proportional to the logarithm of the energy of the cosmic ray $\implies X_{max}^{EM} = \lambda_r \ln\left(\dfrac{E_0}{\xi_c^{EM}}\right)$

1.2.2 Hadronic Cascades

Charged hadrons may either interact several times with atmospheric nuclei, feeding the hadronic cascade with more hadrons and other secondary particles, or decay. As for the EM case, the number of particles in a hadronic cascade increases until a critical energy is reached, marking a maximum in the development of the hadronic component. For hadrons, the critical energy is defined as the threshold, below which the probability of a hadron to decay surpasses the probability of a further interaction. The probability of a hadron decaying is also dependent on the atmospheric depth

- since the air density decreases exponentially with altitude, the probability for a hadron to decay before interacting is higher in the upper atmosphere. After reaching the critical energy, the hadronic cascade thins out, transferring its energy into the EM and the muonic components. The main decay modes of pions and kaons (as the most numerous hadrons), along with their branching rations, are shown in Tab. 1.2.

$\pi^{\pm}$ ($\tau \approx 2.6 \times 10^{-8}$ s)	BR	$K^{\pm}$ ($\tau \approx 1.238 \times 10^{-8}$ s)	BR
$\longrightarrow \mu^{\pm} + \nu_{\mu}/\overline{\nu_{\mu}}$	0.99988	$\longrightarrow \mu^{\pm} + \nu_{\mu}/\overline{\nu_{\mu}}$	0.6351
$\longrightarrow e^{\pm} + \nu_{e}/\overline{\nu_{e}}$	0.00012	$\longrightarrow \pi^{\pm} + \pi^{0}$	0.2116
		$\longrightarrow \pi^{\pm} + \pi^{\pm} + \pi^{\mp}$	0.0559
		$\longrightarrow \pi^{0} + e^{\pm} + \nu_{e}/\overline{\nu_{e}}$	0.0482
		$\longrightarrow \pi^{0} + \mu^{\pm} + \nu_{\mu}/\overline{\nu_{\mu}}$	0.0318
		$\longrightarrow \pi^{\pm} + \pi^{0} + \pi^{0}$	0.0173

Table 1.2: Decay modes of charged pions and kaons, along with their rounded branching rations (BR) and mean lifetimes τ.

The Heitler-Matthews model provides an extension to the Heitler model, with the goal of describing the main features of hadronic cascades in EAS. The model divides the atmosphere into layers of thickness $\lambda \ln(2)$, where λ is the pion interaction length. After traversing one layer, a proton of energy E_0 produces k new pions (with the same energies), from which two thirds are charged and one third neutral (as shown in Fig. 1.5). Neutral pions immediately decay into two photons, feeding the EM component, while the charged pions traverse another atmospheric layer and repeat this process. After n generations, the amount of energy transferred to the EM component is:

$$E_{\rightarrow EM} = E_0 \Big[1 - \Big(\frac{2}{3} \Big)^n \Big]. \tag{1.3}$$

Thus, after only four interactions (with four interaction lengths corresponding to roughly half the vertical height of the atmosphere), about 80% of the shower energy is transferred into the EM component. Charged pions are created until their energy drops below the critical energy ξ_c, subsequently decaying into muons. The critical energy decreases with the primary hadron energy, corresponding to 30 GeV at $E_0 = 10^{14}$ eV and 10 GeV at $E_0 = 10^{17}$ eV. Generally, at higher altitudes, the values of the critical energy are higher than the ones mentioned, so very inclined showers (i.e., $\theta > 70^{\circ}$), which develop in a less dense atmosphere for a prolonged time, have more energetic muons as a result of the decay of higher-energy pions.

The description above only applies for a simplified case, where the inelasticity of the collisions is not taken into account. To remedy this, the model introduces a parameter $\kappa < 1$, which represents a fraction of the available energy for particle production. Therefore, at each interaction, only κE_0 is distributed between new pions, modifying the relations accordingly.

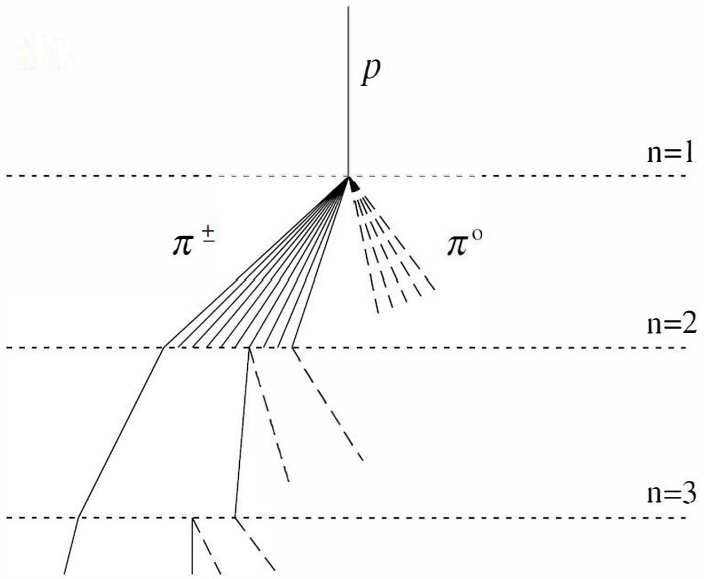

Figure 1.5: Schematic illustration of the Heitler-Matthews model, depicting the development of the hadronic cascade, taken from [17]

1.2.3 Muonic Component

As previously mentioned, practically all charged pions and a large portion of kaons ultimately decay into muons, creating a distinct shower component. A muon is a second generation lepton, with a rest mass of approximately 105.66 MeV, about 207 times heavier than the electron. Similarly, muon interactions are governed by Quantum Electrodynamics, interacting with other particles through the electromagnetic and weak forces. On the other hand, muons are unstable particles, with a mean lifetime of roughly 2.2 μs. At the speed of light, this would result in them travelling on average for about 660 m before decaying. However, due to the relativistic effects, namely, the time dilation (of the muon lifetime from the Earth's reference frame) and the length contraction (of the Earth's atmosphere from the muon's reference frame), many muons can traverse through the entire atmosphere, reaching detectors at the ground level and even several meters underground for the case of the most energetic muons.

Muons are the direct decay products of the shower hadrons and carry relevant information about the hadronic interactions in EAS. In the decay processes, muons receive, on average, about 79% of the parent pion or 52% of the parent kaon energy [21], while having two to three orders of magnitude longer decay lengths, as shown in Table 1.3. Although both the EM and the muonic components originate from the hadronic cascades, they differ significantly in their development. Since a muon is significantly heavier than an electron, its radiative losses by bremsstrahlung radiation are much smaller than those for electrons[3]. Additionally, muons have a smaller

[3]The intensity of bremsstrahlung radiation is inversely proportional to the square of the particle's mass.

cross-section for pair production and suffer less from multiple scattering than electrons, leading to a virtual non-existence of muonic cascades in air showers. While the EM cascade does not further influence the muonic component (apart from a rare photon conversion into a muon-antimuon pair), the lowest-energy muons, on the other hand, decay into electrons and positrons, marking another contribution to the EM cascade.

	γ []	λ_{decay} [m]
Pion - $E_{pion} = 1$ GeV	7,16	55,7
Pion - $E_{pion} = 100$ GeV	716,49	5 570
Kaon - $E_{kaon} = 1$ GeV	2,03	7,5
Kaon - $E_{kaon} = 100$ GeV	202,56	753
Muon - $E_{muon} = 1$ GeV	9,46	6 234
Muon - $E_{muon} = 10$ GeV	94,64	62 337

Table 1.3: Rounded Lorentz factors $\gamma = \dfrac{E}{m_x c^2}$ and mean decay lengths $\lambda_{decay} \approx \gamma \tau_x c$ for pions, kaons and muons of different energies.

According to the Heitler-Matthews model, all charged pions, whose energy falls below the critical energy, decay into muons. The number of muons produced in a hadronic cascade, as predicted by the Heitler-Matthews model, is expressed as:

$$N_\mu = \left(\frac{E_0}{\xi_c}\right)^\beta, \tag{1.4}$$

where $\beta = \ln n_{ch} / \ln n_{tot}$ represents a fraction of logarithms of the number of charged pions over the total number of pions produced in the hadronic cascade. The default value of β is assumed to be 0.85, while Monte Carlo simulations predict a shift of this parameter to higher values (0.92-0.95). This discrepancy in the β values can be remedied by the model by including the inelasticity of collisions and a slight tuning of its parameters. In any case, the number of muons grows slower than linearly with the cosmic ray energy, which is important for the treatment of air showers initiated by a heavier nuclei, as described below.

1.2.4 Superposition Model

The superposition model is a further extension to the Heitler-Matthews model and is used to describe the development of hadronic cascades for nucleus-induced showers, i.e., cosmic rays with atomic mass $A > 1$. According to the model, the energy E_0 of a cosmic ray is distributed equally among all of its A nucleons. The development of a hadronic cascade is then treated as a superposition of A independent showers, each with an initial energy of $\dfrac{E_0}{A}$, developing simultaneously in the atmosphere. This is a reasonable approximation, as for any nuclei, the binding energy per nucleon does not exceed 9 MeV, far less than average pion interaction energies in an air shower ($\sim$ 100 GeV). An EAS is then treated as a sum of A showers initiated by protons.

The number of muons N_μ^A in this air shower is thus modified to

$$N_\mu^A = A^{1-\beta} N_\mu^p, \tag{1.5}$$

where N_μ^p is the number of muons in a proton shower, governed by (1.4). This means that, for $\beta = 0.85$ (0.95), an air shower initiated by an iron nuclei is going to have 1.8 (1.2) times more muons than a proton-initiated shower of the same cosmic ray energy.

1.2.5 Atmospheric muons

Naturally, muons lose energy during their propagation through the atmosphere. The muon's critical energy in air, defined as the energy at which the ionization losses of muons are equal to their radiative losses, amounts to 1114 GeV (the value being averaged over the corresponding gasses) [22]. Comparing this value to the average muon energies at the sea level, depicted in Fig. 1.6, it can be seen that in the atmosphere, muons predominantly lose their energy by ionization. Additionally, most muons have energies between hundreds of MeV to tens of GeV, and, thus, they can be considered as minimum ionizing particles (MIP), losing on average around 1.8 MeV every g/cm^2. This can be formally written as:

$$\frac{\mathrm{d}E_\mu}{\mathrm{d}X} = -a, \tag{1.6}$$

where $a \approx 1.8\,\mathrm{MeV/g\,cm^{-2}}$. Accordingly, a muon propagating vertically throughout the whole atmosphere and reaching sea level loses approximately 1.8 GeV.

It is worth mentioning that the spectrum shown in Fig. 1.6 applies for all muons detected from an average of EAS with the same zenith angles, resulting in an average spectrum of muons. Muons from this spectrum are called atmospheric muons and are not correlated with muons from individual air showers. In all cases, atmospheric muons are considered as a background for individual EAS measurements. On the other hand, they can also be used to calibrate the trigger rates of air shower detectors.

After traversing through the atmosphere, particles from air showers continue to lose energy as they propagate through the ground. Roughly four meters of rock provide the same column depth as the whole atmosphere [24], resulting in a quick absorption of hadrons and EM particles, statistically leaving only energetic muons and neutrinos, which can be detected by underground detectors. The critical energy of muons in standard rock amounts to 693 GeV [22] and, therefore, ionization still remains the main process of energy loss for most muons traversing through the ground.

In more detail, the main processes, by which the underground muons dissipate their energy, are shown in Fig. 1.7. Ionization losses represent the most relevant process for muons with energies up to 100 GeV. Above this energy, radiative losses become relevant. These are bremsstrahlung (short dashed curve), direct pair-production (long

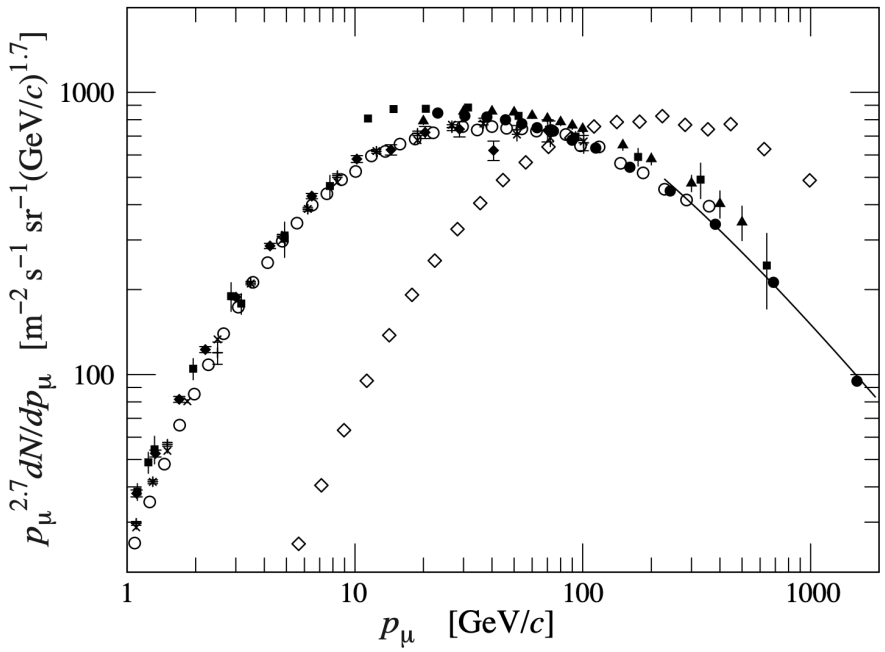

Figure 1.6: Average momentum spectra of muons at sea level, taken from [23]. All markers, except for $\diamond$, represent measurements of showers with $\theta = 0°$, while $\diamond$ stands for showers of $\theta = 75°$.

dashed curve), and photoproduction (dotted curve). Bremsstrahlung photons are emitted as a result of muons being deflected in the EM fields of the atomic electrons or nuclei. During photoproduction and direct pair-production, a muon emits a virtual photon, which either hadronically interacts with surrounding matter, creating new hadrons (photoproduction), or produces an electron-positron pair (direct pair-production). These radiation processes are relevant only for very energetic muons that reach deep into the Earth's crust, and therefore are not relevant in our next discussions.

1.3 EAS Detection

The detection of EAS can be performed using various techniques and detector types. The most common technique is to deploy an array of surface detectors, which measure the lateral density and arrival time distribution of the shower particles that reach the ground level. The footprint of an EAS grows with the cosmic-ray energy and, thus, at the highest energies, the lateral spread of the shower particles can be in the order of several kilometers. Given the strong suppression of the cosmic-ray flux at the end of the cosmic-ray spectrum, detection areas in the order of thousands of kilometers are required to gather relevant statistics. As more recent techniques, experiments have been exploiting the Cherenkov and fluorescence light, emitted by the shower particles in the atmosphere, to measure the longitudinal development of EAS.

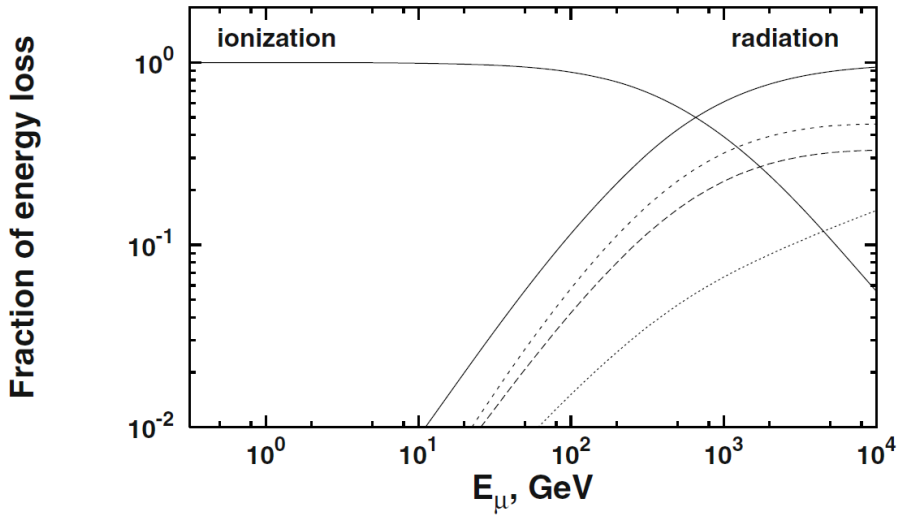

Figure 1.7: Ionization and radiation losses of muons underground with respect to the muon energy, taken from [24]. The dotted lines in the upward order represent photoproduction, bremmstrahlung and direct pair-production.

Below, the different detection techniques and types of detectors are described in greater detail.

1. **Surface detector arrays**: Most commonly comprised of scintillation detectors, but other experiments, like the Pierre Auger Observatory, opted for water-Cherenkov detectors (WCD). They are sensitive to charged shower particles, mostly electrons, positrons, and muons, and also to very high-energy photons that can convert to electron-positron pairs inside the WCD detectors. These detectors are deployed over large areas, forming a geometrical grid, from which the arrival time distributions and the particle density of the lateral distribution of EAS are measured. The reconstruction of the shower geometry (zenith and azimuthal angles) and the position of the shower core, i.e., the intersection point of the shower axis[4] with the ground, is obtained through different trigger times and the total signal measured by each detector.

   Examples of experiments utilizing ground arrays of scintillation and water Cherenkov detectors are the Telescope Array and the Pierre Auger Observatory, respectively. The latter is the world's largest cosmic ray detector, implementing a water-Cherenkov detector array comprising an area of 3000 km^2. Currently, the Pierre Auger Observatory is enhancing its surface detector array with complementary detectors, which will allow disentangling the electromagnetic from the muonic component of EAS. Namely, each WCD is being equipped with a scintillation detector on top and a radio antenna. Additionally, as part of the AMIGA (Auger Muons and Infill for the Ground Array) upgrade [25], scintillation detectors are being buried next to nearby WCD detectors in a smaller grid, which will allow a direct and independent measurement of the

[4] A continuation of the cosmic ray arrival direction in the atmosphere.

muon component of EAS. In the past, experiments like AGASA [26] used lead shielding on their scintillation detectors to absorb the electromagnetic component and directly measure the muonic component in the GeV-TeV range.

2. **Fluorescence telescopes**: Fluorescence light is produced in the atmosphere following the excitation of nitrogen molecules by the shower electrons and positrons. The fluorescence light is emitted isotropically and its detection by the telescopes depends on the atmospheric conditions. Fluorescence telescopes are used for the measurement of the longitudinal electromagnetic profiles of EAS, providing an almost calorimetric estimation of the EAS energy, provided that the fluorescence yield, which depends on the atmospheric conditions, is well estimated. Fluorescence telescopes are a part of hybrid observatories like the Pierre Auger Observatory or the Telescope Array, which can detect EAS using both techniques. Both experiments use the so-called hybrid events, i.e., EAS detected simultaneously by both techniques, to provide a complete picture of the EAS development, encompassing the lateral and longitudinal EAS profiles.

3. **Cherenkov telescopes**: Cherenkov light is produced whenever a charged relativistic particle propagates faster than the phase velocity of light in a medium. In EAS, the emission of Cherenkov light happens primarily due to the shower electrons and positrons, which comprise the bulk of the cascade. Unlike the fluorescence light, the Cherenkov light is collimated, and the detector must be aligned with the shower axis. Typically, the detection of EAS at the lowest energies is done via the Cherenkov light, as only above 10^{17} eV showers start to produce enough fluorescence light to be detected by this technique.

While the lateral development of EAS is measured by the ground arrays, their longitudinal development can be detected by Cherenkov or Fluorescence telescopes. Contrarily to the ground arrays, which have nearly 100% duty cycle, Cherenkov and Fluorescence telescopes can only operate during clear nights with only a small fraction of moonlight, which reduces their duty cycle to a maximum of $\sim$15%.

1.4 Mass Composition of Cosmic Rays

As opposed to the cosmic-ray energy spectrum, the relative abundances of various cosmic-ray nuclei are still unclear. To assess how various nuclei contribute to the overall mass composition, it is convenient to show the cosmic ray energy spectrum for individual cosmic ray species. Conventionally, the base spectrum is multiplied by a factor of $E^{2.6}$, which aids to enhance the distinct spectral features. We show this scaled spectrum, comprised of data from several experiments, for individual cosmic ray nuclei groups in Fig. 1.8.

The spectrum is shown for four elemental groups: proton, helium, oxygen, and iron, where, typically, intermediate and heavy mass elements are grouped into the oxygen and iron groups, respectively. While the lowest-energy cosmic rays are primarily

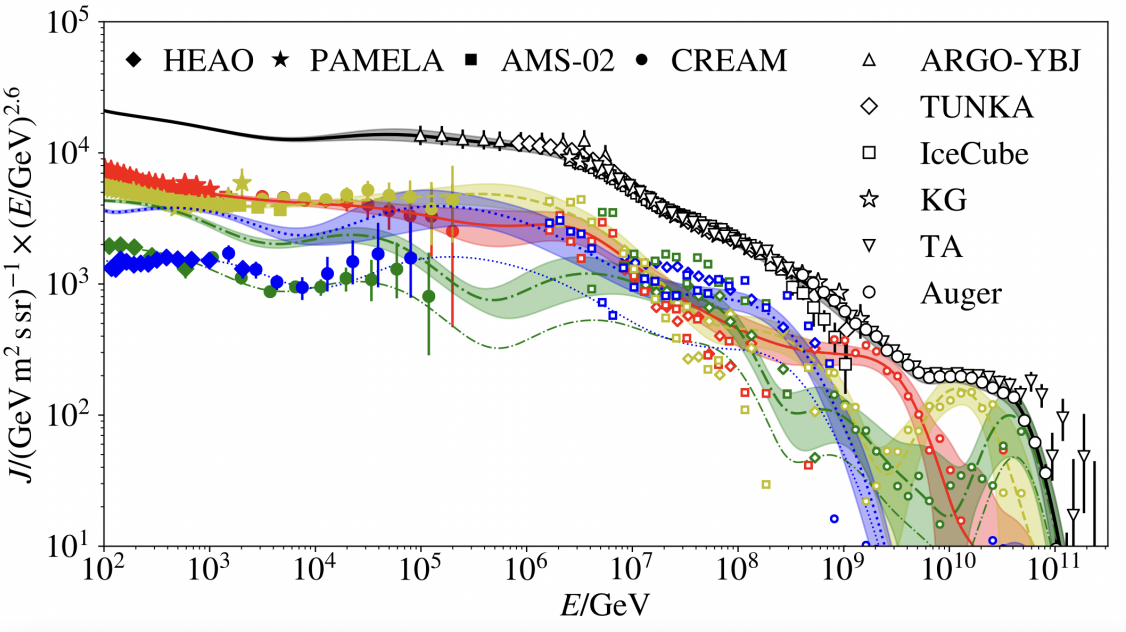

Figure 1.8: The energy spectrum of cosmic rays, comprised of experimental measurements (data points) and the data-driven Global Spline Fit model (lines and bands), see [27] for details and the data citations. The all-particle spectrum is in black, while colored objects stand for spectra of different elements/elemental groups: Red for protons, yellow for helium, green for oxygen/oxygen group and blue for iron/iron group. For oxygen and iron, fluxes of these nuclei and combined fluxes of nuclei with similar atomic number are depicted (combined fluxes have bands around the lines, which represent a model variation of one σ).

light nuclei, their mass composition evolves with increasing energy. Fig. 1.8 suggests that, with accordance to the Peters cycle, the proton and helium fluxes dip at the knee area, while the second knee is accompanied by a dip in the iron group. Another visible feature indicates the proton flux having a maximum and then dropping right before the ankle, such that the toe predominantly consists of heavier elements. At the highest energies, before the flux suppression, the mass composition seems to change dramatically, while the all-particle spectrum remains featureless. We note that, at the lowest energies, the nuclear mass composition of cosmic rays can be directly measured by space experiments. On the other hand, for EAS, the mass composition must be inferred from comparisons of air-shower observables with Monte Carlo simulations, which must precisely treat the development of EAS in the atmosphere and detectors' responses. The largest source of systematic uncertainties arises from the predictions of high-energy hadronic interaction models, which are extrapolated several orders of magnitude above the energies achieved at man-made particle colliders.

Currently, one of the most widely used observable for determining the mass composition of cosmic rays is the depth of the shower maximum, denoted as X_{max}, i.e., the atmospheric slant depth at which the development of EAS reaches its maximum. The determination of the mass composition using $\langle X_{\mathrm{max}} \rangle$ and $\sigma_{X_{\mathrm{max}}}$ observables in the energy range 10^{16} - 10^{20} eV, as measured by different experiments, is shown

in Fig. 1.9. There, it can be seen that both the $\langle X_{\mathrm{max}} \rangle$ and $\sigma_{X_{\mathrm{max}}}$ measurements suggest, first, a transition from a heavy to a light-dominated cosmic ray mass composition between the knee and the ankle regions, and again a shift towards a heavier composition for $E > 10^{19}$ eV. Complementary to Fig. 1.8, the observed transition of the cosmic ray mass composition between the knee and the ankle seems to indicate the exhaustion of the sources to accelerate, first, protons, then helium, and then successively heavier nuclei, as suggested by the Peters cycle.

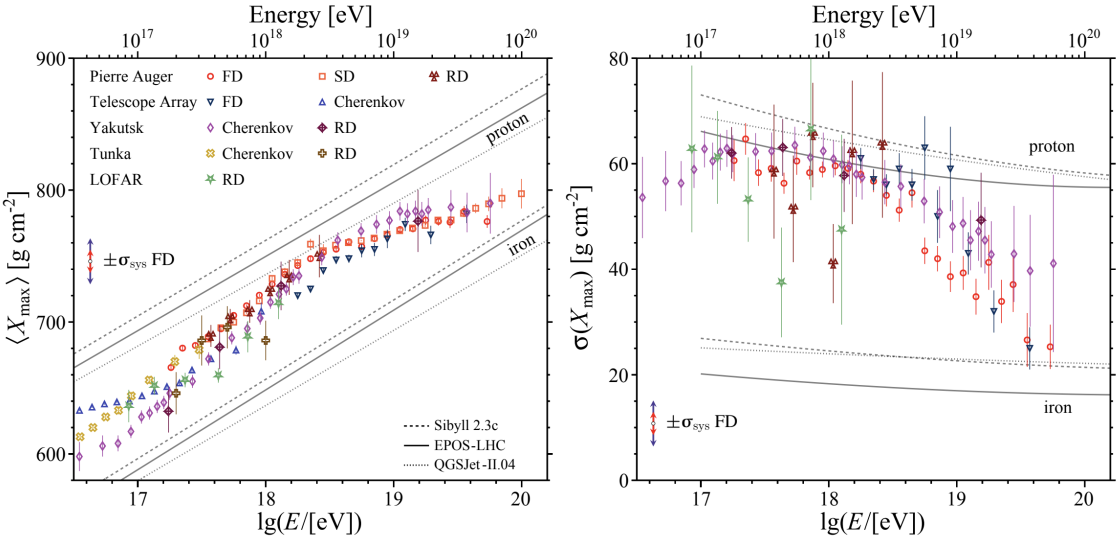

Figure 1.9: Measurements of $\langle X_{\mathrm{max}} \rangle$ (left) and $\sigma_{X_{\mathrm{max}}}$ (right) compared to the predictions for proton and iron nuclei made by hadronic interaction models Sibyll-2.3c, EPOS-LHC and QGSJET-II.04, taken from [28].

Due to the limited duty cycle of the fluorescence telescopes, the number of events collected at the $E > 10^{19.5}$ eV is not yet statistically significant to accurately determine the mass composition using the X_{max} measurements. It is possible to infer the mass composition from the lateral distribution of the shower particles, measured by the surface detector arrays. Having duty cycles of nearly 100%, and much less strict event selection criteria, allows scientists to utilize about one order of magnitude larger statistics than from fluorescence telescopes. However, the interpretation of the mass composition using these observables is often not consistent with the X_{max} measurements, as the high-energy hadronic interaction models account for the largest source of systematic uncertainties in the current analyses. Thus, to gain additional insight into the mass composition of cosmic rays, it is desirable to measure different mass-composition sensitive observables, which come with their own sets of statistics and systematic uncertainties. One example is the reconstruction of the longitudinal profiles of muons and their maxima X^{μ}_{max}, which will be the topic of the following chapters.

1.5 EAS Simulation

Data collected from EAS experiments can only be interpreted by resorting to air shower and detector simulations. These simulations are based on the Monte Carlo method and exploit all the relevant knowledge of Quantum Electrodynamics (QED) and Quantum Chromodynamics (QCD), combined with phenomenological modelling of the air shower development. While there is good agreement between data and theory in the description of the EM component (governed mostly by QED), the hadronic processes in EAS are not precisely described and cannot be calculated from the first principles. Consequently, various models of hadronic interactions need to be implemented in air shower analyses. Current high-energy models of hadronic interactions, most notably EPOS-LHC [29], QGSJET-II.04 [30] and Sibyll-2.3d [31], are calibrated to the data from man-made particle accelerators, particularly the Large Hadron Collider (LHC). However, in the context of high-energy hadronic interactions in EAS, it is not sufficient due to several reasons, namely:

- The highest center-of-mass energy currently reached at man-made colliders is 13.6 TeV in proton-proton collisions at the LHC. However, the center-of-mass energies in the highest-energy cosmic ray collisions are in the order of 100 TeV (the "Oh-My-God" particle registering more than 650 TeV), at least one order of magnitude above those of particle colliders.
- There is a significant lack of data regarding interactions where, in the energy range of interest, the bulk of particles are created in the very forward regions. Additionally, data from collisions of nuclei with oxygen or nitrogen are also scarce.

The hadronic interaction models must therefore be extrapolated in multiple ways, resulting in relatively large systematic uncertainties. Furthermore, differences between the models themselves, arising from different foundations upon which they are built, lead to different interpretations of experimental data. Therefore, an improvement of the current models of hadronic interactions is another major goal in the field of cosmic-ray physics.

One of the most known software for simulating the development of air showers is CORSIKA [32]. The cosmic ray community has been using and actively updating it for over 30 years. It incorporates much of the available knowledge regarding air shower propagation in the atmosphere, enabling studies of EAS initiated by almost any cosmic ray with energies ranging from 10^{12} eV to values larger than 10^{20} eV. Carefully tracking every particle in an EAS along its trajectory, the simulation process is comprehensive, albeit time-consuming, as the computing time scales approximately with the cosmic ray energy. To counteract the large computing time typical for modeling EAS, the process of "thinning" is implemented. It effectively selects only one particle from a subset of particles whose energies fall below an adjustable threshold, assigning it a corresponding weight and discarding the rest of the subset. This significantly decreases the required computing time, making it a widely used tool in EAS studies (more details are given in [32]).

In CORSIKA, positional and timing observables are described with the use of a Cartesian coordinate system, where the positive x-axis points to the magnetic north, the positive y-axis to the west, and the positive z-axis in the upward direction. In this configuration, the x-axis and y-axis define the so-called detector plane, while the zero-point of the z-axis is located at sea level. The zenith angle θ of an EAS is given by an angle between the primary particle momentum vector and the z-axis, while the azimuthal angle is measured between the positive x and the horizontal component of the primary particle momentum vector. For clarity, the coordinate system is shown in 1.10.

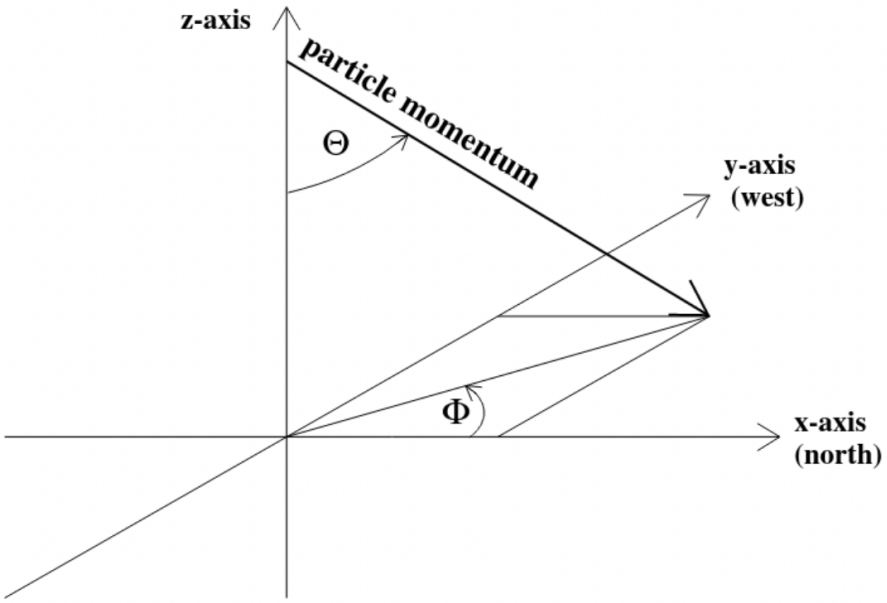

Figure 1.10: Default CORSIKA coordinate system.

The modelization of the atmospheric profiles in CORSIKA is represented by five atmospheric layers of varying density, resulting in five relations that connect the atmospheric altitude with the mass overburden $T(h)$, specifically:

$$T(h) = a_i + b_i e^{-\frac{h}{c_i}}, \qquad i = 1, 2, 3, 4 \tag{1.7}$$

and

$$T(h) = a_5 - b_5 \frac{h}{c_5}, \tag{1.8}$$

where the parameters a_j, b_j and c_j, $j \in \hat{5}$, are inferred from balloon and satellite measurements at several sites across the globe, where the main EAS experiments are installed. Relation (1.8) represents the last layer of the atmosphere, where the mass overburden vanishes at $h = 112.8$ km. The mass overburden T is related to the slant atmospheric depth X, given by (1.2), by the following relation:

$$X = \frac{T}{\cos\theta}, \tag{1.9}$$

where we neglect the Earth's curvature (a reasonable assumption for $\theta < 70°$) and thus can assume that X rises with $\sec\theta$.

Chapter 2

Reconstruction of the Production Depth of Muons in EAS

The depth of the shower maximum, retrieved from the measurements of the electromagnetic longitudinal profiles of EAS, is one of the most-utilized mass-composition-sensitive observables in cosmic-ray physics. Similarly, we can acquire relevant information about cosmic-ray nuclear mass composition from the maxima of the longitudinal profiles of the production depths of muons.

Although the electromagnetic and muonic profiles originate from two distinct shower components and reflect different physics processes, their development is intrinsically connected to the hadronic cascade, which might explain why their shape is similar, as shown in Fig. 2.1. While the electromagnetic longitudinal profiles are a measurement of the almost calorimetric energy deposited by the shower electrons and positrons in the atmosphere, the muonic longitudinal profiles reflect the production rate of muons in the atmosphere (i.e., the number of muons produced at each slant depth X). Also, they are measured differently. While the longitudinal profiles of the electromagnetic component can be directly detected by fluorescence telescopes, the muonic longitudinal profiles can only be reconstructed from the arrival time of muons measured by ground and underground detectors.

The motivation for this thesis is two-fold. Namely, we want to investigate the possibility of inferring the mass composition of cosmic rays from the muonic longitudinal profiles and, additionally, of gaining deeper insight into the hadronic interactions taking place in EAS. The grounds for this work can be found in [6, 7, 33]. All of the previous results are based on the following observations:

1. **Muons deviate little from their mother particles**: As shown in [33], 10 GeV (50 GeV) muons deviate only by 0.1° (0.03°) from their parent pions. Similar values are found for kaons and other hadrons.

2. **Muons propagate in nearly straight lines to the ground:** Muon trajectories are very little affected by bremsstrahlung or multiple scattering. Additionally, the Earth's magnetic field only plays a role for very inclined air

showers (θ>80°), during which muons traverse significantly larger distances and where the Earth's curvature cannot be neglected.

These observations imply that muons carry relevant information about their production points and, consequently, their parent hadrons. Thus, muons (and their longitudinal production profiles) can, to some extent, reveal the mechanisms of high-energy hadronic interactions occurring in the higher parts of the atmosphere, which would otherwise be unavailable to us.

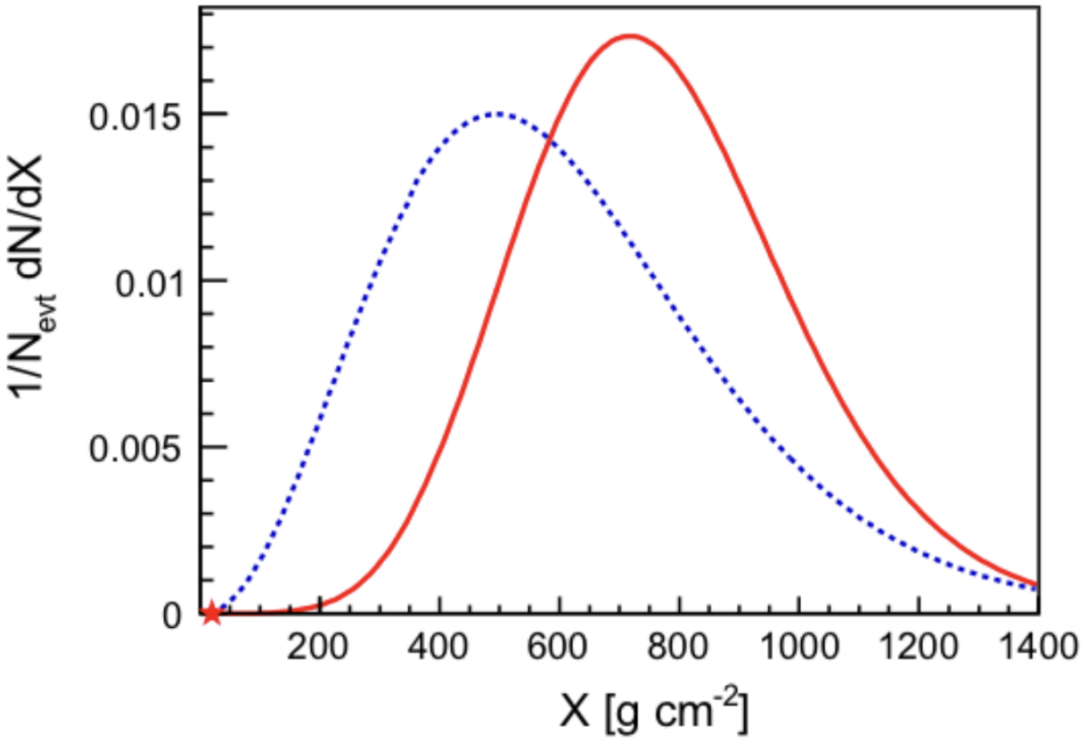

Figure 2.1: Illustration of the longitudinal profiles of the EM (red) and the muonic (blue) components, taken from [34].

Due to the fashion it is defined, the muonic longitudinal profile is conventionally referred to as the Muon Production Depth (MPD) distribution, which will be adopted from now on.

We distinguish two kinds of MPD distributions. The first one, called the *true* MPD distribution, is defined as the longitudinal profile of the production depth of all muons produced in an EAS above some energy E^{μ}_{kin}. This distribution is not affected by the effects of muon propagation and decay. Therefore, it faithfully reflects the parent hadronic component and, ideally, its shape would not be affected by the detector effects, as we will explain below. However, the *true* MPD distribution cannot be directly reconstructed without resorting to Monte Carlo simulations to estimate the fraction and distribution of the muons that decay before reaching the ground detectors. Thus, we define a second MPD distribution, called the *apparent* MPD distribution, which consists of only those muons that reach the ground level, and

that can be effectively detected. However, the *apparent* MPD distribution depends on the shower geometry, muon propagation effects, and observation conditions. In the following chapters, we will solely focus on the *apparent* MPD distribution, which we will call the MPD distribution from now on.

We note that the the Muon Production Depth is fundamentally a feature of individual muons, i.e., its reconstruction is performed on a muon-by-muon basis. We introduce this reconstruction process in the next section.

2.1 The Arrival Time Model

The current method of the MPD reconstruction is thoroughly described in [6, 7, 33]. Here, we summarize its main ideas. We first introduce the following notation regarding the geometry of an air shower (as illustrated in Fig. 2.2):

- Shower axis: An extrapolation of the incident cosmic ray's trajectory into the atmosphere.
- Shower core: The point where the shower axis intersects the ground.
- Shower plane: A hypothetical plane, perpendicular to the shower axis, propagating along the shower axis at the speed of light. It is where the shower particles are contained, which is why we also call it the (plane) shower front. The air-shower coordinate system used in our work is defined on the shower plane.

The model uses a Cartesian coordinate system, with the origin centered at the shower core. The x-and y-axes define the shower plane, with the y-axis parallel to the ground, and the z-axis coincides with the shower axis. Alternatively, a corresponding cylindrical coordinate system may be utilized, defined by:

$$r = \sqrt{x^2 + y^2}, \tag{2.1}$$

and

$$\xi = \arctan\left(\frac{x}{y}\right), \tag{2.2}$$

where r represents the *distance from the shower core* and ξ the polar angle in the shower plane.

When a muon is produced in the atmosphere, the corresponding z-coordinate indicates its production distance (called the Muon Production Distance) measured along the shower axis. Since muons are massive particles and do not propagate parallel to the shower axis, they will arrive delayed with respect to the arrival time of the shower front at the detector coordinates (r, ξ). We call this delay the muon arrival time. This is the only quantity measurable by the arrays of particle detectors and is the main idea behind the MPD reconstruction. It can be decomposed into several components, caused by various effects. Below, we will detail on the several caused of muon delay.

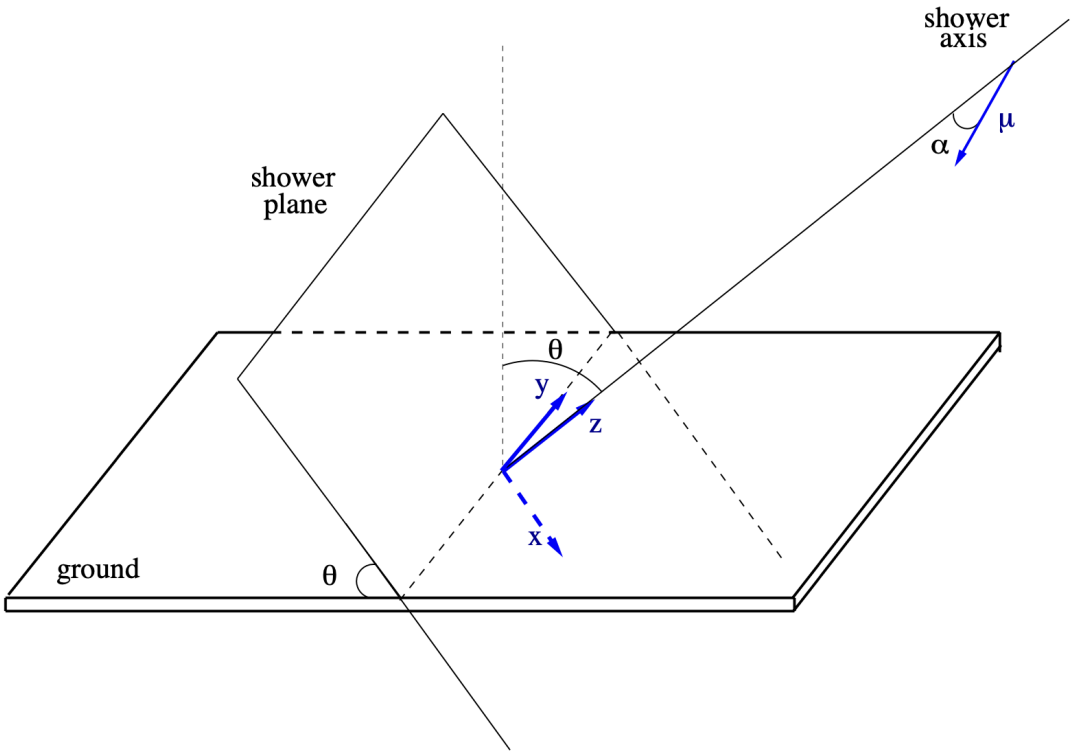

Figure 2.2: The geometry used in the standard MPD analysis, taken from [35].

Additionally, the basis of the reconstruction of the MPD depends on two assumptions: First, muons are created in the shower axis. This assumption was adopted in [35], arguing that the transverse production position of muons is confined to a few tens of meters from the shower axis, a small distance compared to the span of current EAS experiments. As a second simplification, we assume that muons travel along straight lines, as mentioned above.

2.1.1 Geometric Delay

Muons inherit the transverse momentum p_t of their parent hadrons, deviating from the shower axis by an angle $\alpha = \arcsin\left(\frac{p_t}{p_{tot}}\right)$, where p_{tot} is the muon total momentum. Consequently, muons will arrive at the ground with some delay with respect to the arrival time of the shower plane. Assuming, initially, that muons travel at the speed of light, the origin of this delay is only related to their trajectories. Therefore, we call it the *geometric delay*, and define it as

$$\tau_g = \frac{1}{c}[l^2 - (z - \Delta)^2], \tag{2.3}$$

where $\Delta = r \tan\theta \cos\xi$ is the distance from the muon impact point at the ground to the shower plane and $l = \sqrt{(z-\Delta)^2 + r^2}$ is the total distance traversed by a muon.

The relation can be inverted to get

$$z = \frac{1}{2}\Big(\frac{r^2}{c\tau_g} - c\tau_g\Big) + \Delta, \tag{2.4}$$

which represents a one-to-one mapping between the Muon Production Distance z and the geometric delay. In a final step, we compute the Muon Production Depth X from Eq. (1.2), namely:

$$X = \int_z^\infty \rho(z')\mathrm{d}z'. \tag{2.5}$$

2.1.2 Kinematic Delay

Since muons are massive particles, they cannot propagate at the speed of light. Thus, muons accumulate a further source of delay, called the *kinematic delay*, which is directly linked to their kinetic energy. The kinematic delay of a muon traveling over a distance l is given by

$$\tau_{kin} = \frac{1}{c}\int_0^l dl'\Big(\frac{1}{\beta(l')} - 1\Big), \tag{2.6}$$

where $\beta = v/c$, and v is the velocity of the muon. Since muons lose energy during their propagation, β will vary along the muon's path. As mentioned in Chapter 2, in the range of energies of interest, muons behave as minimum ionizing particles, i.e., suffer constant energy losses in the medium by ionization. The energy of muons at the detector E_f can be written as a function of their energy at production E_i, as $E_f = E_i - \rho a l$, where ρ is the traversed density, and $a \simeq 1.8\,\mathrm{MeV/g\,cm^{-2}}$ is the electronic stopping power of muons. Evaluating the integral, the kinematic delay then reads as:

$$\tau_{kin} = \frac{1}{c\rho a}\left(\sqrt{E_i^2 - m_\mu^2 c^4} - \sqrt{E_f^2 - m_\mu^2 c^4}\right) - \frac{l}{c}. \tag{2.7}$$

2.1.3 Further Sources of Delay

Apart from the geometric and the kinematic delays, which are the dominant sources of delay, several other effects contribute to the total arrival time of muons. Here, we mention two other types of delay. The first one, called the *geomagnetic delay* τ_B, arises from the deflection of muons in the Earth's magnetic field. In [36], a model for the geomagnetic treatment of muons has been described, while in [35], it was applied to the delay analysis. There, it was shown that this delay only becomes important for very long muon trajectories, corresponding to very inclined, almost horizontal ($\theta \geq 80^\circ$) air showers.

The second delay results from the multiple scattering of muons on nuclei. In [35], it was shown that for $r > 100$ m, this contribution, denoted as τ_{MS}, is deemed negligible, with a comment that if muons are treated separately, this effect might occasionally become important.

Additional sources of delays might be introduced by the approximations made by the Arrival Time Model, but their contributions are negligible with respect to the mentioned ones. To illustrate the contributions by the individual delays, we show their dependence on the distance from the shower core in Fig. 2.3.

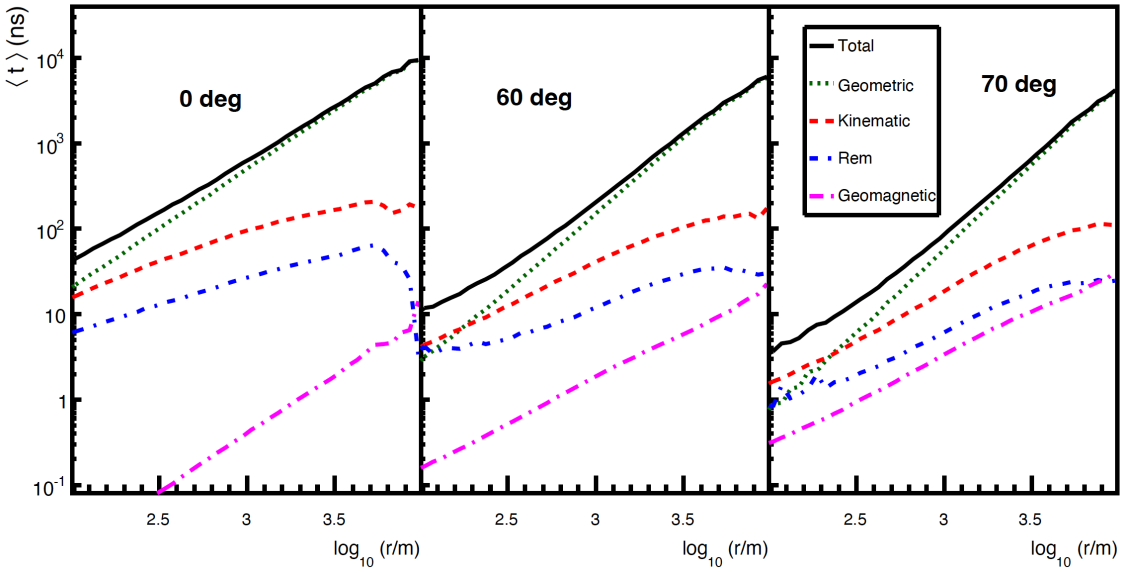

Figure 2.3: The contributions of different types of delay with respect to the distance from the shower core, depicted for 3 showers with different zenith angles. The "Rem" label corresponds to τ_{MS} and the remaining sources of delay combined. Taken from [35].

2.1.4 The Process of MPD Reconstruction

The total delay of muons can be expressed as

$$\tau = \tau_{geom} + \tau_{kin} + \tau_B + \tau_{MS} + ... \approx \tau_{geom} + \tau_{kin}, \tag{2.8}$$

which, within the scope of the model, can be reasonably approximated as a sum of the geometric and the kinematic delays. We can substitute (2.8) into relation (2.4) to get

$$z = \frac{1}{2}\Big(\frac{r^2}{c\tau - c\tau_{kin}} - (c\tau - c\tau_{kin})\Big) + \Delta. \tag{2.9}$$

Thus, to determine the Muon Production Distance/Depth, a precise estimation of the kinematic delay is crucial. This has proven to be a difficult task, since the kinematic delay is intimately related to the muon energy spectrum in EAS, which is largely unknown, given the difficulties and cost of such enterprise. Therefore, in this and the previous works, the kinematic delay estimation was obtained from parametrizations of the energy spectrum of muons obtained from Monte Carlo simulations of EAS.

In the standard MPD analysis ([6]), the mean kinematic delay is given by the following parametrization:

$$< \tau_{kin} >= \frac{1}{2c}\frac{r^2}{l}\varepsilon(r, z-\Delta), \tag{2.10}$$

where the function ε is given by

$$\varepsilon(r,z) = p_0(z)\Big(\frac{r}{[m]}\Big)^{p_1}, \tag{2.11}$$

with p_0 being parametrized as

$$\begin{aligned}\log_{10}(p_0(z)) = &-0.6085 + 1.955\log_{10}(z/[m]) - 0.3299\log^2_{10}(z/[m]) + \\ &+ 0.0186\log^3_{10}(z/[m])\end{aligned} \tag{2.12}$$

and p_1 as

$$\log_{10}(p_1) = -1.176. \tag{2.13}$$

The expression (2.10) is then substituted into (2.9), replacing the kinematic delay with the average kinematic delay. Due to the shape of the parameter p_0, the equation for the Muon Production Distance cannot be solved analytically. Therefore, the reconstruction of the Muon Production Distance z consists of an iterative process: First, we get an estimation of z from equation (2.9), neglecting the contribution of the kinematic delay, i.e., $\tau_{kin} = 0$. We then use this value of z in equation (2.10) to estimate the average kinematic delay, which we again plug into (2.9). This can be done several times, but the convergence of the method is quick such that only two iterations are necessary. After the final value of z is calculated, we transform it into the correspondent atmospheric depth X using the relation (1.2).

2.1.5 Applications and Limitations

The method is applicable for the MPD reconstruction only if a detector or a subsequent analysis is able to distinguish muon signals from contributions of other EAS particles. Additionally, while the method could, in principle, be applicable for air showers with a wide range of zenith angles, it was only implemented for air showers with $\theta \in (55°, 65°)$, utilizing the water-Cherenkov detectors of the Pierre Auger Observatory [37]. It was the only way to acquire an almost pure muonic signal in the detectors, as, for lower zenith angles, the contribution of the electromagnetic to the total signal measured by the WCDs was too high to allow a reliable reconstruction. Furthermore, it was shown in [33] that the relation between the uncertainties of the Muon Production Distance δz and the intrinsic time resolution of a given detector δt depends on the distance from the shower core r, namely:

$$\frac{\delta z}{z} = 2c\frac{(z-\Delta)^2}{zr^2}\delta t. \tag{2.14}$$

For distances close to the shower core, the uncertainty of the production distance diverges, therefore, a cut in r has to be applied. For a given precision e_z of $\frac{\delta z}{z}$, the

value of the cut r_c is given by:

$$r_c = \frac{\sqrt{\frac{2z_{max}c\delta t}{e_z}}}{1 + \sqrt{\frac{2c\delta t}{e_z z_{max}}} \tan\theta \cos\xi}, \tag{2.15}$$

where z_{max} is the highest possible production point of muons in a particular shower. For example, in [37], where the MPD was reconstructed using data from the Surface Detector of the Pierre Auger Observatory, a cut of $r_c = 1700$ m was imposed (given by the features of WCDs). The value of this cut decreases with decreasing zenith angle, as for lower θ, muons are created deeper in the atmosphere and thus z_{max} is smaller. The cut also has an impact on the contribution of the kinematic delay, as far from the shower core, it can be considered as a correction to the geometric delay (see Fig. 2.3). This is a helpful feature of the imposed cut, because it justifies the reasoning of assuming zero kinematic delay in the first iteration during the MPD reconstruction process. On the other hand, the number of muons in the reconstruction is drastically reduced by this cut. In [37], about 50 muons were left for EAS with cosmic ray energies of $10^{19.5}$ eV after applying all cuts. The r-cut is considered as one of the main setbacks of the current MPD reconstruction model and also one of the main motivation points for this thesis, as our proposed model will try to lower the cut to desirable values.

Chapter 3

Machine Learning

Machine Learning (ML) is a rapidly growing field in computer science that enables computers to learn from data without being explicitly programmed. This technology has been gaining popularity in recent years, as it has shown to have significant potential in a wide range of applications, including image and speech recognition, natural language processing, recommendation systems, and autonomous vehicles, among others. The increasing availability of large datasets, along with advances in hardware and software technologies, have contributed to the growth of ML, making it one of the most promising fields in (computer) science. With the ability to learn from vast amounts of data, ML can provide insights, predictions, and solutions that would otherwise be difficult or impossible to obtain using traditional programming techniques.

In recent years, machine learning has emerged as a powerful tool in cosmic-ray physics, allowing researchers to process large amounts of data and extract meaningful insights [5, 38]. Machine learning algorithms can help identify patterns and correlations in cosmic-ray data that may not be apparent using traditional statistical methods, while improving the efficiency and accuracy of the data analysis. The objective of this chapter is to provide an overview of machine learning and the algorithm at the heart of the proposed MPD reconstruction method, the Gradient-Boosted Decision Trees (GBDT).

3.1 Basic Concepts of Machine Learning

Machine learning is an application of the field of Artificial Intelligence that encompasses various, albeit all data-driven, algorithms. Based on their methodology of learning from data, ML algorithms can be broadly classified into three categories: supervised, unsupervised and reinforcement learning. Supervised learning algorithms are trained on labeled data, i.e., independent variables are supplied to the algorithm alongside the corresponding dependent variables we wish to model. Typical supervised learning applications are classification or regression problems. As the name suggests, unsupervised learning algorithms are, on the other hand, applicable for unlabeled data, being often used for clustering methods. The concept of reinforce-

ment learning is different, as the algorithm tries to learn by interacting with a chosen environment and receiving feedback in a form of rewards or punishments. In this thesis, we aim to reconstruct a quantity from different observables acquired by Monte Carlo simulations and study the final model's quality on various datasets. Therefore, we will focus on supervised learning and its characteristics.

3.1.1 Supervised Learning

The main goal of supervised learning is finding a function $f : X \longrightarrow Y$ that maps given input variables (called ***features***) $x \in X$ to the correct output (called ***label*** or ***target***) $y \in Y$. In other words, a supervised learning algorithm tries to solve the equation

$$y = f(x). \tag{3.1}$$

The essence of supervised learning lies in the method of finding a solution to (3.1): An algorithm tries to find a model $\hat{f}(\mathrm{x})$ that best fits the target function $f(\mathrm{x})$ by minimizing the expected value of a so-called objective function O:

$$\hat{f} = \underset{F}{\operatorname{argmin}}\, \mathbb{E}[O(f, F)]. \tag{3.2}$$

The optimization of O is done iteratively in a number of steps, hence the "learning" label. The form of O varies with respect to the problem at hand: For example, depending on the type of the target variable, supervised learning can be distinguished into two categories: classification- and regression-type problems. While classification involves predicting a discrete target, regression involves predicting continuous target values. Objective functions are different for each of these categories, and since this work focuses on a 1D regression-type problem, we will prefer explaining the characteristics of machine learning on regression examples from now on. We will also set $\boldsymbol{x} \in \mathbb{R}^m$ and $y \in \mathbb{R}$, where m is the number of features.

In order to acquire $\hat{f}$, a general structure of implementing a supervised learning model is typically followed:

1. **Problem definition:**
   The first step is to define a problem to be solved. This means selecting the desired target and assessing the state of the available features, be it the relations to the target variable or their type, such as text, images, or numerical data.

2. **ML algorithm selection:**
   Depending on the aim of the problem, there are many different types of supervised learning algorithms to choose from, such as GBDT or Neural Networks [39]. Various machine learning algorithms have their own advantages and disadvantages and are suitable for different data structures. For example, Neural Networks are suitable for problems in computer vision, language processing or speech recognition, while algorithms based on Decision Trees have upper hand in dealing with tabular data [40].

3. **Data pre-processing:**
   In order for the model to learn from the features, they must be pre-processed to ensure they are in a format that can be utilized by the algorithm. Different algorithms require different data pre-processing, but, generally, the task involves:

   - *Data cleaning*: Identifying and handling missing or erroneous data values, such as outliers, noise, or incorrect data types. This is done by either removing them or replacing them with appropriate values.
   - *Feature & Target Engineering*: Selecting and transforming features in a way that captures the underlying patterns and relationships in the data. The model's performance can be significantly impacted by the quality of the features, as well as the form of the target variable. It also includes encoding of categorical variables (converting them to numerical forms) and formatting data types such as dates, times, and currencies to ensure the consistency of the data.
   - *Data transformation*: Transforming data into a format that is appropriate for analysis. This includes scaling, normalization, and optionally converting continuous data into discrete data by grouping values into categories or bins. This is sometimes done to simplify the analysis of complex datasets.
   - *Data reduction*: Reducing the size of the dataset while retaining its important features. This is done through feature selection, or data under- or over-sampling.

4. **Data splitting:**
   A ML model first has to learn patterns in the given data in order to be able to make predictions on another dataset. Thus, the available data need to be split into two parts - the ***training dataset*** and the ***testing dataset***. The training dataset is used for the model to learn the corresponding underlying data patterns, while the testing set is used to evaluate the model's performance after the training has concluded.

5. **Training and evaluating the model:**
   Training a model involves a feedback loop where the model makes predictions on the input data, and the output is compared to the corresponding label. The model then adjusts itself based on the error between the two in order to minimize it. This feedback loop helps the model improve its performance over time. After the model has been trained, its performance is evaluated by using the testing dataset. This involves measuring various metrics between the predicted output and target, which generally depend on a given problem type.

6. **Tuning the model:**
   If the model's performance is not satisfactory, the model's tunable parameters, called ***hyperparameters***, need to be adjusted, either by hand or optimization algorithms. During the optimization process, models are trained with different sets of hyperparameters, choosing the best performing one. After a satisfactory

optimization, the model is ready to be implemented in predictions on other (possibly unlabeled) datasets.

During the training phase, a model needs to quantify the error between the predicted output and the corresponding target. This is done via a so-called ***loss function***, which takes the predicted output and the target as inputs and outputs a single number (or a vector) as a measure of error. The most commonly used loss function in regression problems is the ***Mean Squared Error***

$$MSE(a, b) = \mathbb{E}[(a - b)^2] = \frac{1}{n}\sum_{i=1}^{n}(a_i - b_i)^2, \tag{3.3}$$

where n is the number of datapoints in a given dataset.

3.1.2 The Bias-Variance Trade-off

An important quality measure for a newly constructed ML model is its ability to learn a function that can make accurate predictions on new, unseen data. The generalization ability of a model is closely related to the concepts of ***underfitting*** and ***overfitting***, and the phenomenon of the so-called ***bias-variance trade-off***.

Bias and variance are two sources of error that can affect the performance of a machine learning model. An illustration of the concept is depicted in Fig. 3.1.

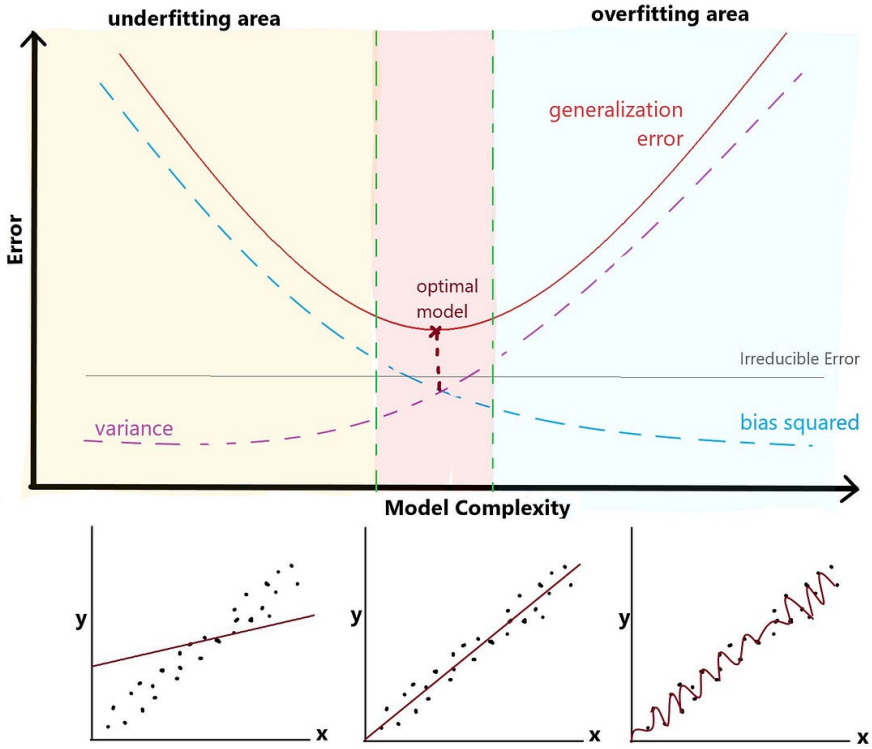

Figure 3.1: An illustration of the Bias-Variance Trade-off, taken from [41].

Bias refers to an error introduced by approximating the problem at hand with a simpler model. If $\hat{f}(\boldsymbol{x})$ is the approximate solution to (3.6), then

$$\text{bias}[\hat{f}(\boldsymbol{x})] = \mathbb{E}[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x}). \tag{3.4}$$

Models with high bias fail to capture the complexity of the data and thus lead to higher errors on both training and testing datasets. On the other hand, models with low bias correspond to a good fit to the training dataset.

Variance, on the other hand, describes an error introduced by the model being too sensitive to the training data, or, alternatively, it represents the amount by which the model output would change if different training data were used. Variance of a model is described as

$$\text{var}[\hat{f}(\boldsymbol{x})] = \mathbb{E}[\hat{f}^2(\boldsymbol{x})] - (\mathbb{E}[\hat{f}(\boldsymbol{x})])^2. \tag{3.5}$$

A model with high variance is overly complex and captures noise in the training data, which implies that it will not perform well on previously unseen (testing) data. Correspondingly, low variance indicates that the model has the ability to generalize to other datasets.

Additionally, in most real-world applications, due to various sources of noise accompanying the data, a random error term ε is present in (3.1), called the ***irreducible error***. Therefore, most of the time, a function to be learned has the following form:

$$y = f(\boldsymbol{x}) + \varepsilon(\boldsymbol{x}), \quad \mathbb{E}[\varepsilon] = 0, \text{var}(\varepsilon) = \sigma_\varepsilon^2. \tag{3.6}$$

Considering all three sources of errors and taking the MSE as our error measure, the equation for the bias-variance trade-off reads as:

$$MSE(\hat{f}(\boldsymbol{x}), f(\boldsymbol{x})) = (\text{bias}[\hat{f}(\boldsymbol{x})])^2 + \text{var}[\hat{f}(\boldsymbol{x})] + \sigma_\varepsilon^2. \tag{3.7}$$

From (3.7), we see that optimally, both bias and variance should be as low as possible. In most cases, however, it is not possible to achieve both, as one of them tends to increase as the other decreases (as depicted in Fig. 3.1). Instead, the model needs to be optimized so to prevent both underfitting (high bias, low variance - model fails to infer fundamental information from data) and overfitting (low bias, high variance — over-complicated model, does not generalize to unseen data). While underfitting can be caused by having too few features, using a model that is too simple or training the model for too short a time, overfitting is often caused by exactly the opposite, having too many features, using a model that is too complex, or training the model for too long.

To prevent underfitting and overfitting, a separate dataset, called the ***validation dataset***, is often used to evaluate the performance of a ML model during training. It acts as a separate testing dataset, but its error is evaluated before each adjustment of the model during the training, rather than at its end. This way, the change in the validation error can be monitored, letting the model be trained as the validation

error decreases, preventing underfitting. On the other hand, when the validation error stops decreasing, the training of the model can be stopped to prevent overfitting (which is the principle of the so-called early-stopping algorithm). Like this, an optimal model, given the selected hyperparameters, is found. The validation dataset can also be used for evaluating a model for a given set of hyperparameters when tuning the model, allowing us to select the right combination of hyperparameters in order to get an optimal model.

3.2 Gradient-Boosted Decision Trees

Gradient Boosted Decision Trees (GBDT) is a powerful machine learning algorithm used for regression and classification problems, particularly standing out when dealing with tabular datasets. GBDT has gained widespread popularity in recent years due to its ability to handle complex, high-dimensional data and non-linear relationships, while producing highly accurate predictions. The basic object in the GBDTs algorithm is a so-called decision tree, a hierarchical model that maps out possible decisions and their consequences using a flowchart-like structure. The process of gradient boosting then combines multiple decision trees (also labeled weak-learners), creating one strong-learner. Apart from being an interpretable and versatile algorithm, GBDT is readily available in popular machine learning libraries such as XGBoost [42], and LightGBM [43], which makes it easy for developers and data scientists to implement it in their projects. The next subsections are dedicated to describe GBDTs in more detail.

3.2.1 Decision & Regression Trees

A decision tree is a model used in decision-making processes, implemented via a tree-like structure. From the simplest illustration point, we can imagine it as a visual tool used to make decisions by mapping out all possible choices and their likely consequences (see Fig. 3.2). It effectively operates by partitioning a given parent dataset into small subsets of datapoints that share some key characteristics. A decision tree consists of the following components:

- **Node:**
  A "point" where a decision is made on how to split a given dataset into subsets, based on the values of one of the input features. Each node contains a set to be split, a splitting criterion based on one of the input features and two or more branches, each corresponding to a possible outcome of the split. Depending on a node's position in a tree, we can label it as:
  - *Root Node*: The topmost node in a decision tree, where a first decision is made. It is often labeled as the level 0 node, where levels, counted from the root node downwards, represent the depth of a tree
  - *Decision/Internal Node*: A node belonging to level $1, 2, ..., n-1$, where n is the overall depth of a tree

– *Leaf (Node)*: A final node in a tree (level n node), at which a label or a value is assigned to datapoints satisfying all criteria that lead to this node.

- **Branch:**
  A "line" connecting two nodes, corresponding to a possible outcome of the split made in the node above it. Branches of a decision tree can be interpreted as a set of if-then rules which explain how the input features are related to the output variable, with each path through the tree corresponding to a different set of conditions that lead to a specific outcome. The number of branches emanating from a decision node depends on the chosen split conditions and the number of possible values the corresponding feature can take. In all of the following text, we will assume only two branches growing from each node.

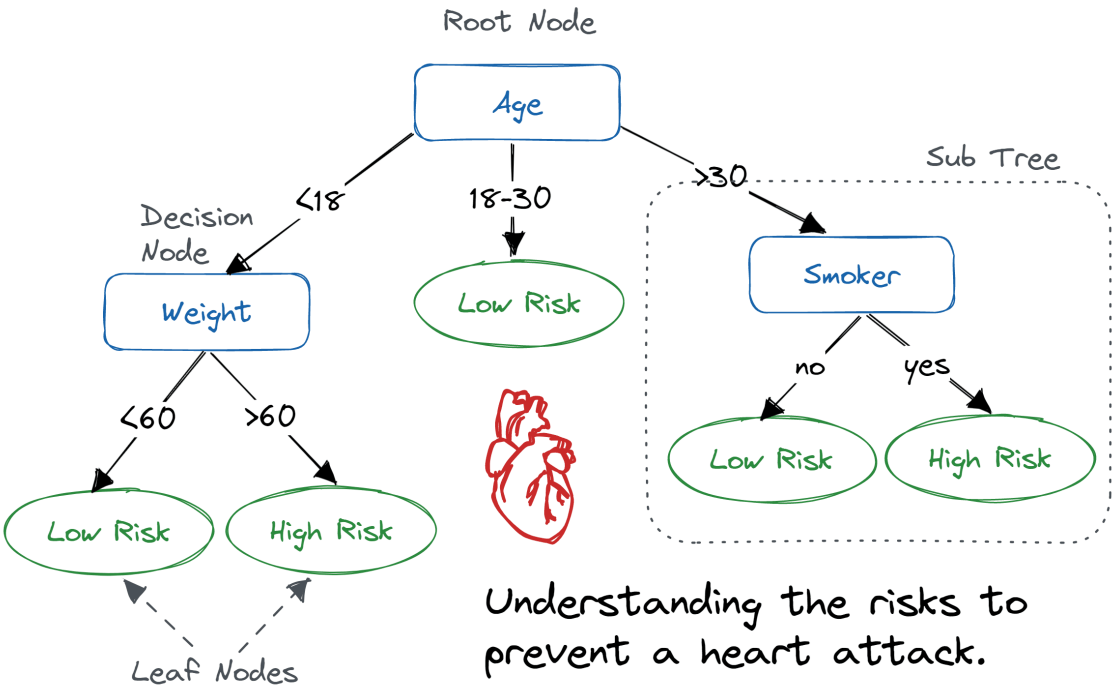

Figure 3.2: An illustration of the structure of a decision tree, taken from [44].

In mathematical terms, we first assume a dataset $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^{n}$ of independent identically distributed tuples $(\boldsymbol{x}_k, y_k)$, where $\boldsymbol{x}_i \in \mathbb{R}^m$, $y_i \in \mathbb{R}$, m is the number of features and n the overall number of datapoints. A decision tree can be expressed as a function $\hat{f}$ of input features $\boldsymbol{x}$ as

$$\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{J} v_j \mathbb{1}_{R_j}(\boldsymbol{x}), \tag{3.8}$$

where J is the total number of leaves in the tree, R_j is the j-th leaf (i.e., a set bounded by all splitting criteria through which the leaf arose), v_j is the value assigned to all datapoints belonging to R_j (depending on the target) and $\mathbb{1}$ is the indicator function defined as

$$\mathbb{1}_{R_j}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in R_j \\ 0 & \text{if } \boldsymbol{x} \notin R_j. \end{cases} \tag{3.9}$$

Assuming a subset $S \subseteq \mathcal{D}$ present at a given node, a split s is defined as a pair of subsets $\{S_1, S_2\}$, $S_1 \cup S_2 = S$, which can be written as

$$s = \Big\{\{(\boldsymbol{x}, y)|x^j < t, (\boldsymbol{x}, y) \in S\}, \{(\boldsymbol{x}, y)|x^j \geq t, (\boldsymbol{x}, y) \in S\}\Big\}, \tag{3.10}$$

where x^j, $j \in \{1, ..., m\}$, is the feature according to which the split is made, and t is the threshold which defines the split. We define the pair (x^j, t) as the "criterion". In the process of building a decision tree, splits are made on newer and newer nodes until some terminal condition is met, e.g., when the set in each leaf contains only one class of data or when a tree reaches a pre-determined depth level.

The output value v_j, corresponding to the j-th leaf, depends on whether we deal with a classification or a regression, hence regression trees, problem:

- *Classification*: The target variable y is discrete, i.e., R_j contains datapoints with targets corresponding to classes of data (e.g., 1, 2, ...). When there is only one class present in R_j, the decision tree predicts this class as the output v_j for each datapoint satisfying the split conditions leading to R_j. If there are multiple classes present in R_j, the majority class is taken as the output.
- *Regression*: The target variable y is continuous. Therefore, for a given leaf R_j, v_j is chosen as the mean value of all target values in R_j.

The main goal of a decision tree is to find the optimal splitting decisions throughout the tree that maximize the overall information gain. If, at a given node, we choose some specific criterion A, then the information gain, IG, is defined as

$$IG(S, A) = I(S) - \sum_i \frac{|S_i|}{|S|} I(S_i), \tag{3.11}$$

where $|S|$ and $|S_i|$ denote the number of datapoints in S and S_i, respectively, and where I is the so-called impurity metric, whose form depends on the type of problem to be solved. For regression analysis, we usually take

$$I(S) = MSE(S) = \frac{1}{|S|} \sum_j (y_j - \bar{y})^2, \tag{3.12}$$

where y_j are the target values in S and $\bar{y}$ is their mean value (similarly for the subsets S_i).

The basic algorithm for finding the best splitting decision at each node, corresponding to the maximal information gain, proceeds as follows:

1. Separately, for each feature x^j, a given dataset is sorted such that the values of the respective feature are sorted in ascending order. Therefore, we end up with m separate datasets.

2. In each newly created dataset, we sequentially select the threshold t to be the average value between each pair of adjacent values of the sorted feature, i.e., $t = \frac{x_k^j + x_l^j}{2}$, with x_k^j and x_l^j adjacent. Combined with the condition above, we end up with a total of $m \cdot (n-1)$ thresholds.

3. For each t, we calculate the information gain from (3.11) and select the criterion with the maximum value of IG to do the splitting at the given node.

For large datasets, this "greedy" algorithm does not scale well, which is why, in advanced GBDTs libraries, the so-called histogram splitting algorithm is implemented. The algorithm first bins a feature according to which a split is to be made, and then proceeds according to the basic algorithm. By considerably reducing the number of potential splitting points to be investigated, the algorithm speeds up the splitting process at the price of losing some of its flexibility (a risk which is largely mitigated when dealing with large datasets).

A regression tree is built node-by-node, until user-specified criteria are met. Various GBDT libraries vary in the *order* by which the nodes are constructed, the two most commonly used strategies being depth-wise and leaf-wise tree growths. While depth-wise expansion splits all nodes at a given depth before adding more levels, the leaf-wise approach always splits the node that maximizes the information gain, which can lead to unbalanced trees (see Fig. 3.3). If there is not a specific criterion to stop the construction of a tree (i.e., complete trees are built), the two approaches will produce the same result. Generally, depth-wise growth performs better for smaller datasets, where leaf-wise-built trees tend to overfit. In comparison, leaf-wise growth tends to excel when dealing with larger datasets, outperforming the level-wise growth [45].

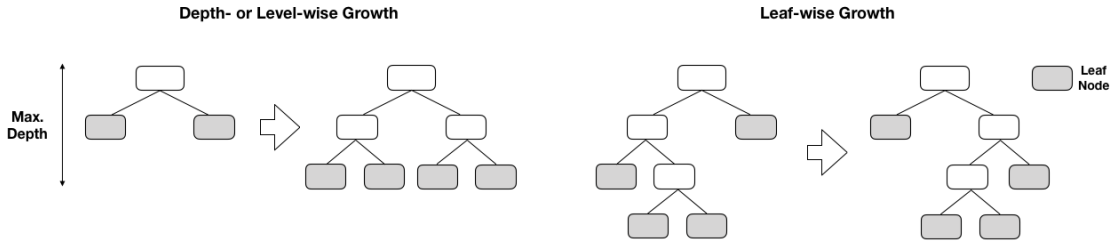

Figure 3.3: Illustrations of Growing Strategies used in creating decision trees, taken from [45].

As for the algorithm's advantages, a single regression tree is easily interpretable. Unlike neural networks, often seen as a "black box" model, we can consider single regression trees as a "white box" model, requiring less intense data pre-processing than other machine learning models. On the other hand, it has significant disadvantages: It is prone to extreme overfitting, as, without any stopping mechanism, a tree could just create a leaf for each of its training datapoints, which is why decision trees are generally high-variance models. Additionally, decision trees are not robust models, since small variations in data may give rise to completely different trees.

Both these disadvantages can be mitigated by combining multiple regression trees into one model, one such method being the gradient boosting.

3.2.2 Gradient Boosting

Gradient boosting is one example of the so-called ensemble learning, a type of supervised learning method based on combining many weak models (or weak learners) into a single powerful one. A weak learner is, in classification settings, defined as any model that performs better than a simple coin toss (i.e., a random guess). The gradient boosting method combines weak learners sequentially, as each subsequent learner attempts to learn relevant information from the previous one. In the GBDT algorithm, a weak learner is a decision/regression tree that has not been fully grown. Each tree tries to improve upon the errors made by the previous trees, being the errors quantified by a loss function L. The final prediction is made by combining the predictions of all the weak trees. Consequently, the GBDT, as opposed to a single tree, is a significantly more robust algorithm. It is less prone to overfitting and can model more complex relationships between features and a target with significantly better accuracy.

As it was written earlier in this chapter, a supervised learning algorithm learns by optimizing a problem-specific objective function O, trying to find a "good-enough" model $\hat{f}(\boldsymbol{x})$ of the target function $f(\boldsymbol{x})$. In the GBDT algorithm, O is in the form

$$O = L + \Omega, \tag{3.13}$$

where Ω stands for the so-called regularization terms, which control the complexity of individual trees. We first look at a simplified case where $\Omega = 0$. Since the GBDT algorithm works by sequentially adding decision trees to an ensemble model, we can write $\hat{f}$ as

$$\hat{f}^{(K)} = \sum_{k=1}^{K} \gamma_k, \quad \gamma_k \in \mathcal{T}, \tag{3.14}$$

where K is the total number of trees used in the ensemble (not known beforehand) and $\mathcal{T}$ is the set of all possible decision/regression trees. γ_k is an output of a regression tree, defined by (3.8). The algorithm is first initialized by a constant value, typically the mean value of the target, and then, the training itself proceeds in an iterative manner: After each tree is added at a step t, the overall model is updated via the relation

$$\hat{f}^{(t)} = \sum_{k=1}^{t} \gamma_k = \hat{f}^{(t-1)} + \gamma_t. \tag{3.15}$$

The aim of each new tree γ_t is to optimize (i.e., to minimize) the given objective function. Therefore, at each step t, we want to add a tree which satisfies

$$\gamma_t = \underset{\gamma}{\operatorname{argmin}}\, O(f, \hat{f}^{(t)}) = \underset{\gamma}{\operatorname{argmin}}\, L(f, \hat{f}^{(t-1)} + \gamma). \tag{3.16}$$

Depending on the specific form of the chosen loss function, (3.16) is relatively difficult to solve. Thus, as a common practice (see [42]), a second order Taylor expansion of

L is made so that

$$\gamma_t \approx \underset{\gamma}{\operatorname{argmin}} \left[L(f, \hat{f}^{(t-1)}) + \sum_{i=1}^{n} \Big(g_i \gamma(\boldsymbol{x}_i) + \frac{1}{2} h_i \gamma^2(\boldsymbol{x}_i) \Big) \right], \tag{3.17}$$

where

$$\begin{aligned} g_i &= \frac{\partial L(f(\boldsymbol{x}_i), \hat{f}^{(t-1)}(\boldsymbol{x}_i))}{\partial [\hat{f}^{(t-1)}(\boldsymbol{x}_i)]}, \\ h_i &= \frac{\partial^2 L(f(\boldsymbol{x}_i), \hat{f}^{(t-1)}(\boldsymbol{x}_i))}{\partial [\hat{f}^{(t-1)}(\boldsymbol{x}_i)]^2} \end{aligned} \tag{3.18}$$

are the gradient and the diagonal elements of the Hess matrix of the loss function, respectively. If we substitute the tree structure (3.8) into (3.17) and drop the $L(f, \hat{f}^{(t-1)})$ term, as it does not contribute to the minimization process, for each leaf R_j of the t-th tree we get

$$v_j \approx \underset{v}{\operatorname{argmin}} \left[(\sum_{i \in R_j} g_i) v + \frac{1}{2} (\sum_{i \in R_j} h_i) v^2 \right] = - \frac{\sum_{i \in R_j} g_i}{\sum_{i \in R_j} h_i} \tag{3.19}$$

and finally

$$\gamma_t \approx - \sum_{j=1}^{J} \frac{\sum_{i \in R_j} g_i}{\sum_{i \in R_j} h_i} \mathbb{1}_{R_j}(\boldsymbol{x}). \tag{3.20}$$

The relation (3.20) provides a general method of determining which tree to add at each step t. If we take $L = MSE$, then

$$\begin{aligned} g_i &= 2\big[\hat{f}^{(t-1)}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\big], \\ h_i &= 2, \end{aligned} \tag{3.21}$$

and

$$\gamma_t \approx \sum_{j=1}^{J} \frac{\sum_{i \in R_j} r_i}{|R_j|} \mathbb{1}_{R_j}(\boldsymbol{x}), \quad r_i = f(\boldsymbol{x}_i) - \hat{f}^{(t-1)}(\boldsymbol{x}_i). \tag{3.22}$$

The value of each leaf is the mean value of the negative gradient $\boldsymbol{r}$, also called the residual. Therefore, the method is clear: We train a regression tree with the original features $\boldsymbol{x}$, but where the target variable is the negative gradient $\boldsymbol{r}$, and we take this tree to be γ_t. This means that each new tree tries to learn on the residuals from the old ensemble and adds its newly predicted residuals to the overall model. Together, the trees form a strong ensemble model, capable of predicting complex and high-dimensional relationships with state-of-the-art accuracy. For illustration purposes, an example of how a GBDT model fits a 1D function is depicted in Figure 3.4.

As another method of preventing overfitting, the GBDT algorithm can be implemented with regularization terms which punish overly complex trees so that the algorithm will tend to select a model which employs simpler trees. In modern GBDT libraries such as XGBOOST or LightGBM, the basic tree complexity term can be expressed as

$$\Omega(\gamma) = \sum_{j=1}^{J} \alpha |v_j| + \frac{\lambda}{2} v_j^2, \tag{3.23}$$

where the α and λ parameters are called the $L1$ and $L2$ regularization hyperparameters, respectively. If we substitute (3.23) into (3.13) and subsequently solve for γ_t, then for each leaf v_j, $j \in \{1, ..., J\}$, we get

$$v_j \approx -\frac{\sum_{i \in R_j} g_i \pm \alpha}{\sum_{i \in R_j} h_i + \lambda}, \tag{3.24}$$

where the plus sign applies if the nominator is larger than zero and the minus sign otherwise, bringing the optimal leaf value closer to 0. It can be seen that each regularization parameter lowers the value of the leaf in a different fashion, which is why both hyperparameters are used frequently in model tuning.

The last regularizing hyperparameter we mention is the *learning rate* $\varepsilon \in (0, 1\rangle$, which scales the output of all trees, resulting in a slower but ultimately more accurate training sequence. If this hyperparameter is implemented, we can generally write the whole GBDT model as

$$\hat{f}^{(K)} = \varepsilon \sum_{k=1}^{K} \gamma_k, \quad \gamma_k \in \mathcal{T}. \tag{3.25}$$

There is a large number of tunable hyperparameters available in large GBDT libraries, all of which controlling some part of the tree-building process. It is therefore important to precisely investigate which hyperparameters to tune, as it allows the model to better generalize to new data, which is the ultimate goal of machine learning.

3.2.3 The LightGBM Library

LightGBM is a high-performance gradient-boosting framework that has gained popularity in recent years. Developed by Microsoft, it is designed to be memory-efficient, fast, and scalable. One of the key features of LightGBM is its ability to handle large datasets efficiently. It uses a technique called Gradient-based One-Side Sampling (GOSS) to reduce the computational cost of gradient calculations by retaining the instances with large gradients and undersampling those with smaller gradients. This reduces the memory and computation requirements, enabling LightGBM to handle datasets that are too large for other algorithms. Another unique feature of the framework is the Exclusive Feature Bundling (EFB), which groups together highly correlated features to further reduce the memory usage and speed up the training process. Additionally, LightGBM supports parallel processing, allowing it to take advantage of multi-core CPUs and distributed computing environments. All of these features, with the addition of its native use of the histogram splitting algorithm, makes LightGBM often significantly faster than many other popular machine learning algorithms.

The algorithm has many hyperparameters to be set before the training process begins. They control the behavior of the algorithm and play a crucial role in determining the accuracy of the model. The hyperparameters can be divided into two categories: tree-based and boosting-based.

- **Tree-based:**
  - *Number of leaves:* Assigns the maximum number of leaves in a tree. A higher value will lead to a more complex model, which can result in overfitting.
  - *Maximal depth of tree:* Controls the maximum depth of a tree. Similarly as for the number of leaves, a higher value will result in a more complex model, which is prone to overfitting.
  - *Fraction of features used:* Decides the fraction of features to be randomly sampled for each tree, which may help to prevent overfitting by reducing the correlation between trees. It is also called *Colsample by Tree*.
  - *Fraction of data used:* Decides the fraction of data to be randomly sampled for each tree, which might help to reduce overfitting by creating more diverse trees. It is also named *Subsample*.
- **Boosting-based:**
  - *Learning rate:* Controls the step size at each iteration of the training process. A smaller learning rate will result in slower convergence but can result in better generalization. Corresponds to the hyperparameter ε in (3.25).
  - *Minimal Child Weight:* Is used to control the minimum sum of instances (hessian) required to continue splitting a node in a decision tree. It corresponds to the sum $\sum_{i \in R_j} h_i$ in (3.24). If this sum in a node is less than a predefined number, then a split at this node will not be made, and the node will become a leaf node. It can help to prevent overfitting by reducing the number of unnecessary splits in the tree, especially when dealing with noisy data.
  - *L1 (L2) Regularization:* Involves adding a penalty term to the objective function that is proportional to the absolute value (square) of the leaf values in the model. This penalty term encourages the model to use fewer features by shrinking the coefficients of less important features towards zero. Corresponds to the hyperparameter α (λ) in (3.24).

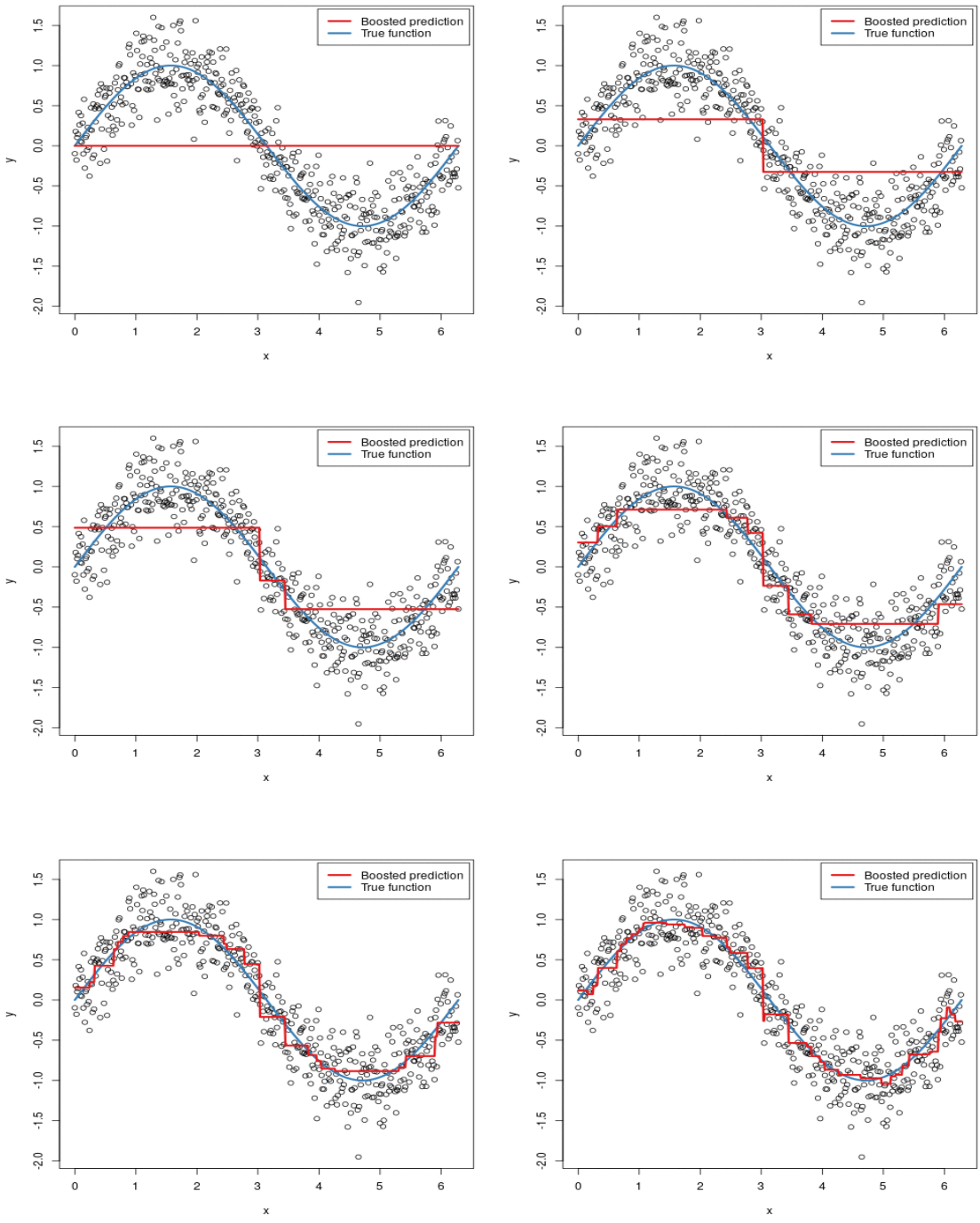

Figure 3.4: Illustrative process of a GBDT model fitting the function $\sin x$.

Chapter 4

MPD Reconstruction

In this chapter, we introduce a new approach for the reconstruction of the MPD, based on the aforementioned Gradient-Boosted Decision Trees (GBDT) algorithm, using the LightGBM library. We first describe the data used in the reconstruction process, the technical details of the machine learning (ML) model and its hyper-parameters, and pre-processing applied to the data to obtain improved reconstruction performance. In the second part of this chapter, we apply the model to EAS datasets with different characteristics. Initially, we analyze an EAS simulation library using fixed energy and zenith angle and later transition to a more realistic scenario, considering libraries with continuous zenith angle and energy. We test the accuracy of the GBDT model, first to the muon-by-muon reconstruction of the MPD and then to the reproduction of the overall MPD distribution. While the first gives us key information about the method's accuracy, the latter is the observable, which we can effectively measure with an array of particle detectors. After validating the method, we apply it to different primary particle species, hadronic interaction models, and higher cosmic ray energies and investigate the corresponding effects.

4.1 Simulations and Data Preparation

The analysis is conducted with Monte Carlo simulations of EAS, using the CORSIKA v.7.7402 software. The particle transport and interactions of the shower particles below 80 GeV, what we call the "low-energy hadronic interactions model", was handled by the FLUKA v. 2020.0.6 package [46], interfaced within CORSIKA. A summary of the simulation libraries used in this study is given in Table 4.1. We divide the MPD analysis into three phases, each testing the proposed ML model on increasingly real-situation datasets:

1. EAS with fixed values of zenith angles θ and primary particle energy E_{prim},
2. EAS with continuous values of θ, sampled from a uniform distribution in $\sin^2(\theta)$, and fixed primary particle energy E_{prim},
3. EAS with continuous values of θ and E_{prim}.

E_{prim} [eV]	θ [°]	Primary[1]	Model	Thinning[2]	# showers
Fixed: $= 10^{17.0}$	Fixed: $= \{0°, 60°\}$	Proton	QGSJET-II.04	Yes	200
			EPOS-LHC		200
			SIBYLL-2.3d	No	40
		Iron	QGSJET-II.04	Yes	200
			EPOS-LHC		200
			SIBYLL-2.3d	No	40
	Continuous: $\in (0°, 65°)$	Proton	QGSJET-II.04	Yes	1000
			EPOS-LHC		500
		Iron	QGSJET-II.04		1000
			EPOS-LHC		500
Continuous: $\in (10^{18.5}, 10^{19.0})$	Continuous: $\in (0°, 65°)$	Proton	QGSJET-II.04	Yes	500
			EPOS-LHC		500
			SIBYLL-2.3d		500
		Iron	QGSJET-II.04		500
			EPOS-LHC		500
			SIBYLL-2.3d		500

Table 4.1: A summary of all air showers implemented in the MPD analysis.

4.1.1 Domain Transformations & Data Cuts

The data structure implemented in the proposed ML model is based on the current MPD reconstruction method and, therefore, we follow the initial data pre-processing and muon selection according to the Arrival Time model, described in Chapter 2.

First, we are interested in reconstructing muons that can reach ground detector arrays, i.e., that do not decay in the atmosphere. Thus, for the EAS simulations in CORSIKA, we set the ground altitude of 1452 m, corresponding to the average altitude of the Surface Detector of the Pierre Auger Observatory, and instruct CORSIKA to output only those muons that reach the set altitude. In this work, we also assume an ideal detector response.

Second, the coordinate system used by CORSIKA refers to the plane of the detector, which we call the Detector Plane system - DP. However, we use the so-called Shower Plane system - SP in standard cosmic-ray analyses. Therefore, we must perform the necessary coordinate transformations from the DP to the SP on our data. This procedure is done in three stages, described as follows:

1. In CORSIKA, by definition, the shower core, i.e., the point where the shower axis intersects the ground, is set at the DP coordinates $(0, 0, z_{\text{ground}})$, where $z_{\text{ground}} = 1452\,\text{m}$.

[1]From now on, we will call the EAS-initiating cosmic ray the **primary particle** or simply the **primary**.

[2]Given the huge number of secondary particles generated in an EAS, typical Monte Carlo simulations apply the "thinning" algorithm, which significantly aids in reducing the required CPU time. See Chapter 1, section 1.5.

2. The transformation between the DP and the SP system (x_{SP}, y_{SP}, z_{SP}), where z_{SP} is the Muon Production Distance, is performed via the following relations (utilizing the corresponding DP coordinates Φ and θ - see Fig. 1.10):

$$\begin{aligned} \varphi &= \Phi + \pi \\ x_{SP} &= x_{DP}\cos\varphi\cos\theta + y_{DP}\sin\varphi\cos\theta \\ y_{SP} &= -x_{DP}\sin\varphi + y_{DP}\cos\varphi \\ z_{SP} &= \frac{z_{DP}}{\cos\theta} \end{aligned} \tag{4.1}$$

3. A final transformation into a corresponding cylindrical coordinate system is performed:

$$\begin{aligned} r_{SP} &= \sqrt{x_{SP}^2 + x_{SP}^2} \\ \xi_{SP} &= \arctan\left(\frac{x_{SP}}{y_{SP}}\right), \end{aligned} \tag{4.2}$$

where r_{SP} is usually called the *distance from the shower core*. Additionally, we construct two additional variables:

$$\begin{aligned} \Delta &= r_{SP}\tan\theta\cos\xi_{SP} \\ l &= \sqrt{r_{SP}^2 + (z_{SP} - \Delta)^2}, \end{aligned} \tag{4.3}$$

with Δ representing a distance between the muon ground impact point and its projection in the shower plane, and l is the total distance traversed by a muon in the atmosphere.

Since we only perform our analysis in the SP system, we will omit the SP index in the subsequent formalism. A simplified illustration (for $\xi = 0°$) of the SP coordinate system is displayed in Figure 4.1 for clarity.

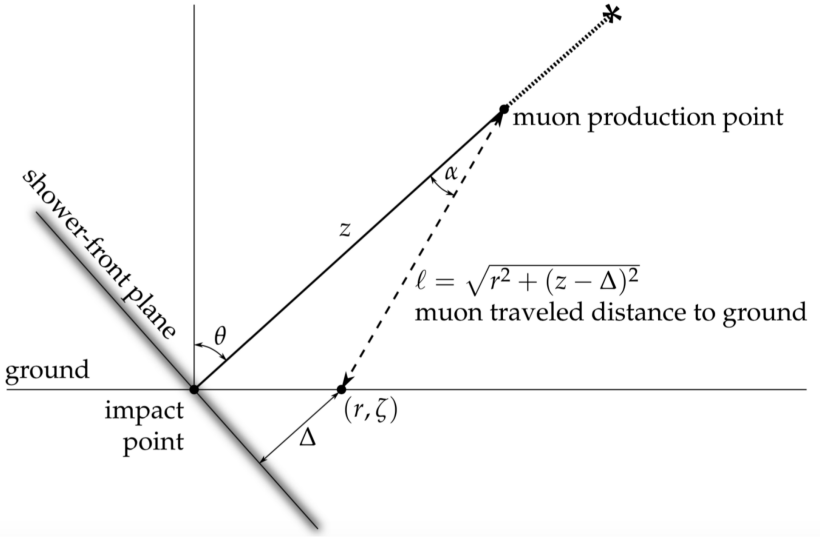

Figure 4.1: An illustration of the Shower Plane system, depicted for $\xi = 0°$.

The second modification to the raw CORSIKA data concerns the reference time system. On our simulations, the origin of our time reference system is set at the first interaction of the cosmic ray with an air molecule. Therefore, the timing information of the particles arriving at the ground level reflect the absolute time t_{abs} elapsed since the first interaction, at $z = z_{first}$. In the standard MPD analysis, however, we work in the Shower Plane coordinate system, in which the time origin t_{ref} is set by the impact of the shower core at the ground, with the shower particles arriving with a given delay with respect to the arrival time of the shower-plane front. The reference time t_{ref} is the time required for a hypothetical particle propagating at the speed of light, parallel to the shower axis, to travel from the first interaction point with $(t_{abs}, z) = (0, z_{first})$ to the ground. At the ground level, the reference time is given by

$$t_{ref} = \frac{z_{first}}{c} \tag{4.4}$$

and our respective time delay t for each muon reads

$$t = t_{abs} - (t_{ref} - \frac{\Delta}{c}), \tag{4.5}$$

where the $\frac{\Delta}{c}$ term arises from the asymmetrical propagation of an air shower with $\theta \neq 0°$ (for each muon, we shift the z-coordinate by Δ so the shower plane coincides with the ground impact point - see Fig. 4.1). As is common jargon in the MPD field, we will refer to the total delay of muons t as the *arrival time*.

Lastly, we apply specific cuts for the aims of our work:

1. In the previous work, the reconstruction of the MPD was applied to the surface detector array of the Pierre Auger Observatory [37]. However, it could only be used for detectors at radial distances larger than 1700 m, due to the reconstruction quality requirements given by the relation (2.14). This resulted in the necessity of selecting EAS with energies above 20 EeV, which would contain enough muons to reconstruct the MPD profiles. Our objective is to dramatically reduce the applied radial cut from $r_{cut} \geq 1700\,\mathrm{m}$ to $r_{cut} \geq 200\,\mathrm{m}$, allowing us to reconstruct EAS with lower energies.

2. In order to reconstruct the MPD, it is necessary to distinguish the muonic component of EAS from the EM background. In [37], EAS with zenith angles ranging from $55° - 65°$ were used to satisfy this requirement, as in such inclined events, the EM contamination was largely absorbed in the atmosphere. We aim to extend the MPD reconstruction to lower zenith angles, which can be done by tailoring the MPD reconstruction setup to arrays of buried scintillation detectors, as in the case of AMIGA [25]. In this case, the 2.3 m of soil above the detector provides a vertical mass overburden of $\sim 540\,\mathrm{g\,cm^2}$, allowing for a complete shielding from the EM component of EAS and acquiring a pure muonic signal. As muons also lose energy while propagating underground, a lower threshold E_{th} on the muon energies needs to be imposed that quantifies

the required energy to reach 2.3 m below ground. We can find this limit by solving the equation (1.6):

$$E_{th} = a\rho_g l_g, \tag{4.6}$$

where ρ_g is the ground density and l_g is the distance a muon traverses in the ground. Following the AMIGA setup, we arrive to $E_{th} \approx \frac{1\text{GeV}}{\cos\theta}$, which, from now on, we implement in our analysis.

4.1.2 Machine Learning Setup

As mentioned at the beginning of the chapter, we use the GBDT algorithm via the LightGBM library as our machine learning (ML) model of choice. While GBDT models are known to perform well "out of the box", to fully unlock their potential, a thorough hyperparameter tuning and feature engineering must be implemented in the pipeline. In our case, the hyperparameters were optimized using the Hyperopt library [47], an open-sourced Python package designed to find the hyperparameters which best optimize a custom objective function f_O. The objective function was chosen to reflect our goals: reconstruct the MPD muon-by-muon and reproduce the shape of apparent MPD distribution. The form of our objective function is as follows:

$$f_O = MSE(X) \cdot MSE(f_X), \quad MSE(y) = \frac{1}{n}\sum_{i=1}^{n}\Big((y_i^{pred} - y_i^{true})^2\Big), \tag{4.7}$$

where X denotes the MPD values and f_X represents frequencies of binned X values, with a constant bin width of $20\,\text{g cm}^{-2}$. The superscripts *pred* and *true* represent the model predictions and the target values, respectively. The search range of individual hyperparameters and their optimal values found by Hyperopt are displayed in Table 4.2.

Hyperparameter	*Tuning Range*	*Optimal Value*
Learning Rate	[0.001, 1]	0.695
Number of Leaves	{2, 3, ..., 1000}	987
Maximal Depth of Tree	{1, 2, ..., 20}	11
Minimal Child Weight	{0, 1, ..., 5000}	1376
L1 Regularization	[0, 100]	5.0
L2 Regularization	[0, 100]	55.1
Colsample by Tree	[0, 1]	0.279
Subsample	[0, 1]	0.218

Table 4.2: A summary of the optimal values of hyperparameters (of the MPD ML model) found by Hyperopt, alongside the respective search ranges.

4.1.3 Feature & Target Selection, Feature Engineering

Due to the Arrival Time model being the basis of our MPD model, we choose its core variables as the base input features to our ML model. Our base set of features

consists of $\{\sec\theta, \cos\xi, r, t\}$. On the other hand, choices of targets for our model consist of the Muon Production Depth X, the Muon Production Distance z, the kinematic delay τ_k, or some transformation of the aforementioned targets. From the results of various optimization trials, $\log_{10}(z)$ was chosen to be the optimal candidate for the target.

As any ML model, the GBDT algorithm benefits from feature engineering. Due to the partitioning nature of regression trees and their step-wise output, effectively outputting piece-wise functions, the GBDT algorithm specifically might have trouble combining features in arithmetic operations such as addition or multiplication [48]. Thus, we make new features for the GBDT model to reach its full potential. We again used the Hyperopt package to create and select features that boost the model's performance the most. New features were chosen to be created from the base features by either or both:

1. Transforming individual features through elementary functions
   $\implies x \to \{x^y, \sqrt{x}, \log_{10}(x), \exp(-x)\}, \quad y \in \{2, 3\}$
2. Combining individual features through arithmetic operations $(+, -, \times, /)$

In the optimization process, up to 10 new features could be created. However, in most of our trials, creating more than 5 features did not improve the model's performance. By minimizing the objective function (4.7), the best performing model was found to contain the following set of additional features: $\left\{\frac{ct}{r}, \log_{10}(r), \frac{\log_{10}(r)}{\log_{10}(ct)}\right\}$, where c is the speed of light. The reason of utilizing logarithms might reflect the fact that the r and t distributions are typically skewed with a right tail, i.e., the majority of muons are registered close to the shower core and arriving early [35]. The logarithmic transformation changes the shape of a left-skewed distribution into a one closer to a normal distribution, which might help to boost the performance of the ML model.

4.1.4 Final Data Pre-processing

As described in Chapter 3, we need to separate our data into *training*, *validation* and *test* datasets. As the training dataset, we chose 450 proton- and 450 iron-initiated thinned air showers from the second batch, i.e., with continuous zenith angle values and a fixed primary energy of 10^{17} eV. We chose the QGSJET-II.04 as the underlying hadronic interaction model. From this batch, we used 50 proton- and 50 iron-initiated showers as the validation dataset, defining the training-validation ratio to be 90%/10%. The remaining showers serve as various test datasets.

As a final step in the pre-processing pipeline, we perform data undersampling on the training dataset. The reason behind this decision is to reduce the ML model's bias to prioritizing heavier primary cosmic rays, iron-initiated showers in our case, and their MPD distribution characteristics. The MPD distributions of proton- and iron- initiated air showers have different depths of the shower maxima $X^{\mu}_{\max}$ and number of muons (see the Heitler-Matthews model in Chapter 2), which could introduce additional bias to the model. Therefore, our goal is to undersample the

target distribution so that the corresponding MPD distribution is almost uniformly distributed. With this procedure, we risk losing important information that may be contained in the discarded data, but, on the other hand, we mitigate the risk by having a large training dataset to begin with. For the undersampling process, we use the Imbalanced-learn library [49], which is designed to handle imbalanced datasets by under- and over-sampling the available data. We use the Random-Undersampler algorithm, which finds the most under-represented category, the bin of the MPD distribution in our case, and randomly selects data from all other categories to discard, resulting in a uniform distribution of the target data. At larger zenith angles, the MPD distributions have a long right tail towards large values of X (or small z), implying that we have a smaller amount of data for higher MPD values. To mitigate this effect, we set an upper limit in X above which the undersampling is not performed. The resulting MPD distribution is a uniform distribution up to the upper limit, followed by a right tail, as depicted in Fig 4.2. The upper limit was optimized using the Hyperopt package and was found to be 800 g cm^{-2}.

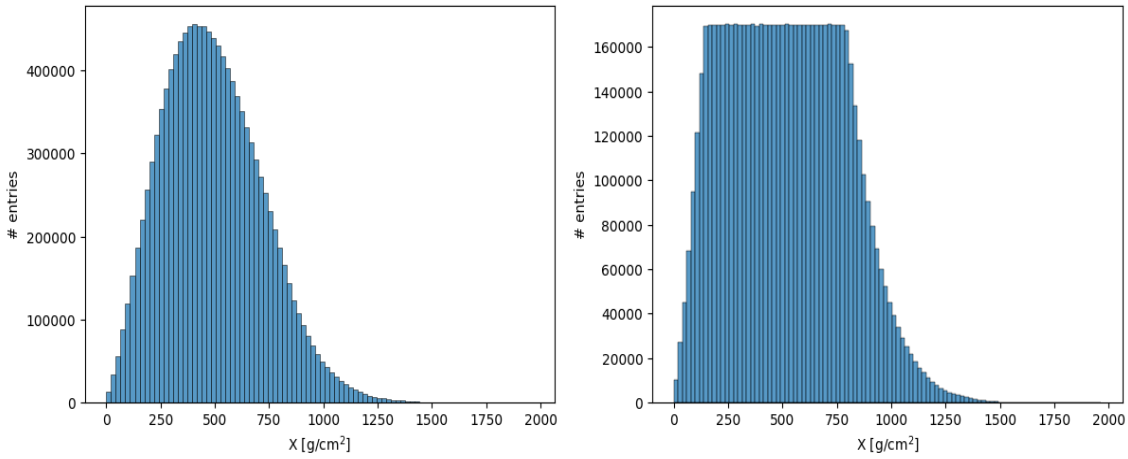

Figure 4.2: An illustration of the undersampling technique implemented in the MPD analysis data pre-processing.

4.1.5 Training Procedure & Model Evaluation

The training procedure is based on an iterative decision tree creation, explained in Chapter 3. The loss function was chosen to be the MSE (see (3.3)), which is evaluated for both the training and validation datasets after each subsequent tree. To counter the overfitting risk, we implement the early-stopping callback, which halts the training after 10 non-improving iterations in the validation dataset MSE. The training for our model stopped after **53** created trees, registering $MSE_{training} = \mathbf{0.02}$ and $MSE_{validation} = \mathbf{0.145}$. The difference between the two values is a consequence of the undersampling of the training dataset.

4.2 Model's Performance Results

In this section, we show the performance of our trained model on the remaining showers, yet to be seen by the GBDT: First, our model is applied to the first batch of air showers with fixed zenith angles and a single primary particle energy. Then, we make predictions on the rest of the second batch, showing results for air showers following a continuous distribution of zenith angles, ranging between 0 and 65 degrees, uniform in $\sin^2\theta$, and fixed primary particle energy. Lastly, we study the model performance applied to a continuous energy and zenith angle distribution of air showers, the dataset which is closer to reality.

We investigate the model's reconstruction performance both from the individual muon perspective (from now on the muon-by-muon treatment) and via the shower-wise treatment, where we compare the shape of the reconstructed MPD distributions with the one from the Monte Carlo and also the depth at which the maximum of the production rate of muons $X^{\mu}_{\max}$ takes place. As explained in Chapter 2, $X^{\mu}_{\max}$ is also a mass-composition-sensitive variable that could be used as an independent and complementary measurement of the nuclear mass composition of cosmic rays in the same fashion as the flagship analyses of the depth of air shower maximum $X_{\max}$. We determine the value of $X^{\mu}_{\max}$ by fitting the MPD distribution with the Gaisser-Hillas function [50], defined as:

$$\frac{dN}{dX} = \frac{dN_{max}}{dX}\Big(\frac{X - X_0}{X^{\mu}_{\max} - X_0}\Big)^{\frac{X^{\mu}_{\max} - X_0}{\lambda}} e^{\frac{X^{\mu}_{\max} - X}{\lambda}}. \tag{4.8}$$

This function has 4 parameters, namely: $X^{\mu}_{\max}$, $N^{\mu}_{\max}$, X_0, and λ, of which we are only interested in the value of $X^{\mu}_{\max}$. Since this function does not describe the tails of the longitudinal profiles, and those are of no use to us, we restrict ourselves to fitting a smaller region around the maximum of the distribution. To find the optimal range where to fit the profiles, we made a scan around the maximum of the distribution and chose the region where we achieved the best reduced χ^2 distribution from the fit. We determine the model's performance shower-wise by comparing the reconstructed and Monte Carlo value of $X^{\mu}_{\max}$.

4.2.1 EAS with Fixed Values of θ and E_{prim}

We start the evaluation of the model's quality using a single primary particle energy, $E_{\text{prim}} = 10^{17}$ eV, and only two zenith angles, 0°, and 60°. This way, we are able to study the model's sensitivity to reconstructing air showers of the "same type", investigating only the physical shower-to-shower fluctuations. Additionally, by studying two very different zenith angles, we are able to see the effects of shower geometries on the model's performance. For each zenith angle, we explore the reconstruction's quality on 100 distinct air showers, both muon-by-muon and shower-wise.

First, we apply the reconstruction to EAS governed by the QGSJET-II.04 model, which allows us to investigate the MPD model's performance without additional systematic uncertainties, which could stem from using a different model of hadronic

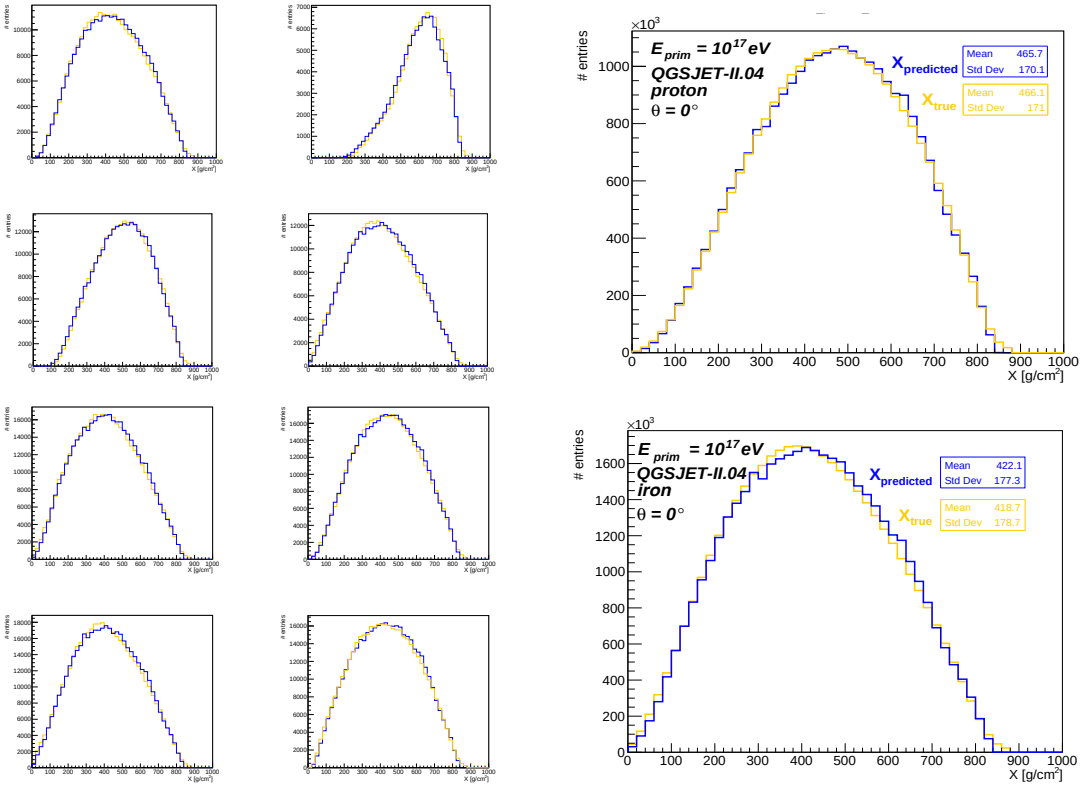

Figure 4.3: Superimposed reconstructed (blue lines) and Monte Carlo (orange lines) MPD distributions for $\theta = 0°$ and $E_{\text{prim}} = 10^{17}$ eV. Left: Examples for the reconstruction of individual showers. Right: Average of 100 showers.

interactions. Figures 4.3 and 4.4 show samples of MPD reconstructions on individual air showers, followed by "averaged" MPD reconstructions on all 100 air showers for both zenith angles and the proton- and iron-induced showers. The muon-by-muon reconstruction characteristics are depicted in Figure 4.5. It can be seen that in the low-zenith case, the reconstruction fares well, correctly matching the shapes of the Monte Carlo MPD distributions, both for the cases of individual air showers and for the averaged one. We assess the model's performance in reconstructing the production depth of each muon through the distribution given by the relation $\Delta X = X_{predicted} - X_{MC}$. From inspection of Figure 4.5, we can see that the muon-by-muon of the MPD reconstruction is practically unbiased, $\langle \Delta X \rangle < 10\,\text{g cm}^{-2}$, while the precision of the reconstruction amounts to $\sigma_{\Delta X} < 80\,\text{g cm}^{-2}$. As we will see, these results comprise a slight upgrade with respect to the current method of MPD reconstruction. For the high-zenith angle showers, the quality of the reconstruction worsens, and a small bias of $\sim 20\,\text{g cm}^{-2}$ is observed for iron-induced showers. Our findings are in agreement with previous studies. However, since we are focused on the zenith angle range of $\theta < 50°$, such studies lay outside the scope of our work. We further observe a degradation of the method's resolution of $\sigma_{\Delta X} \simeq 165\,\text{g cm}^{-2}$, which may be caused by the longer muon trajectories, roughly scaling with $\sec\theta$.

In Figures 4.6 and 4.7, we compare the performance of our method with the standard one. By inspecting of the Figures, it can be seen that our method provides a better

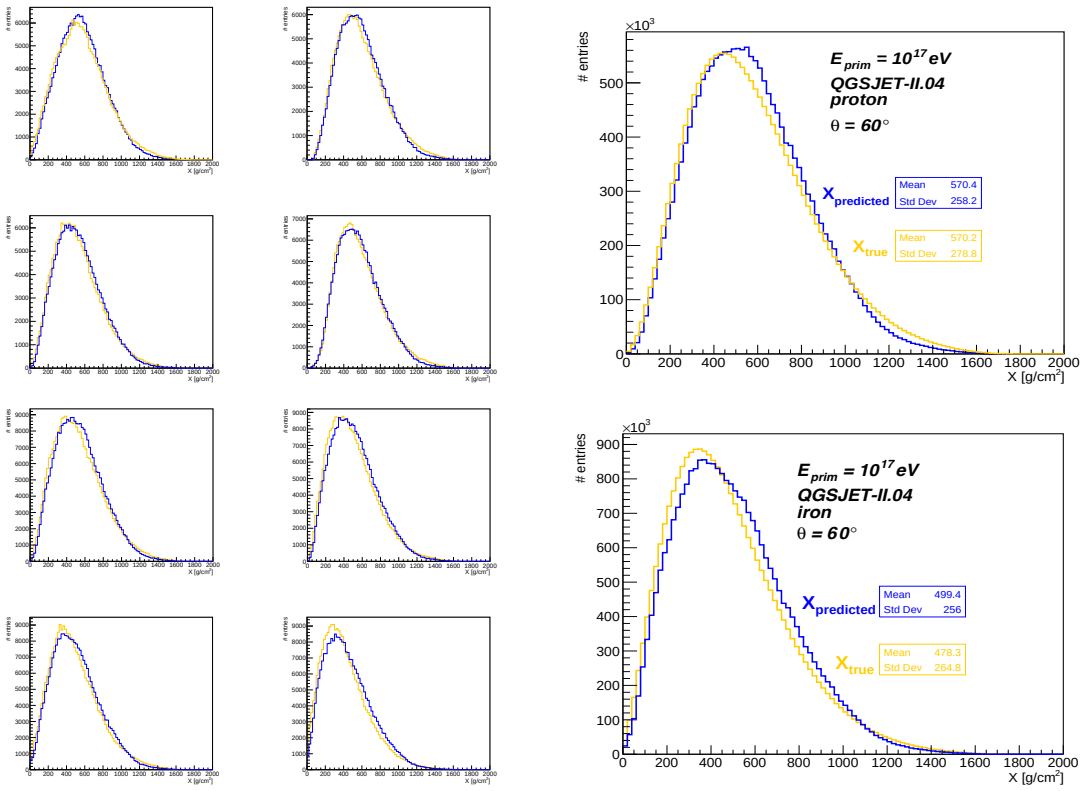

Figure 4.4: Superimposed reconstructed (blue lines) and Monte Carlo (orange lines) MPD distributions for $\theta = 60^\circ$ and $E_{\text{prim}} = 10^{17}$ eV. Left: Examples for the reconstruction of individual showers. Right: Average of 100 showers.

reconstruction of the MPD, considering both the muon-by-muon precision and the reproduction of the overall shape of the MPD distribution. Let us keep in mind that the current method was tuned for zenith angle showers $\theta \sim 60^\circ$ and radial distances of $r \gtrsim 1000$ m. Its main objective was to be applied to a surface detector array of water-Cherenkov stations as the one used by the Pierre Auger Observatory. It is, therefore, not surprising that, when using it at very-low zenith angles, $\theta = 0^\circ$, and $r > 200$ m, the reconstruction fails in most cases.

Even in the case when we restrict ourselves to a subset of muons properly reconstructed by the current MPD model, it still makes biased predictions, with $|\langle \Delta X \rangle|$ being as much as ≈ 55 g cm^{-2}, while our ML model remains almost unbiased. Our earlier reasoning on a slight upgrade in $\sigma_{\Delta X}$ originates here, where we register an improvement in $\sigma_{\Delta X}$ of $\sim 8 - 23$ g cm^{-2}.

Overall, we can make a case for the proposed ML model to have certain advantages over the current MPD reconstruction method. However, its worse reconstruction quality in the high-zenith case needs to be addressed and investigated. A standard way of achieving a better reconstruction is to increase the $r \geq 200$ m cut. The reconstruction of the MPD close to the shower core is not trivial, as, at such small distances, the energy spectrum is more varied than at larger distances (of the order of $r \gtrsim 1000$ m), which makes the estimation of the kinematic delay more challenging. Additionally, the arrival time distribution of the muons close to the shower core

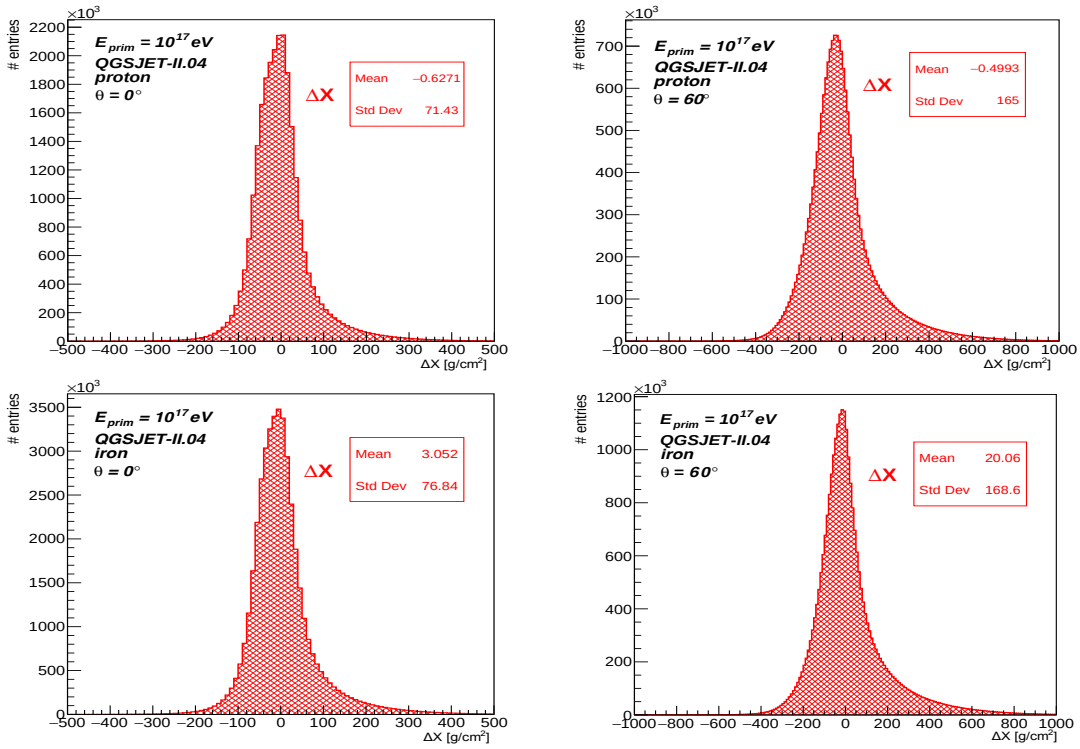

Figure 4.5: Distributions of ΔX for $E_{\text{prim}} = 10^{17}$ eV, for 0° (left) and 60° (right) zenith angles, for proton- (top), and iron-induced showers (bottom).

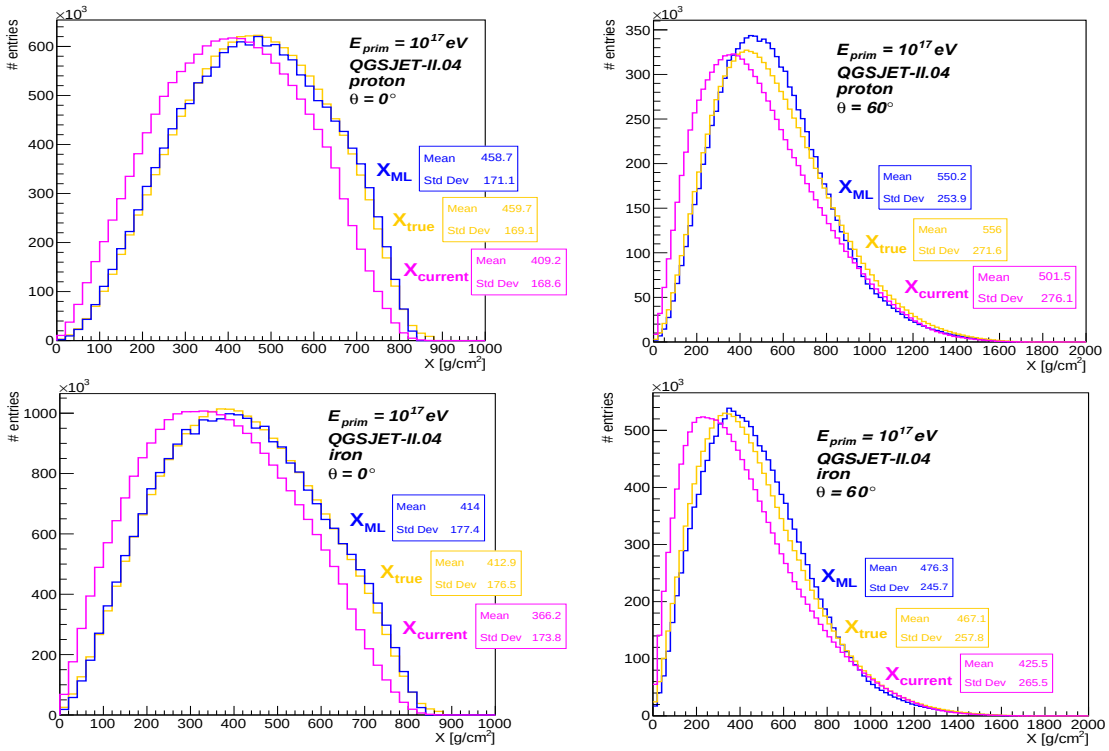

Figure 4.6: Comparison of the MPD distributions, depicting the current model (pink), our ML model (blue), and the Monte Carlo profile (orange) for reference.

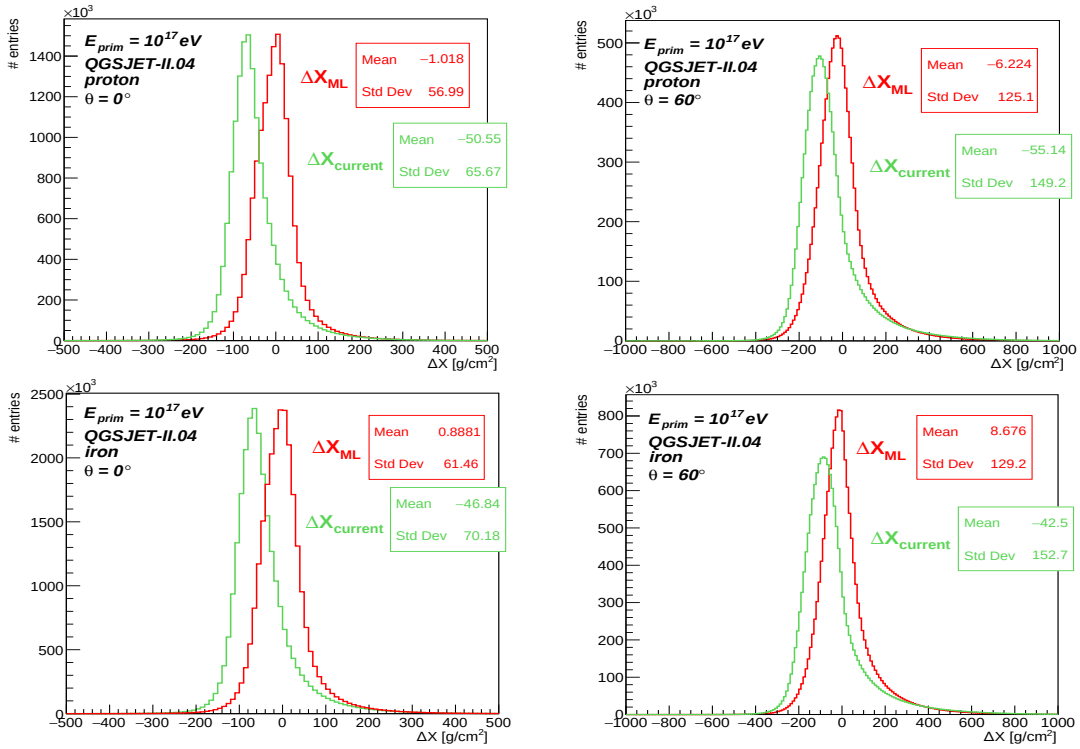

Figure 4.7: Evaluation of the current (green lines) and ML (red lines) reconstruction methods' performances, denoted as $\Delta X_{current}$, and ΔX_{ML}, respectively.

is more compact, which also brings limitations to the reconstruction methods. An alternative to achieve a better reconstruction is to increase the radial cut to 500 or even 1000 m, at the cost of losing a significant fraction of the available muons. We investigate the same reasoning for our ML model in Fig. 4.8.

From the inspection of Fig. 4.8, we verify that the reconstruction quality increases with increasing distance from the shower core. In our case, it may indicate that the ML model is missing a feature that would boost its performance close to the shower core, e.g., the muon energy. However, ground detector arrays of cosmic-ray observatories do not have the ability to measure muon energies and, therefore, making cuts in the distance from the shower core might be an option for a better reconstruction performance[3]. While for $r \geq 500$ m, the improvement in $\sigma_{\Delta X}$ is by one third, for $r \geq 1000$ m, it is by around one half of $\sigma_{\Delta X}$. This is particularly noticeable for air showers with $\theta = 0°$, where the improvement goes up to almost 90 g cm^{-2}. We also stress that the radial cut has an impact on the shape of the MPD profiles. This effect is more noticeable for low-energy showers, i.e., $\theta < 60°$. However, this improvement is achieved at a price of losing $\gtrsim 80\%$ of muons, depending on the zenith angle. Thus, it is imperative to find a balance between the method's precision and the amount of discarded muons. A better-explaining relationship of this phenomenon will be shown in the continuous-zenith subsection.

Next, we investigate the ML model's performance concerning the three main observables at the underground level: the distance from the shower core r, the muon

[3]We note that in the current MPD reconstruction method, $r \geq 1700$ m.

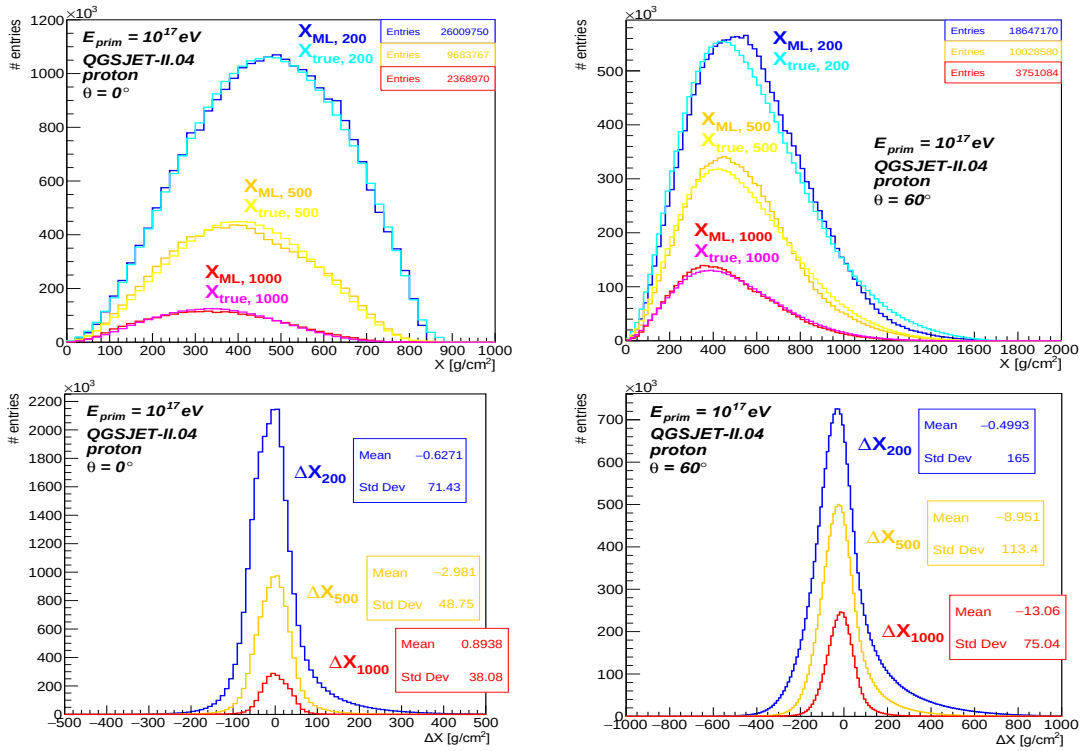

Figure 4.8: Top panel: Monte Carlo and ML reconstructed MPD profiles for the average MPD distribution of 10^{17} eV and proton-initiated showers at $\theta = 0°$ (left), and 60° (right), when applying radial cuts of 200 m (blue), 500 m (orange), and 1000 m (red). Bottom panel: The distribution of $\Delta X = X_{predicted} - Xtrue$ for each radial cuts, shown in the panel above.

arrival time t, and the muon energy E. Here, it is also desirable to compare the model's performance on air showers initiated by different primary cosmic-rays. To get an idea about the differences between proton- and iron-initiated air showers, we show the Monte Carlo MPD distributions for both primary particles in Fig. 4.9. As predicted by the Heitler-Matthews model, iron-initiated showers give rise to ~ 1.8 times more muons than proton-initiated showers. Additionally, the maxima of the respective distributions are different, which is why $X^{\mu}_{\max}$ is the main observable in muon-related mass-composition studies. We note that these two facts are behind our reasoning for applying undersampling to our training data. Otherwise, the ML model might get biased towards the MPD characteristics of iron-induced showers.

In Figures 4.10, 4.11, and 4.12, we show the method's reconstruction bias and precision, $(\langle \Delta X \rangle, \sigma_{\Delta X})$ for proton-, and iron-initiated showers, as a function of the shower core, muon arrival time, and muon energy, respectively. We can immediately observe that the method's reconstruction performance is generally better for iron-initiated showers than for protons. Another observation concerns the "erratic" behavior of $\langle \Delta X \rangle$ with respect to r and t, which might be the consequence of the partitioning nature of the GBDT algorithm. Both observables are used as features in the ML model and, therefore, the predicted target is a multi-dimensional step-wise function of these features. We can see that this behaviour is not replicated in the muon-energy

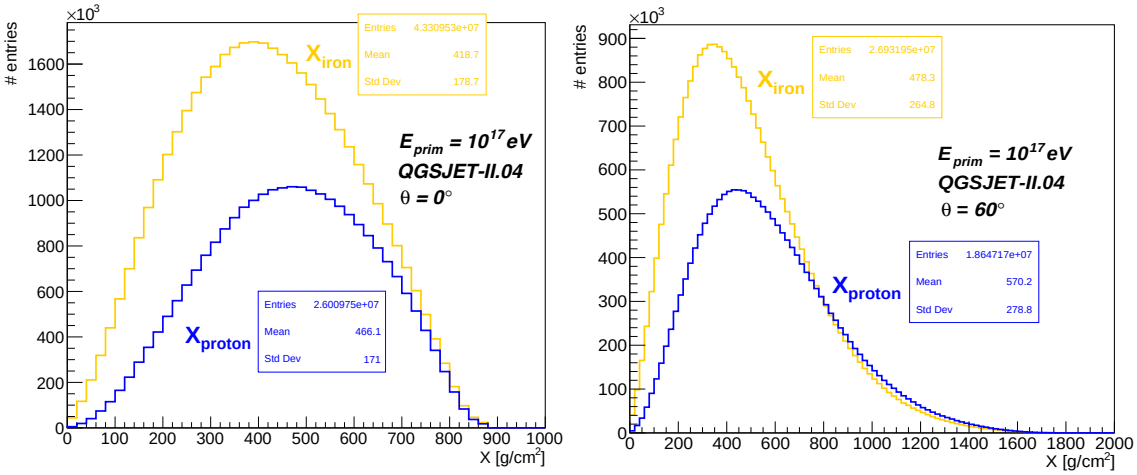

Figure 4.9: Comparison of the Monte Carlo MPD distributions for proton and iron primary particles, $E_{\text{prim}} = 10^{17}$ eV.

dependence, shown in Fig. 4.12. On the other hand, the values of $\sigma_{\Delta X}$, whose relationships mostly resemble smooth curves, predominantly drop with increasing values of the independent variables. This behavior is due to the fact that the observables are correlated into some extent: For example, muons that land close to the shower core also typically arrive earlier than muons landing far from the shower core. Most of the shower muons also land close to the shower core, arrive earlier rather than later and have smaller energies, which is the reason for insufficient statistics at larger values of those observables.

In Fig. 4.12, we can observe that the ML reconstruction is biased with respect to the entire muon energy spectrum. For small muon energies, $\langle \Delta X \rangle$ can be as high as 130 g cm^{-2}, while at high energies, $\langle \Delta X \rangle$ goes down to -110 g cm^{-2}. This is an undesirable effect, which might be caused by the model not having any information about the muon energies. We will investigate this effect in the next chapter, where we build a second ML model that predicts the muon energies from the available features.

To further investigate the model's generalization capacity, we introduce air showers simulated with different models of hadronic interaction, specifically EPOS-LHC and Sibyll-2.3d. As an additional comparison, we have simulated the subset of Sibyll 2.3d showers without activating the thinning algorithm and used these showers to investigate the impact of the thinning algorithm on our results. In Figure 4.13, we show the average Monte Carlo MPD distributions for proton-, and iron-initiated showers, for the three hadronic interaction models. From inspection of Figure 4.13, we observe that both EPOS-LHC and Sibyll 2.3d predict approximately the same X^{μ}_{max} for both primary species. However, QGSJET-II.04 is known for predicting shallower values for X^{μ}_{max}. These observations are in agreement with the results presented by the Pierre Auger Observatory [37] and make us conclude that X^{μ}_{max} is an observable which is sensitive to the hadronic interactions occurring in the shower development. We chose QGSJET-II.04 as our model for training basis since, in the analysis made in [37], QGSJET-II.04 was the model that allowed for a more

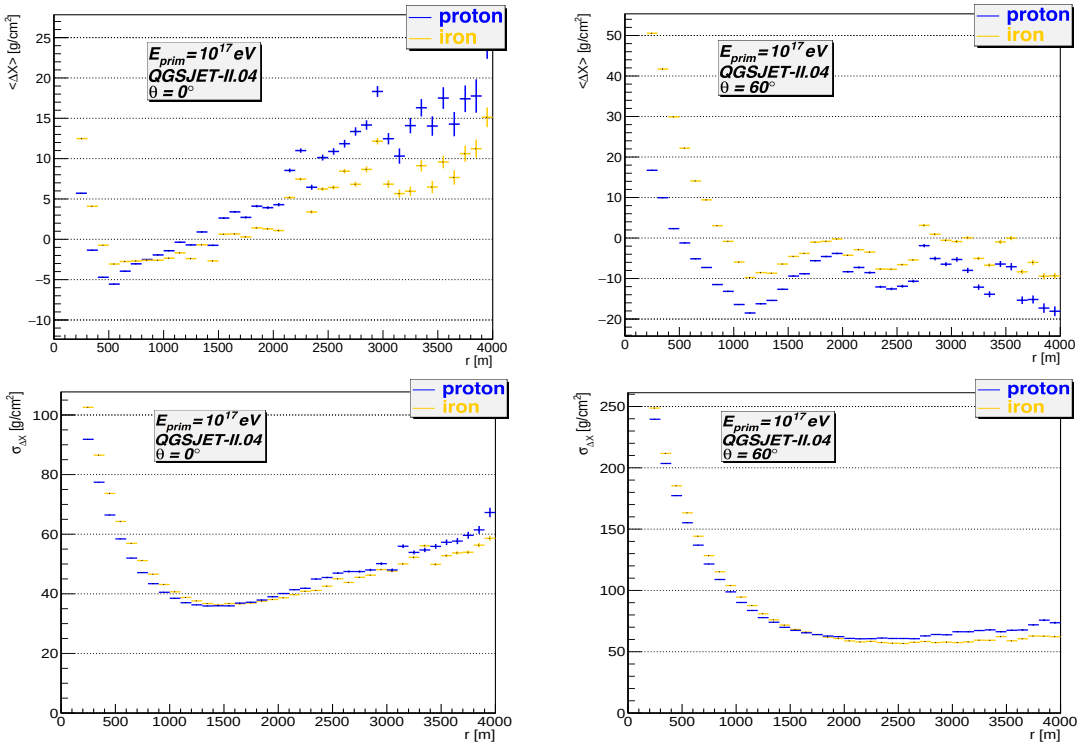

Figure 4.10: Bias (top) and resolution (bottom) of the MPD reconstruction as functions of the shower core distance r, comparing proton and iron primary particles.

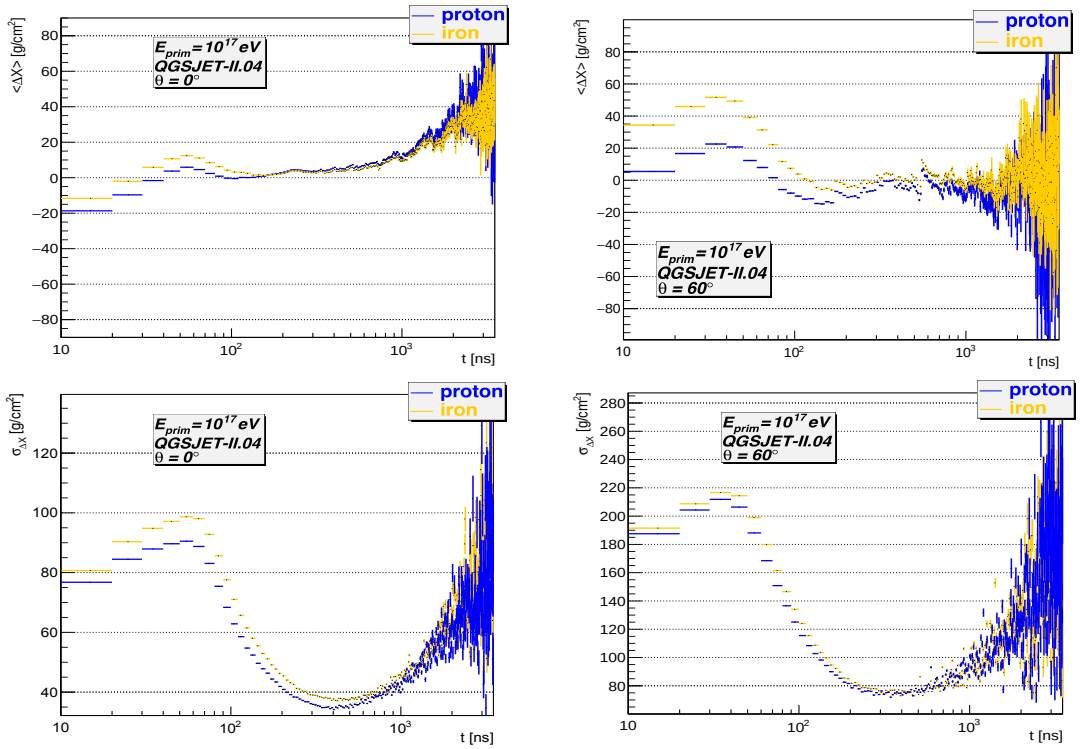

Figure 4.11: Bias (top) and resolution (bottom) of the MPD reconstruction as functions of the muon arrival time t, comparing proton and iron primary particles.

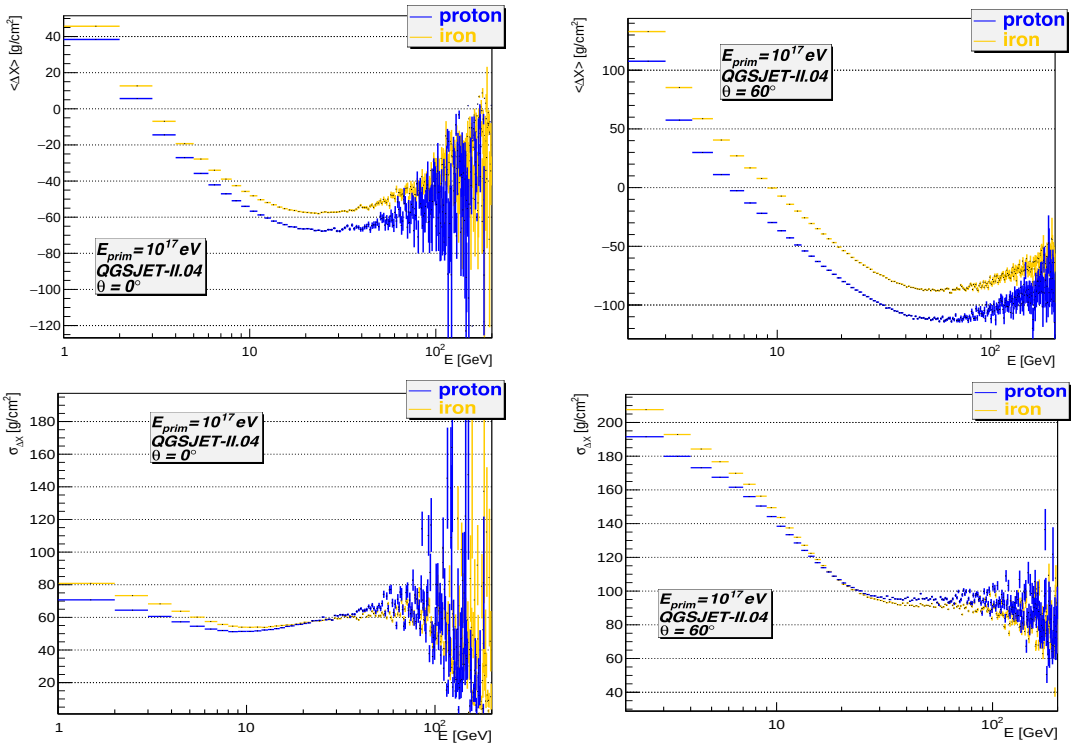

Figure 4.12: Bias (top) and resolution (bottom) of the MPD reconstruction as functions of the muon energy E, comparing proton and iron primary particles.

consistent interpretation between the estimation of the mass composition of cosmic rays using $X^{\mu}_{\max}$, and $X_{\max}$ from the electromagnetic profile. On the other hand, the interpretation of the nuclear mass composition of cosmic rays for $X^{\mu}_{\max}$, when using EPOS-LHC, yielded an unphysical mass composition heavier than Uranium. We will, nevertheless, study all three models to infer whether our ML model can make similar predictions regardless the hadronic interactions model.

Figures 4.14, 4.15 and 4.16 show similar relationships for the reconstruction's bias $\langle \Delta X \rangle$ and resolution $\sigma_{\Delta X}$ as in the proton-iron comparison. From the Figures, we can conclude that the method's performance for all three models is very similar. The differences are subtle, with the method's bias almost predominantly ranging from 0 to 10 g cm^{-2} (the largest being approximately $\langle \Delta X \rangle = 15$ g cm^{-2} between QGSJET-II.04 and Sibyll-2.3d close to the shower core). The resolution $\sigma_{\Delta X}$ is almost identical in all respective relations. As a last note, we do not observe any significant deviation of the Sibyll-2.3d model from the others, which suggests that the thinning algorithm does not significantly impact the ML model's quality of reconstruction.

We now switch from the muon-by-muon treatment of MPD to the shower-wise treatment of individual MPD distributions, through which we are able to infer the mass-composition-sensitive variable $X^{\mu}_{\max}$. By fitting the Gaisser-Hillas function to the MPD distribution, we acquire $X^{\mu}_{\max}$ as detailed in section 4.2. We apply this procedure to the Monte Carlo and reconstructed MPD profiles of proton- and iron-induced showers, with $\theta = 0°$, and $60°$, and plot their $X^{\mu}_{\max}$ distributions. Our results are shown in Figure 4.17. From a closer inspection, we see that the $X^{\mu}_{\max}$ distributions

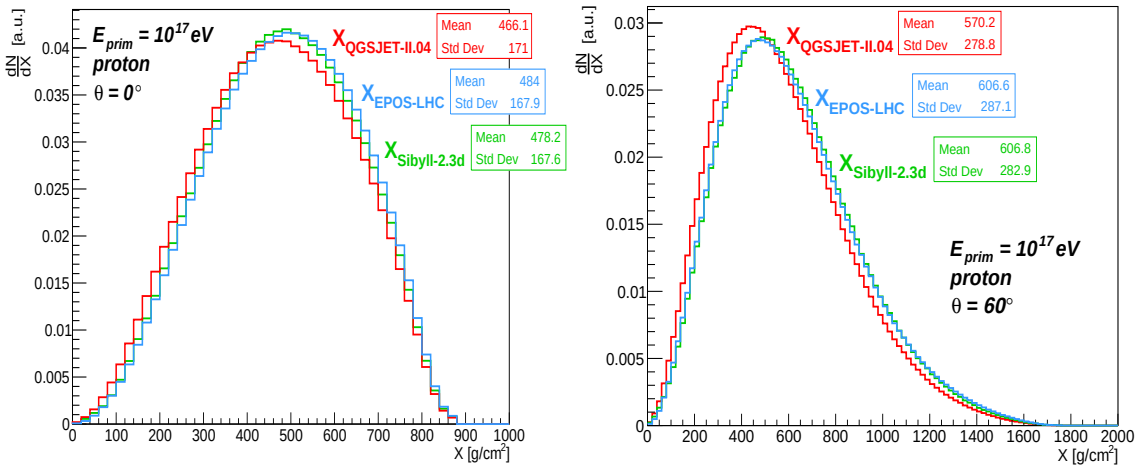

Figure 4.13: Monte Carlo MPD distributions for an average of 100 showers, using the QGSJET-II.04 (red), EPOS-LHC (blue), and Sibyll 2.3d (green) hadronic interaction models. Left: Proton-initiated showers, with $E_{\rm prim} = 10^{17}$ eV, and $\theta = 0°$. Right: Proton-initiated showers, with $E_{\rm prim} = 10^{17}$ eV, and $\theta = 60°$.

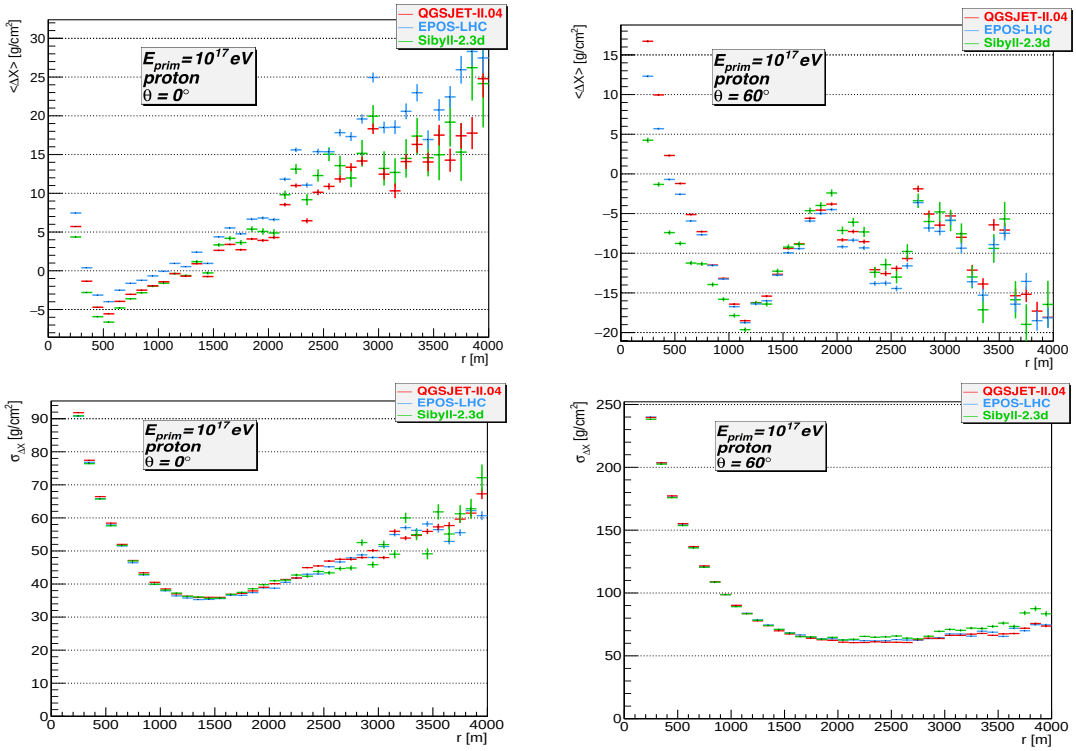

Figure 4.14: Bias (top) and resolution (bottom) of the MPD reconstruction as functions of the shower core distance r for the three models of hadronic interactions.

are wider and occur deeper in the atmosphere for proton-induced showers than for iron, two features which we expect from the superposition model. We further observe that, for the Monte Carlo profiles $X^{\mu}_{\rm max}$ distributions, for $\theta = 0°$, the $\sigma\left(X^{\mu}_{\rm max}\right)$ is $\sim 60\,{\rm g\,cm^{-2}}$ for proton species, and of $25\,{\rm g\,cm^{-2}}$ for iron, values which go inline with the values observed for the shower-to-shower fluctuations observed for $X_{\rm max}$, which

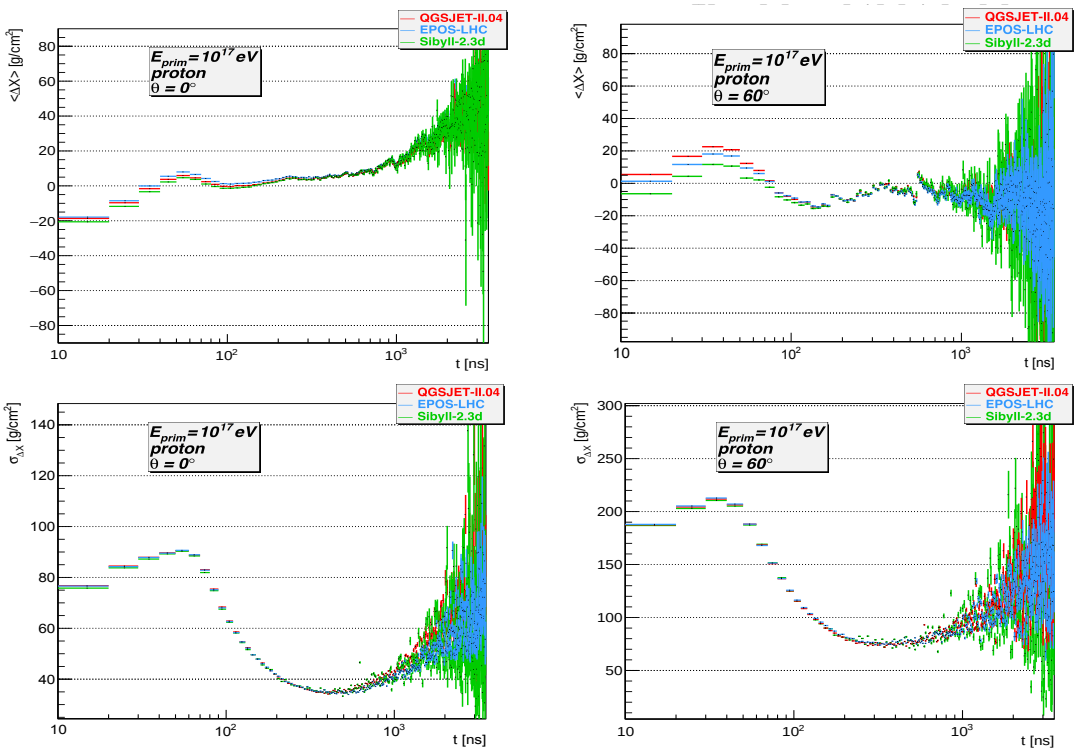

Figure 4.15: Bias (top) and resolution (bottom) of the MPD reconstruction as functions of the arrival time t for the three models of hadronic interactions.

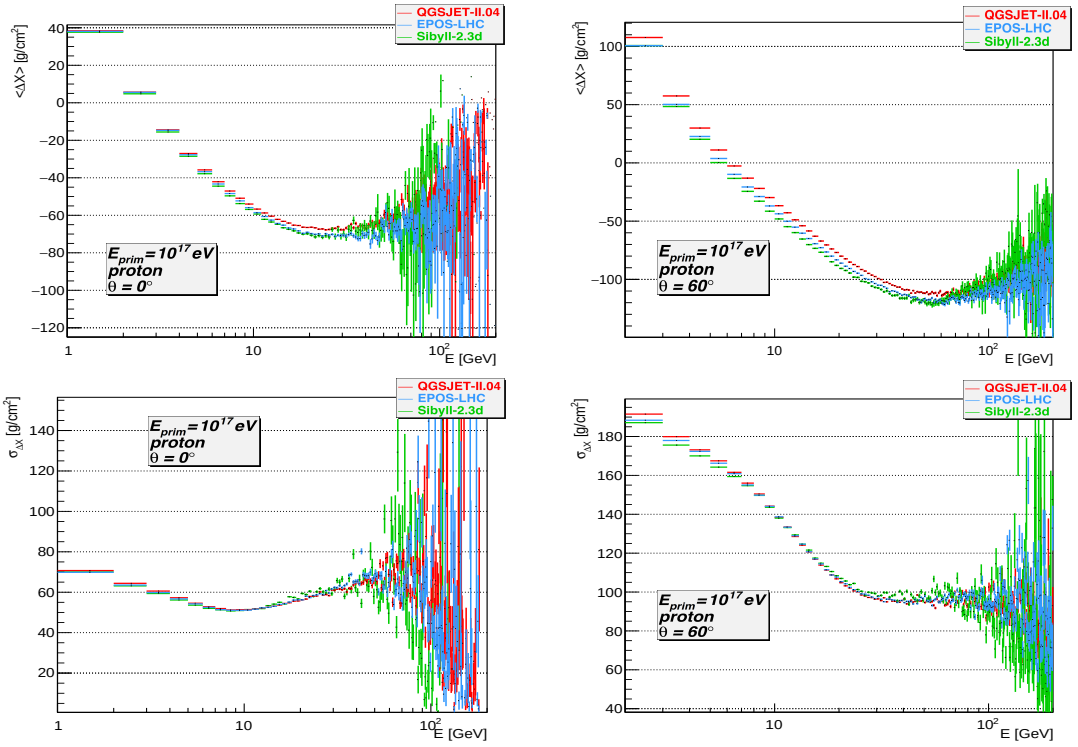

Figure 4.16: Bias (top) and resolution (bottom) of the MPD reconstruction as functions of the muon energy E for the three models of hadronic interactions.

are of $\sim 60\,\mathrm{g\,cm^{-2}}$ for proton species, and of $\sim 20\,\mathrm{g\,cm^{-2}}$ for iron, respectively. However, for $\theta = 60°$, we observe slightly higher values which may be connected to the details of hadronic interactions in the shower cascade. Regarding the distributions of the reconstructed profiles, we observe slight biases towards shallower values of $\langle X^{\mu}_{\mathrm{max}} \rangle$. We also see larger values for $\sigma\,(X^{\mu}_{\mathrm{max}})$, more pronounced for proton-induced showers at $\sim 60\,\mathrm{g\,cm^{-2}}$, which may indicate a poorer quality of the reconstruction. The reconstruction performs (as expected) much better for the low-zenith angle cases, but to be certain, a larger sample of showers will have to go into the analysis. Fig. 4.18 shows the individual X^{μ}_{max} differences between the ML model's predictions and the Monte Carlo MPD distributions. We see that, while the ML model fares well for showers with $\theta = 0°$, the reconstruction of showers with $\theta = 60°$ is biased by almost a factor of two for both primary species. On the other hand, $\langle \Delta X^{\mu}_{\mathrm{max}} \rangle$ accounts to only about 2% of the total range of X values for the respective zenith angles, with $\sigma_{\Delta X^{\mu}_{\mathrm{max}}}$ being even less than that. This is a good sign for the model's quality of reconstruction moving forward.

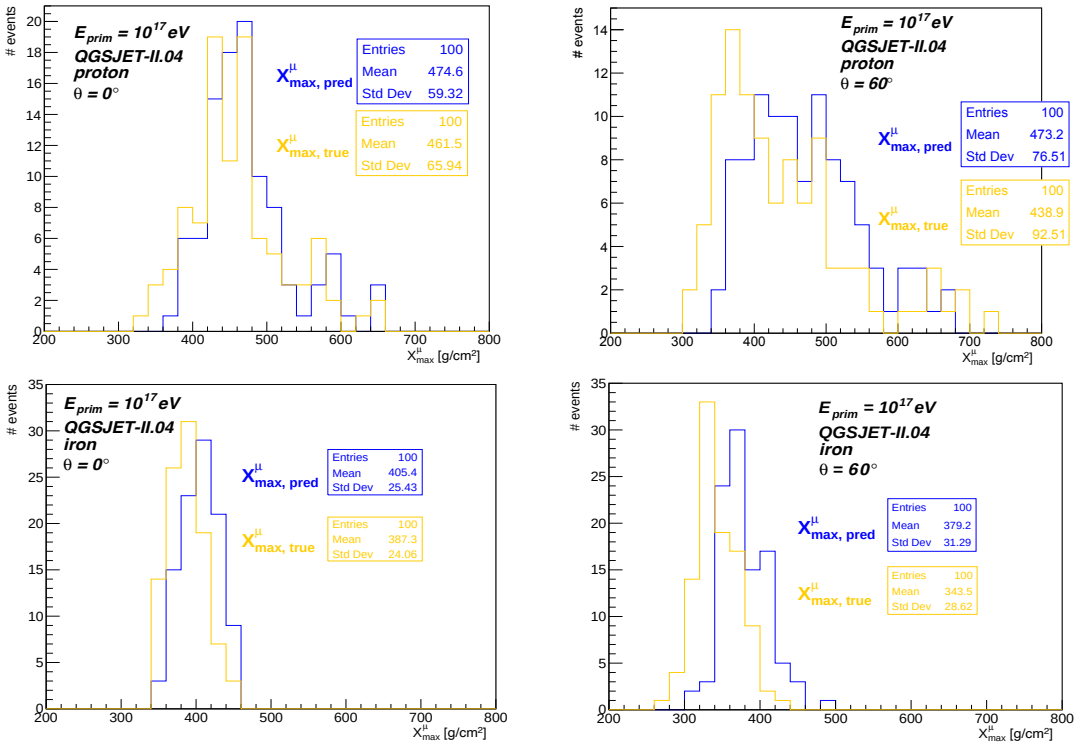

Figure 4.17: Distribution of X^{μ}_{max} from the Monte Carlo (orange) and reconstructed (blue) MPD distributions.

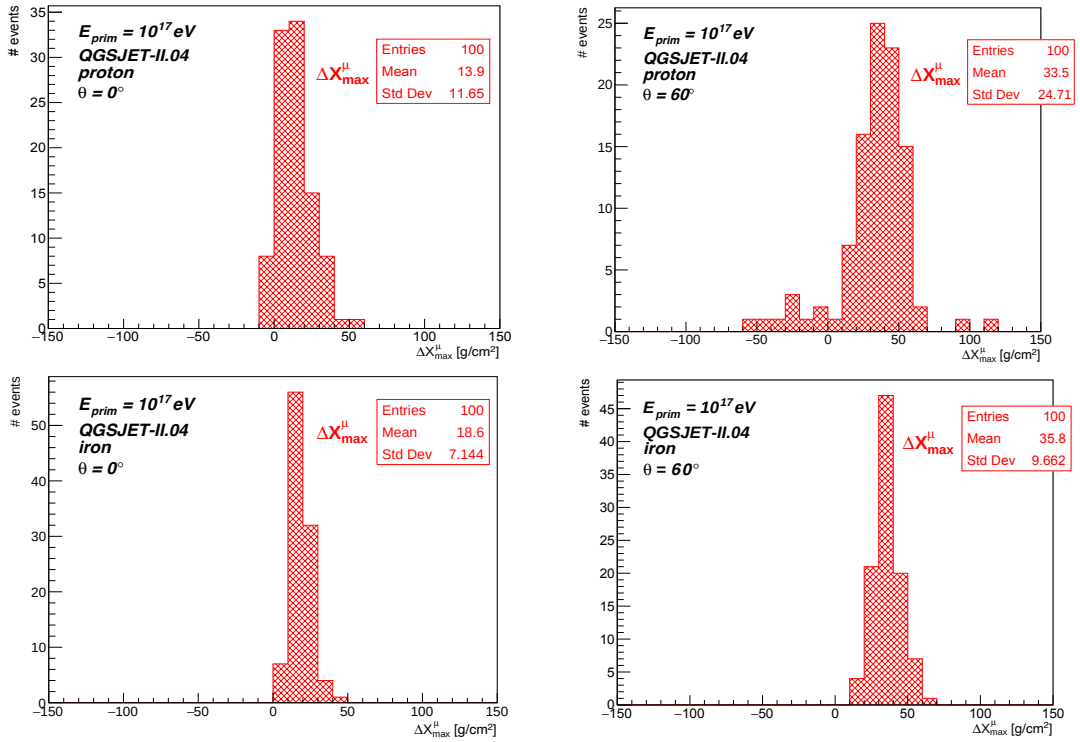

Figure 4.18: Distributions of the $\Delta X^{\mu}_{\max} = X^{\mu,pred}_{\max} - X^{\mu,true}_{\max}$ differences.

4.2.2 EAS with Continuous Values of θ and Fixed E_{prim}

To further our investigation into a more real-case dataset, we will next investigate a sample of air showers with continuous values of zenith angle θ, uniformly distributed in $\sin^2\theta$. This way, we can clearly see the model's reconstruction capability with respect to distinct zenith angles, which, as a variable, is one of the most important quality-deciding criterion in the MPD reconstruction. In this section, we present the ML model's aggregate results as functions of θ.

In Fig. 4.19 and 4.20, we present the mean value and standard deviation of the muon-by-muon reconstruction difference $\Delta X = X_{predicted} - X_{true}$ as functions of θ and how they differ by changing the primary particle and the model of hadronic interactions, respectively. It can be seen that, up to 50°, the reconstruction remains relatively unbiased ($\langle \Delta X \rangle < 15$ g cm^{-2}), independently of the chosen primary particle type or the hadronic interaction model. For higher zenith angles, however, the model's predictions for the iron primary are biased by up to ~ 37 g cm^{-2}. We, however, have to take into account that the range of possible MPD values increases with $\sec\theta$. The bias differences between proton- and iron-induced air showers is mostly small, within 10 g cm^{-2}. However, for $\theta > 60°$, the largest bias accounts to almost 30 g cm^{-2}. We also add that the model consistently predicts higher MPD values for iron-initiated showers. Similarly to what was observed in the previous section, a comparison between QGSJET-II.04 and Sibyll 2.3d, yields similar results, where $\langle \Delta X \rangle < 10\,\text{g cm}^{-2}$, except for the last zenith angle bin. The corresponding standard deviation rises approximately with $\sec\theta$, corresponding to the broadening of the

phase-space of the MPD values. In both cases, the difference in $\sigma_{\Delta X}$ does not exceed 10 g cm^{-2}. This is a hint of the ML model's consistency, while the non-smooth behavior of $\langle \Delta X \rangle$ is a consequence of the model's GBDT-based structure, as argued in the previous section.

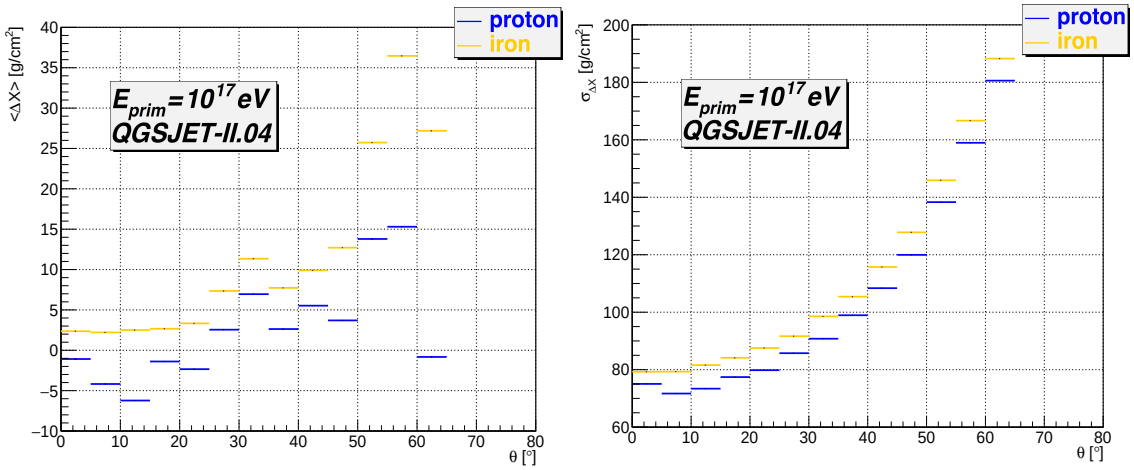

Figure 4.19: Bias (left) and precision (right) of the MPD reconstruction as a function of the zenith angle for $E_{\text{prim}} = 10^{17}$ eV proton- (blue) and iron-initiated showers (orange). The showers are uniformly distributed in $\sin^2\theta$ within the zenith angle range $0° \theta < 65°$.

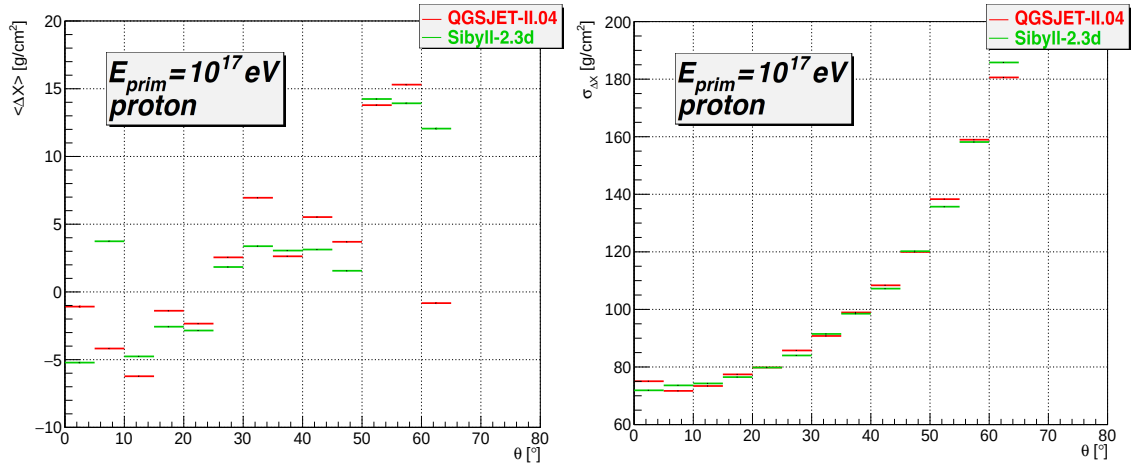

Figure 4.20: Bias (left) and precision (right) of the MPD reconstruction as a function of the zenith angle for $E_{\text{prim}} = 10^{17}$ eV proton-initiated showers using QGSJET-II.04 (red) and Sibyll 2.3d (green) as hadronic interaction models. The showers are uniformly distributed in $\sin^2\theta$ within the zenith angle range $0° \theta < 65°$.

With the continuous library in zenith angle, intriguing relationships between the three characteristic muonic observables (r, t and E) and the zenith angle can be studied. These dependencies are displayed, in the aforementioned order, in Figures 4.21, 4.22, 4.23. We can observe a few similarities in all three behaviors: First, for a given value of the three observables on the y-axes, the reconstruction biases tend

to worsen[4] as we increase θ, as expected. The behaviors of the observables copy the ones from section 4.2.1, magnified approximately by the $\sec\theta$ factor. This effect is particularly visible for the standard deviation plots, which are not affected by the GBDT's "erratic" behavior regarding the model's biases.

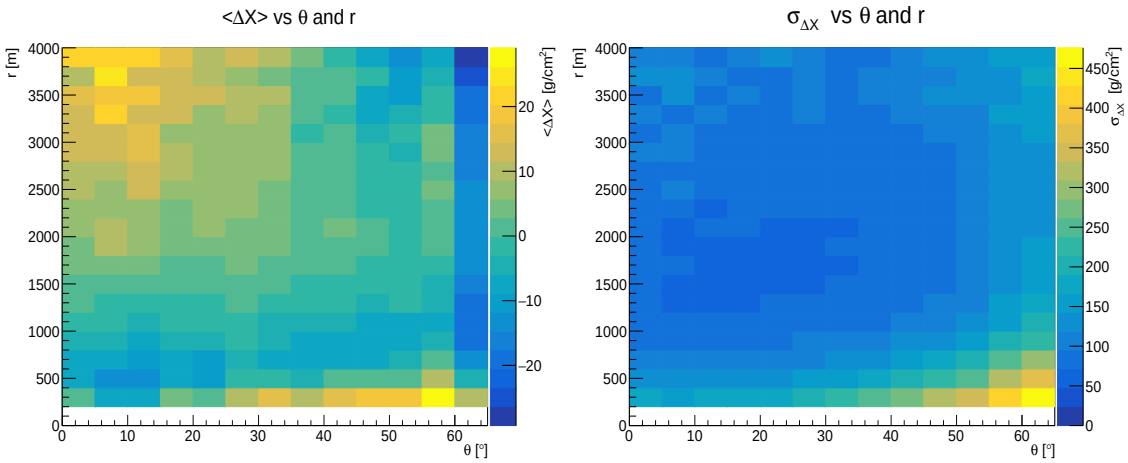

Figure 4.21: The MPD reconstruction characteristics $\langle \Delta X \rangle$ and $\sigma_{\Delta X}$ as functions of the zenith angle θ and the shower core distance r, QGSJET-II.04, $E_{\text{prim}} = 10^{17}$ eV.

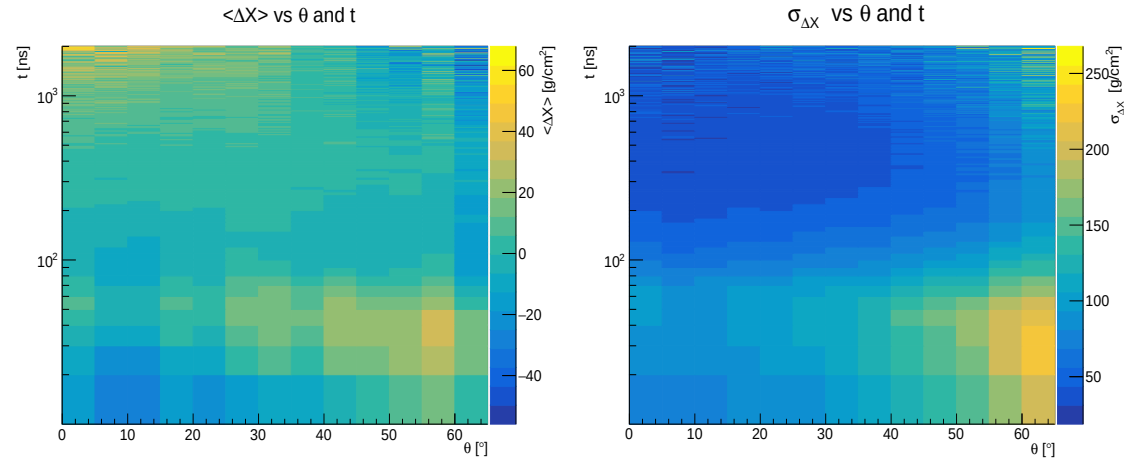

Figure 4.22: The MPD reconstruction characteristics $\langle \Delta X \rangle$ and $\sigma_{\Delta X}$ as functions of the zenith angle θ and the arrival time t, QGSJET-II.04, $E_{\text{prim}} = 10^{17}$ eV.

In the transition to the model's shower-wise performance, we first merge the contributions from all zenith angles and show the zenith-averaged X^{μ}_{max} distributions in Fig. 4.24. We see that the X^{μ}_{max} biases are $\sim 20\,\text{g cm}^2$, a value inline to the one which was obtained in section 4.2.1, where we analyzed showers with fixed zenith angles of 0°, and 60°. The method's performance is on average biased by less than 25 g cm^{-2}, with a resolution of $\sigma_{\Delta X^{\mu}_{\text{max}}} \sim 30 - 33\,\text{g cm}^{-2}$.

[4]We note that as θ increases, the distributions of r, t and E each broaden in range. Thus, for low zenith angles, there are very few muons at large values of the three observables (or none, as we see in 4.23), hence the brighter spots at the right-left corners.

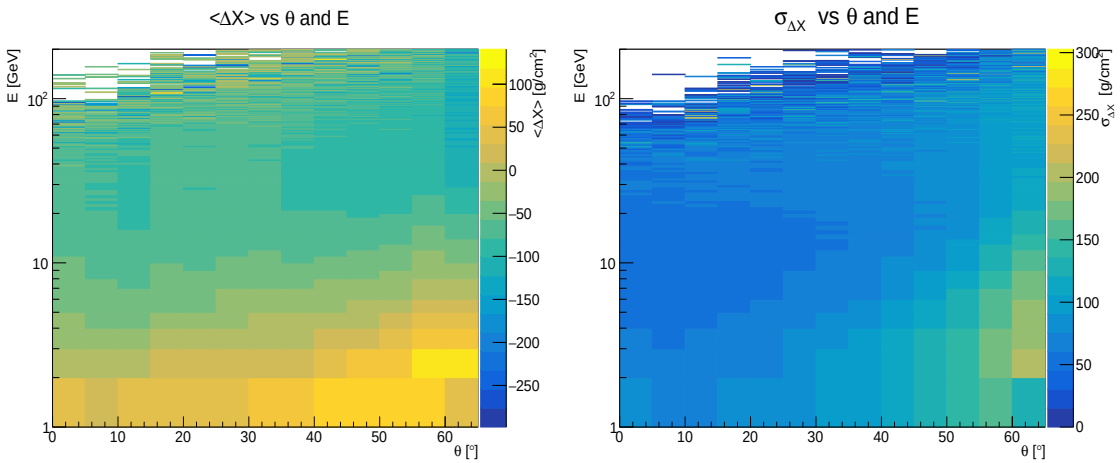

Figure 4.23: The MPD reconstruction characteristics $\langle \Delta X \rangle$ and $\sigma_{\Delta X}$ as functions of the zenith angle θ and the muon energy E, QGSJET-II.04, $E_{\text{prim}} = 10^{17}$ eV.

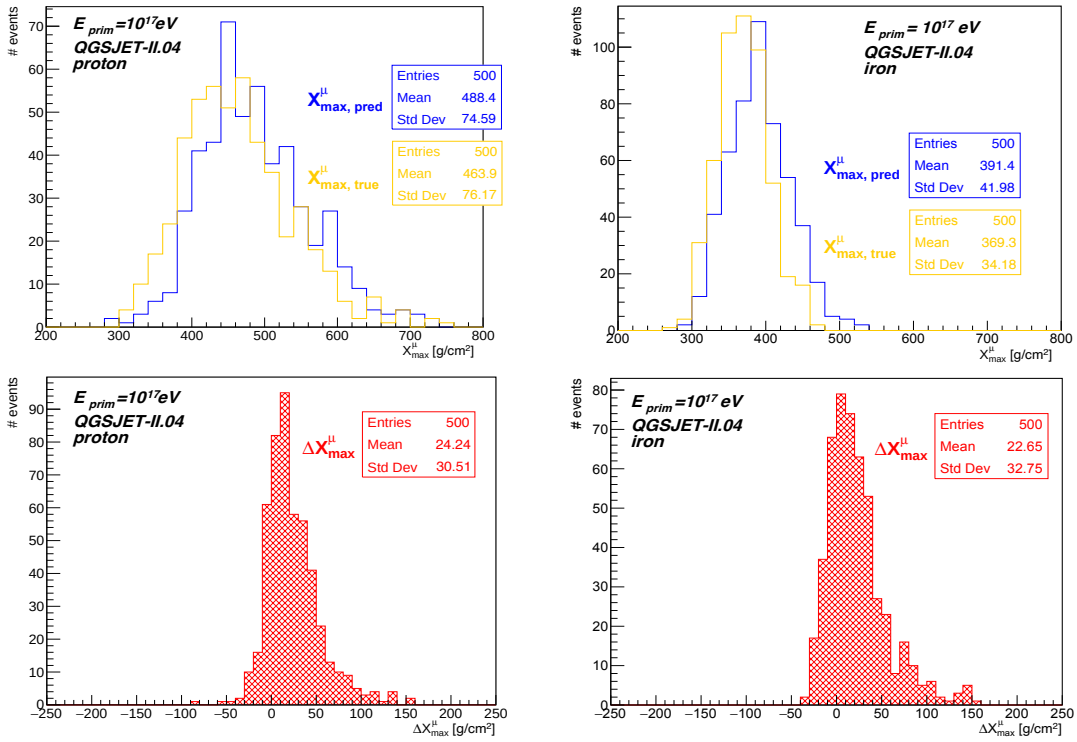

Figure 4.24: Upper panel: Distribution of X^{μ}_{max} from the Monte Carlo (orange) and reconstructed (blue) MPD distributions for proton- (left) and iron-initiated showers (right). Lower panel: Resolution of the X^{μ}_{max} reconstruction, $\Delta X^{\mu}_{\text{max}} = X^{\mu,pred}_{\text{max}} - X^{\mu,MC}_{\text{max}}$ for $E_{\text{prim}} = 10^{17}$ eV proton- (left) and iron-induced showers (right). The showers are uniformly distributed in $\sin^2\theta$ within the zenith angle range $0° \theta < 65°$.

The biases and resolution of the reconstructed X^{μ}_{max} exhibit similar behavior with the zenith angle as the one observed for the muon-by-muon reconstruction. Our results are shown in Figures 4.25 and 4.26. We observe a tendency of a worsening bias and resolution of the reconstructed X^{μ}_{max} for $\theta > 50°$. Below this zenith region,

which is the region of interest for our method, we observe $\langle \Delta X^{\mu}_{\max} \rangle < 30\ \mathrm{g\,cm^{-2}}$, and $\sigma_{\Delta X^{\mu}_{\max}} < 35\mathrm{g\,cm^{-2}}$. Our findings were found to be independent of the type of primary particle and the hadronic interaction model used, allowing us to conclude that our ML model can successfully generalize new data and nicely reconstruct $X^{\mu}_{\max}$.

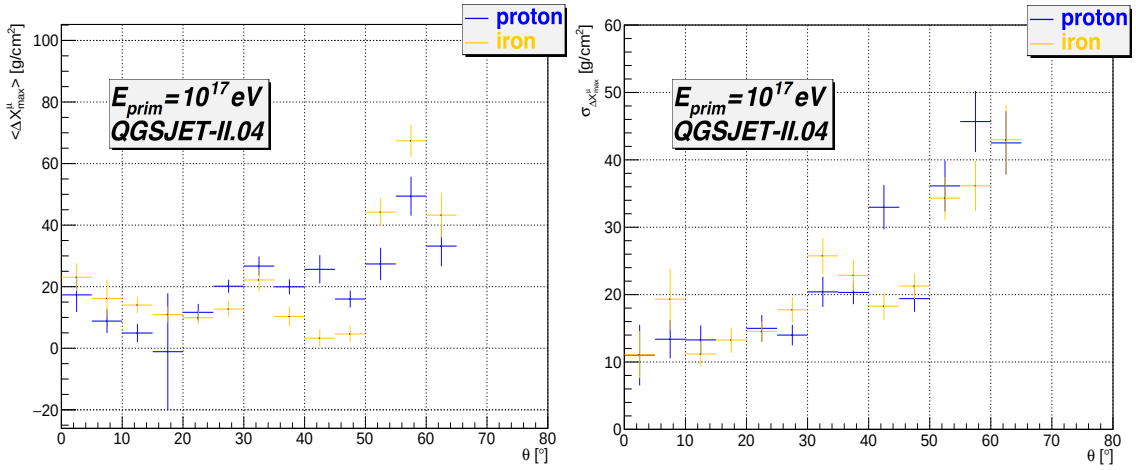

Figure 4.25: Bias (left) and resolution (right) of $X^{\mu}_{\max}$ as functions of the zenith angle for $E_{\mathrm{prim}} = 10^{17}$ eV proton- (blue), and iron-initiated showers (orange), using QGSJET-II.04 as hadronic interaction model. The showers are uniformly distributed in $\sin^2\theta$ within the zenith angle range $0° \theta < 65°$.

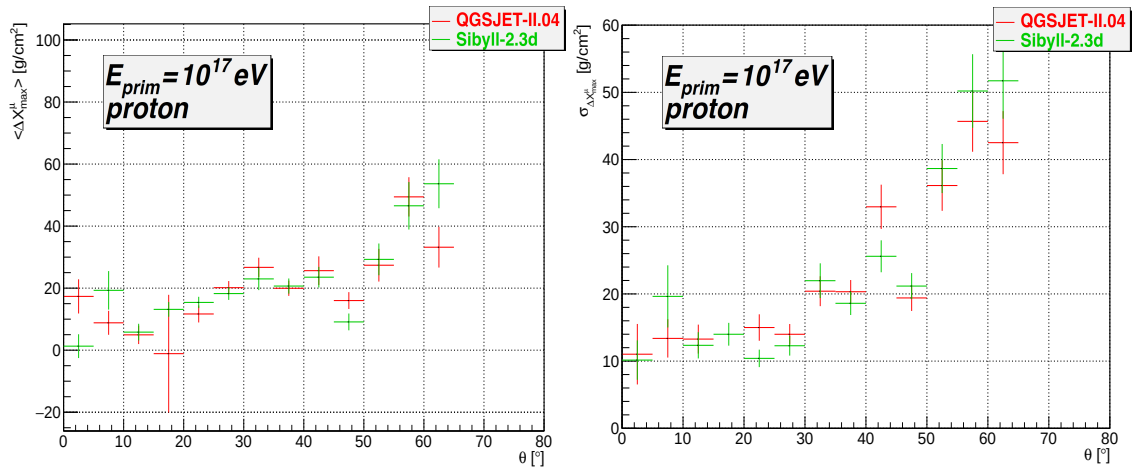

Figure 4.26: Bias (left) and resolution (right) of $X^{\mu}_{\max}$ as functions of the zenith angle for $E_{\mathrm{prim}} = 10^{17}$ eV proton-initiated showers, comparing the QGSJET-II.04 (red) and Sibyll-2.3d (green) hadronic interaction models. The showers are uniformly distributed in $\sin^2\theta$ within the zenith angle range $0° \theta < 65°$.

4.2.3 Continuous EAS library

As a final test, we apply our ML model to the most realistic case used in cosmic-ray physics. In addition to a continuous zenith angle distribution described in the previous section, this last library has a continuous energy spectrum. To further test the ability of the ML algorithm to reconstruct unseen data, we simulated showers in the energy range $10^{18.5}$ eV $\leq E \leq 10^{19}$ eV. According to the Heitler-Matthews model from Chapter 2, one corresponding effect is an increase of muons by a factor of $10^{\beta} - 10^{2\beta}$. Additionally, by increasing E_{prim}, the models of hadronic interactions must further and further extrapolate their predictions, which are based on data from man-made particle colliders. It is, therefore, desirable to test our model at higher energies than the nominal energy at the LHC. If our model performs well, we hope to provide relevant information about hadronic interactions occurring in EAS at $\sqrt{s} > 14$ TeV, the nominal energy at the LHC.

Naturally, we will first study the ML model's performance on the muon-by-muon basis. The dependencies of $\langle \Delta X \rangle$ and $\sigma_{\Delta X}$ on θ and E_{prim} for proton and iron primary particles are shown in Figures 4.27 and 4.28, respectively. We observe that the $\langle < \Delta X \rangle >$ and $\sigma_{\Delta X}$ behavior as a function of the zenith angle is similar to the one observed in the previous subsection for showers with a fixed energy of 10^{17} eV. The method's performance, as a function of the energy of the primary particle, was found to be relatively constant up to $\theta = 50°$, which, again, is not surprising, as we already know that the ML model underperforms for high zenith angles. Additionally, even though we have trained the ML model at lower energies, we obtain comparable values of $\langle < \Delta X \rangle >$ and $\sigma_{\Delta X}$ for $10^{18.5} \leq$ eV $E_{\text{prim}} \leq 10^{19.0}$ eV. This is a hint that the model truly performs the reconstruction muon-by-muon and is not biased towards distribution characteristics such as mean value or median, as is sometimes the case for underdeveloped machine learning models.

Following the structure of the previous sections, we now comment the shower-wise performance regarding the current dataset. In Figures 4.29-4.34, we supply all relevant information on the X^{μ}_{max} observable for every available primary particle (proton, iron) and model of hadronic interactions (QGSJET-II.04, Sibyll-2.3d, EPOS-LHC). We show the reconstructed and Monte Carlo X^{μ}_{max} distributions, the $\Delta X^{\mu}_{\text{max}}$ distribution and relationships between $\langle \Delta X^{\mu}_{\text{max}} \rangle$, $\sigma_{\Delta X^{\mu}_{\text{max}}}$ and θ and E_{prim}.

The X^{μ}_{max} bias and resolution follows the behavior of its muon-by-muon counterpart ΔX: With respect to E_{prim}, they are almost constant, while $\sigma_{\Delta X^{\mu}_{\text{max}}}$ rises almost like $\sec\theta$. We report improved biases for the QGSJET-II.04 model by almost 20 g cm^{-2}, making $\Delta X^{\mu}_{\text{max}}$ almost unbiased, which also applies for Sibyll-2.3d and less for EPOS-LHC. The resulting values of bias and resolution are summarized in Table 4.3.

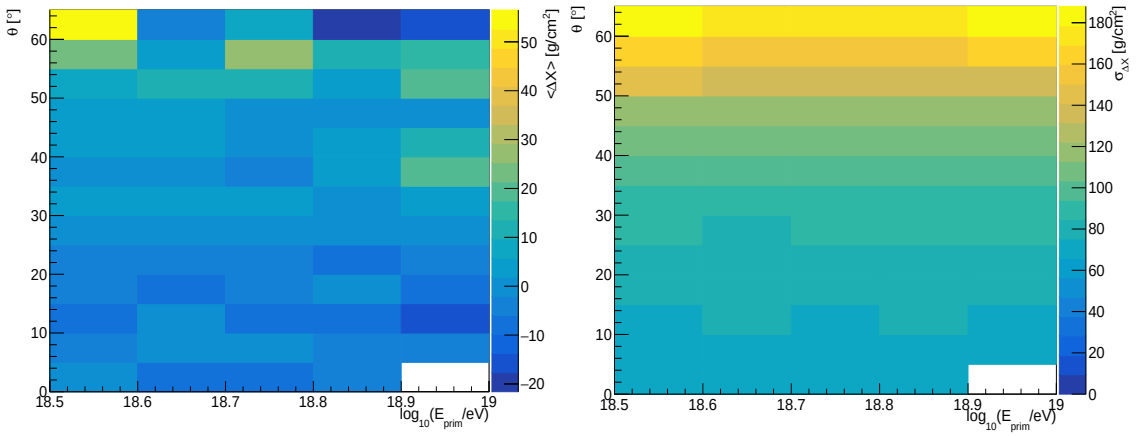

Figure 4.27: The MPD reconstruction bias $\langle \Delta X \rangle$ and resolution $\sigma_{\Delta X}$ as functions of the zenith angle θ and the primary energy, **proton**, $\theta \in (0°, 65°)$, $E_{\text{prim}} \in (10^{18.5}, 10^{19.0})$ eV.

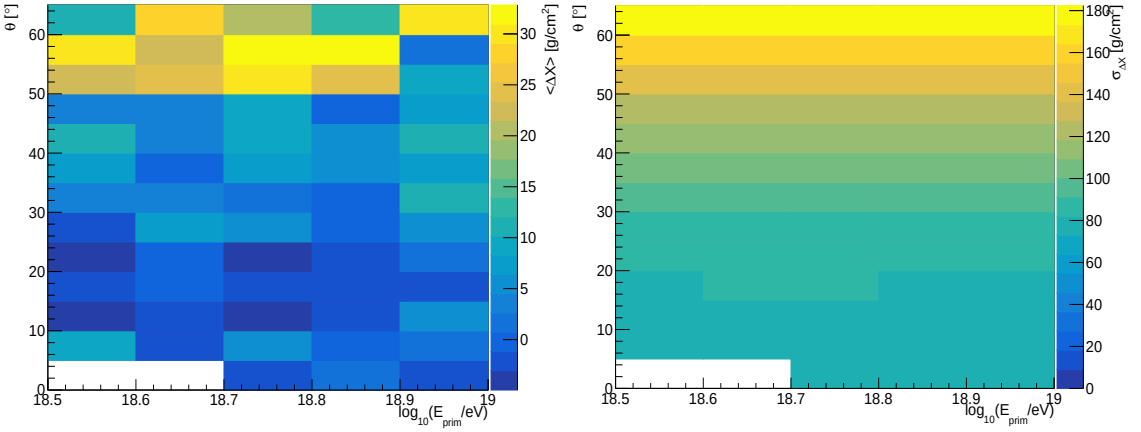

Figure 4.28: The MPD reconstruction bias $\langle \Delta X \rangle$ and resolution $\sigma_{\Delta X}$ as functions of the zenith angle θ and the primary energy, **iron**, $\theta \in (0°, 65°)$, $E_{\text{prim}} \in (10^{18.5}, 10^{19.0})$ eV.

had. model + primary particle	$\langle X^{\mu}_{\text{max}} \rangle$ pred	$\langle X^{\mu}_{\text{max}} \rangle$ true	$\sigma_{X^{\mu}_{\text{max}}}$ pred	$\sigma_{X^{\mu}_{\text{max}}}$ true	$\langle \Delta X^{\mu}_{\text{max}} \rangle$	$\sigma_{\Delta X^{\mu}_{\text{max}}}$
QGSJET-II.04 proton	428	423	68	63	5	31
QGSJET-II.04 iron	356	352	44	31	5	35
Sibyll-2.3d proton	466	467	68	61	-1	29
Sibyll-2.3d iron	385	386	41	28	-1	35
EPOS-LHC proton	477	461	62	57	16	31
EPOS-LHC iron	400	387	36	25	13	30

Table 4.3: Summarized results of the X^{μ}_{max} biases and resolution values for all available data from the dataset with continuous values of θ and E_{prim}.

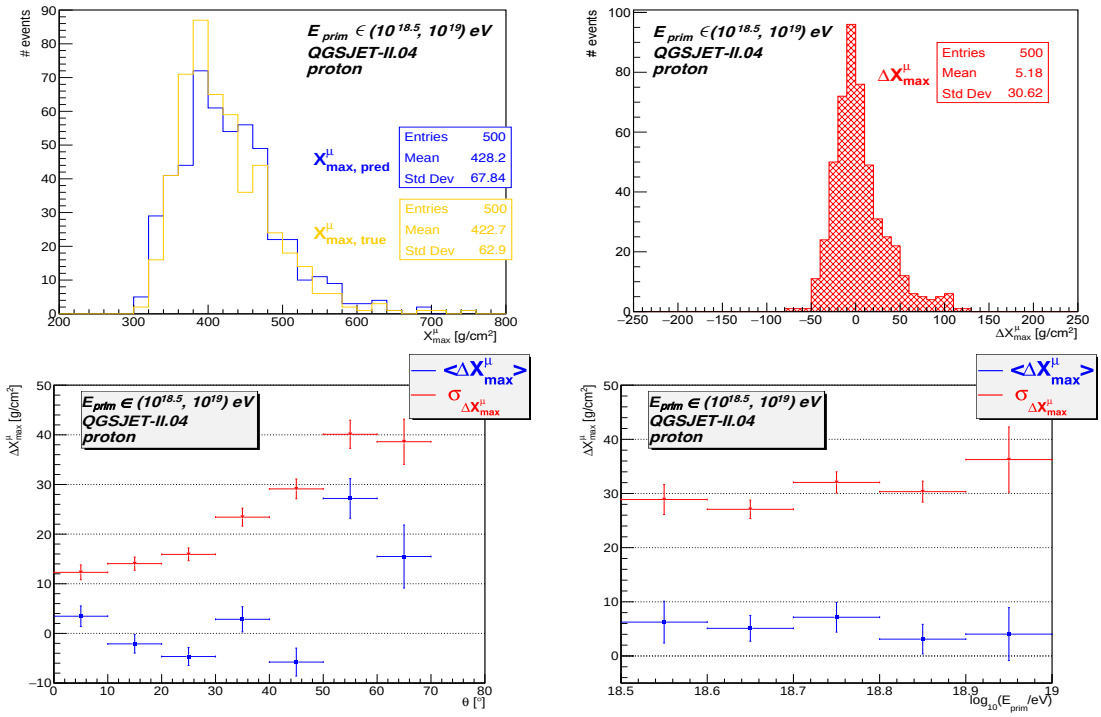

Figure 4.29: **Upper left:** Superimposed $X^{\mu}_{\max}$ distributions. **Upper right:** $\Delta X^{\mu}_{\max}$ distribution. **Lower left:** $\Delta X^{\mu}_{\max}$ characteristics as functions of θ. **Lower right:** $\Delta X^{\mu}_{\max}$ characteristics as functions of E_{prim}. **QGSJET-II.04, proton.**

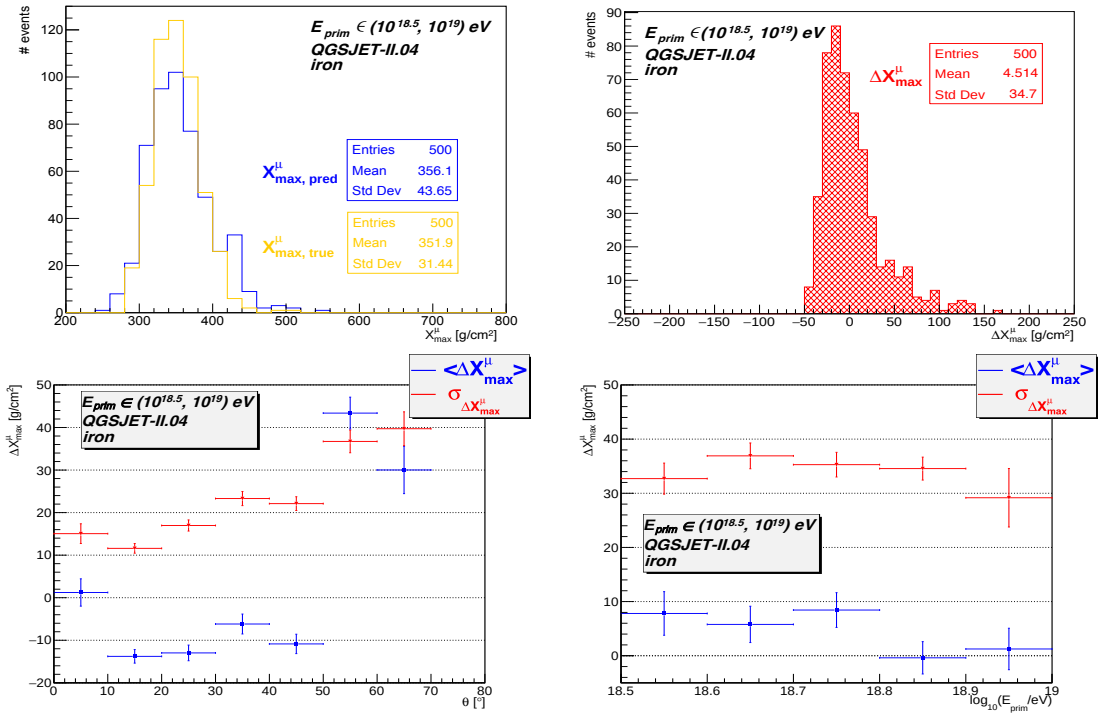

Figure 4.30: **Upper left:** Superimposed $X^{\mu}_{\max}$ distributions. **Upper right:** $\Delta X^{\mu}_{\max}$ distribution. **Lower left:** $\Delta X^{\mu}_{\max}$ characteristics as functions of θ. **Lower right:** $\Delta X^{\mu}_{\max}$ characteristics as functions of E_{prim}. **QGSJET-II.04, iron.**

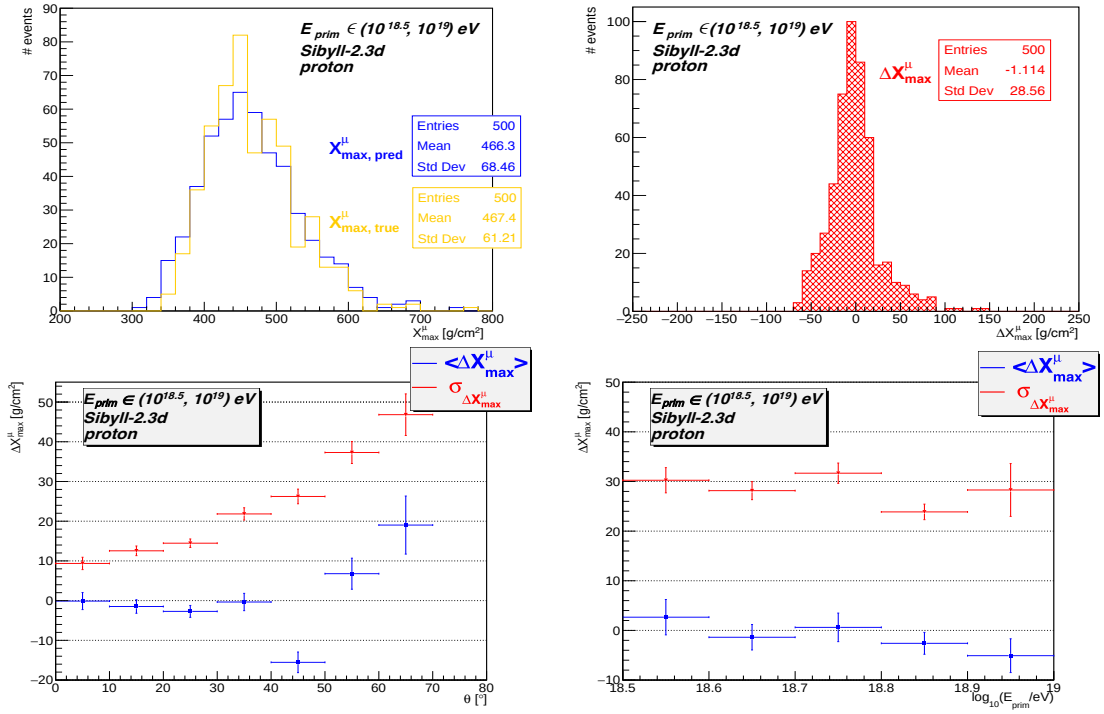

Figure 4.31: **Upper left:** Superimposed $X^{\mu}_{\max}$ distributions. **Upper right:** $\Delta X^{\mu}_{\max}$ distribution. **Lower left:** $\Delta X^{\mu}_{\max}$ characteristics as functions of θ. **Lower right:** $\Delta X^{\mu}_{\max}$ characteristics as functions of E_{prim}. **Sibyll-2.3d, proton.**

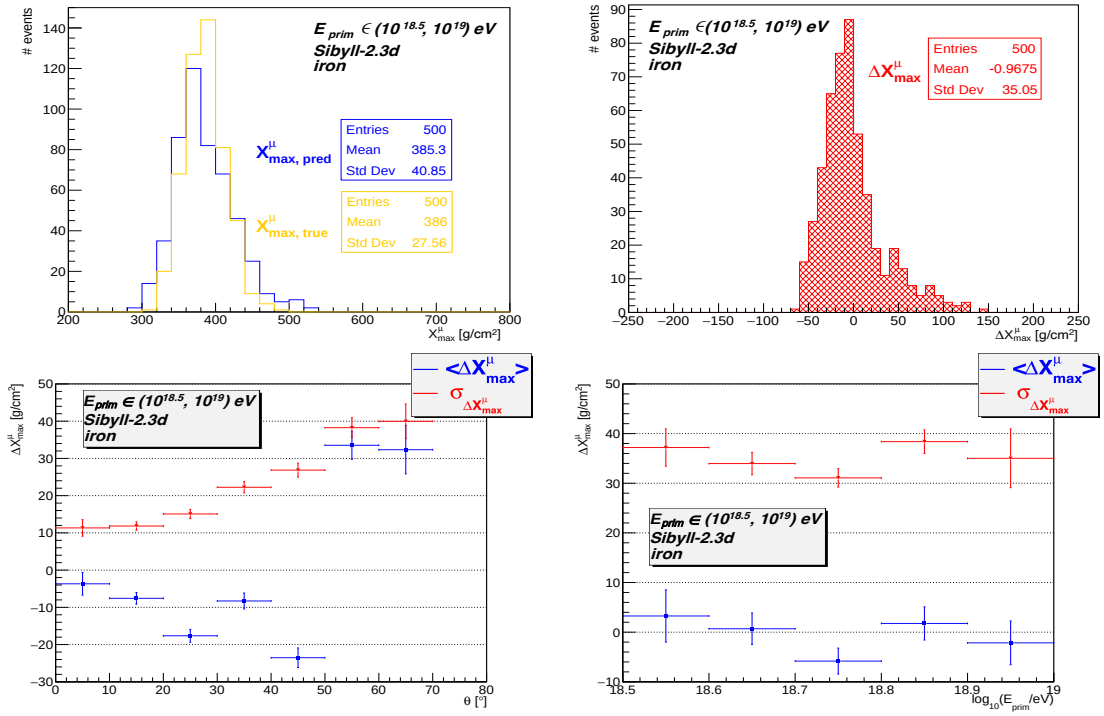

Figure 4.32: **Upper left:** Superimposed $X^{\mu}_{\max}$ distributions. **Upper right:** $\Delta X^{\mu}_{\max}$ distribution. **Lower left:** $\Delta X^{\mu}_{\max}$ characteristics as functions of θ. **Lower right:** $\Delta X^{\mu}_{\max}$ characteristics as functions of E_{prim}. **Sibyll-2.3d, iron.**

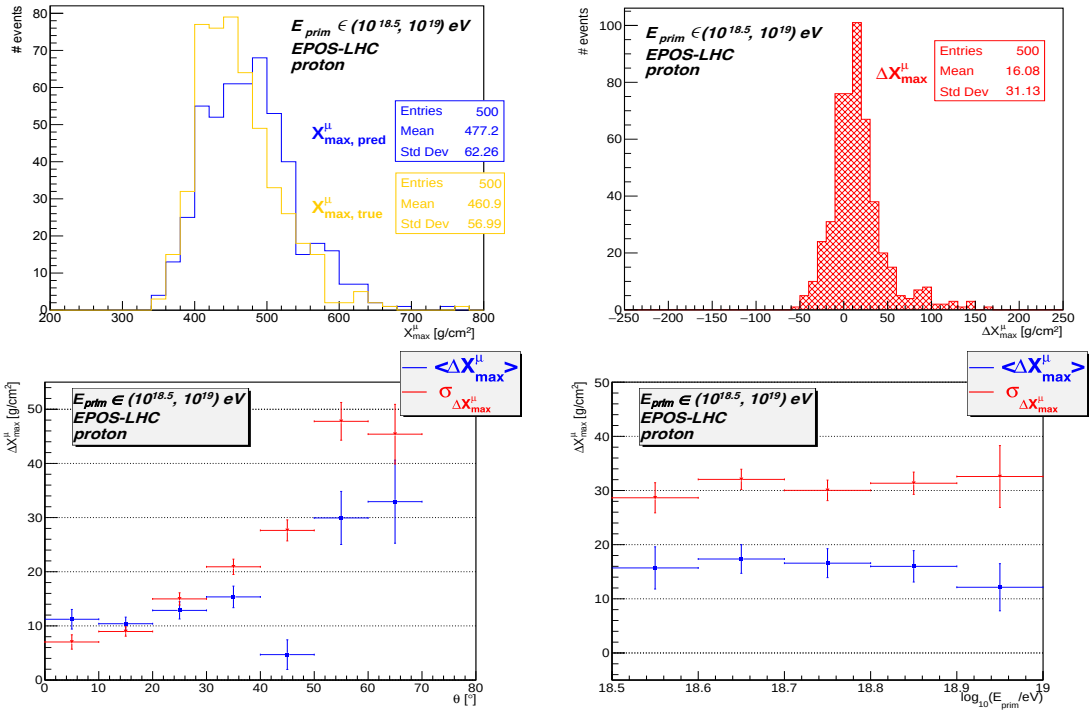

Figure 4.33: **Upper left:** Superimposed $X^{\mu}_{\max}$ distributions. **Upper right:** $\Delta X^{\mu}_{\max}$ distribution. **Lower left:** $\Delta X^{\mu}_{\max}$ characteristics as functions of θ. **Lower right:** $\Delta X^{\mu}_{\max}$ characteristics as functions of E_{prim}. **EPOS-LHC, proton.**

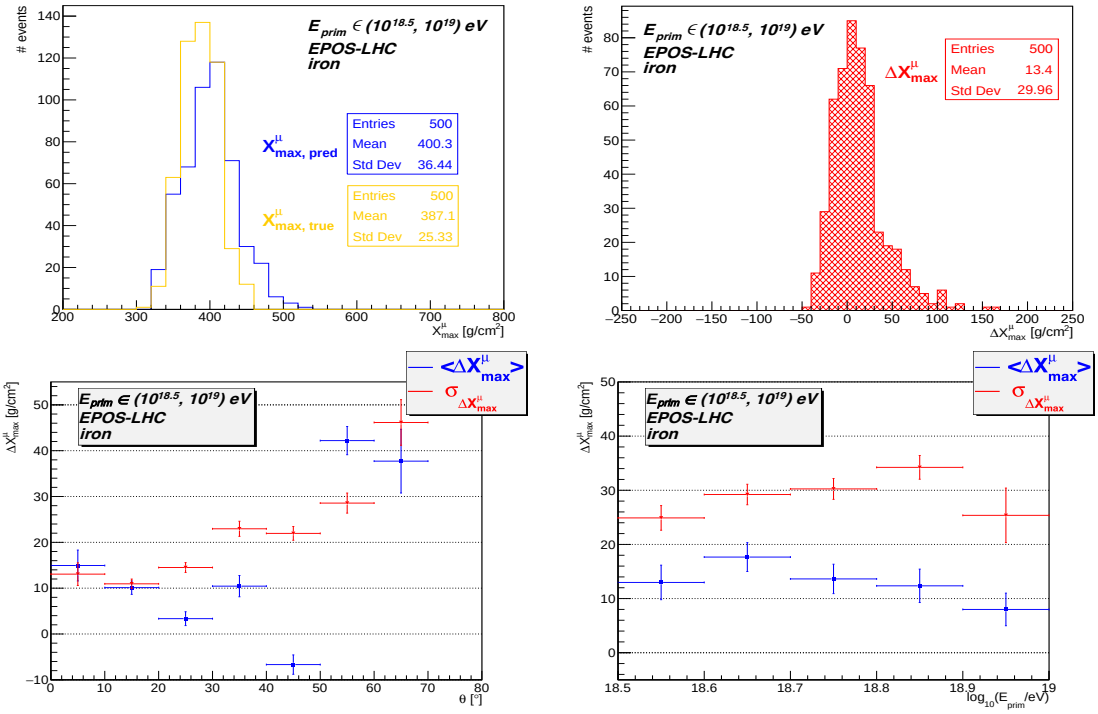

Figure 4.34: **Upper left:** Superimposed $X^{\mu}_{\max}$ distributions. **Upper right:** $\Delta X^{\mu}_{\max}$ distribution. **Lower left:** $\Delta X^{\mu}_{\max}$ characteristics as functions of θ. **Lower right:** $\Delta X^{\mu}_{\max}$ characteristics as functions of E_{prim}. **EPOS-LHC, iron.**

Chapter 5

Muon Energy Reconstruction

In this chapter, we test the predictive power of the previously introduced ML model by attempting to predict the energy of air shower muons. To achieve our goal, we constructed a second ML model, whose target is the muon energy E, with one of the input features being the Muon Production Distance z. We would like to stress that all of our modelings have roots in the original Arrival Time model, which only contains information on the observables at the ground level. However, in this section, we are trying to predict the energies of muons that reach a depth of 2.3 m. Since CORSIKA does not model the particles below the ground, we must approximate the relevant observables underground. From now on, we assume that the geometrical and timing observables, i.e., the distance from the shower core r, the polar angle ξ, and the arrival time t, are unaffected by the underground muon propagation.

In contrast, the propagation underground will affect the muon energy, as described at the beginning of Chapter 4. We proceed as follows: we use the values of our geometrical and timing variables and estimate the muon energy at 2.3 m depth by subtracting from the muon energy given by CORSIKA a given value calculated according to (4.6), assuming a continuous energy loss per unit of traversed matter. As proof of concept, we will present our results for proton-initiated showers using the model of hadronic interactions QGSJET-II.04.

5.1 Data Preparation & Training Performance

Our procedure in creating the energy ML model is similar to the one implemented in Chapter 4. The choice of algorithm is again the Gradient-Boosted Decision Trees, utilized through the LightGBM library. All values of hyperparameters and new features were again found via the Hyperopt library, with a change in the implemented objective function. Now, as we only care about the precision of the reconstructed energy of individual muons, we define the objective function to be

$$f_O = MSE(E_{pred}, E_{true}), \tag{5.1}$$

where the superscripts *pred* and *true* represent the target values predicted by the model and the values of the target itself. Table 5.1 summarizes the searching range

and the optimized values of hyperparameters:

Hyperparameter	*Tuning Range*	*Optimal Value*
Learning Rate	[0.001, 1]	0.393
Number of Leaves	{2, 3, ..., 1000}	308
Maximal Depth of Tree	{1, 2, ..., 20}	9
Minimal Child Weight	{0, 1, ..., 5000}	85
L1 Regularization	[0, 100]	21.2
L2 Regularization	[0, 100]	81.1
Colsample by Tree	[0, 1]	0.443
Subsample	[0, 1]	0.570

Table 5.1: A summary of the optimal values of hyperparameters (of the energy-reconstructing ML model) found by Hyperopt, alongside the respective search ranges.

Our basic set of features consists of $\sec\theta$, $\cos\xi$, r and t, this time supplemented by the Muon Production Distance z. We supply the true CORSIKA values of z into the training data and when we predict the muon energy on test datasets, we use the MPD ML model from the previous chapter to provide the z values. Additionally, the following set of new features was created by Hyperopt for the energy model to yield better results: $\left\{\frac{r^2}{(ct)^2}, \frac{\sqrt{r}}{\sqrt{ct}}, \frac{1}{t}, \frac{\log_{10}(ct)}{r}\right\}$.

The optimal target form was found to be $\log_{10}(E_{ground})$. As discussed above, the final muon energy E at the 2.3 m depth is calculated by subtracting the assumed energy losses throughout the muon propagation in the soil from E_{ground}, according to (4.6). We train the model on 450 proton- and 450 iron-initiated thinned showers with the energy of the primary cosmic ray of 10^{17} eV and continuous zenith angle values, reserving additional 50 showers of each primary as the validation set. 200 proton- and 200 iron-initiated showers with the same characteristics are then used as the test dataset. Since we are interested in the muon-by-muon energy precision, we do not include the undersampling method in the final data pre-processing.

The loss function in model training was again chosen as the MSE. During training, we implemented the early-stopping algorithm, which stopped the training after 43 iterations, recording the train MSE value of **0.296** and validation MSE value of **0.297**[1].

5.2 Energy Reconstruction Results

Our results for a sample from the zenith library ($\{0°, 12°, 25°, 35°, 45°, 60°\}$) are displayed in Figures 5.1 and 5.2. In Fig. 5.1, we show the distributions for:

[1]The train and validation MSE values are almost identical, since we did not use data undersampling - the reader can compare these values with the corresponding values in the previous chapter.

- True muon energy given by CORSIKA E_{true}
- Muon energy predicted by the ML model $E_{ML}^{z=true}$, using true CORSIKA values of z
- Muon energy predicted by the ML model $E_{ML}^{z=pred}$, using predicted values of z (by the first ML model from the previous chapter).

We show the full muon spectra, alongside additional figures focusing on the low-energy part of the respective spectra, where most of the muons land.

First, it can be seen that the "true" and "ML" distributions match well for low-to-middle energies (3-20 GeV). The behavior at the lowest and high energies depends on the zenith angle: As the zenith angle increases, the reconstruction quality decreases for the lowest-energy muons with $E < 2$ GeV and increases for high-energy muons with $E > 20$ GeV. The reason for this might be related to emergence of more energetic muons as zenith angle increases, which broadens the muon energy spectrum. At high zenith angles, there is enough statistics (relative to the shower population) for the model to learn about the high-energy muons and thus reconstruct their energy more easily. Another important factor is that, for inclined showers, i.e., $\theta \sim 60°$, the dominant contribution comes from the geometric delay of muons, which can be almost precisely estimated.

By taking a look at the characteristics of the muon-by-muon energy reconstruction in Fig. 5.2, the mean value and the standard deviation of the difference $\Delta E = E_{ML}^{z=pred}$ - E_{true}, we see that the overall reconstruction is relatively unbiased for all zenith angles, with $\sigma_{\Delta E}$ rising with increasing zenith angle, as expected. On the other hand, $\sigma_{\Delta E}$ accounts to less than 4% of the given energy spectrum range, which is a good sign for this model in terms of its predictive power.

Lastly, we look at how the energy reconstruction bias $\langle \Delta E \rangle$ and resolution $\sigma_{\Delta E}$ evolve with the observables r, t, and E. These dependencies are depicted in Figures 5.3, 5.4 and 5.5, respectively. For the relations between the moments and r and t, the situation is similar and by now well known. Generally, muons impacting far from the shower core and/or late in the shower tend to be less biased and their reconstruction varies less from the respective mean value. As we increase the zenith angle, the quality of the reconstruction worsens. For example, for the model to achieve the same performance for showers with $\theta \sim 60°$ as it is with $\theta \sim 0°$, we would need to consider muons at least ~ 600 m from the shower core and/or cut all muons arriving earlier than 50 ns after the shower core particles.

Regarding the dependence on muon energies, Figure 5.5 confirms what was said above: Low energy muons are reconstructed better than high energy muons, while there is an almost linear dependency between the worsening performance of the ML model and the muon energy. As the zenith angle increases, more high energetic muons are present in the training and the reconstruction for the same muon energies gets better. All-in-all, since low-energy muons dominate the energy spectrum, we can consider the model to be in a relative agreement to a large portion of the data and a hopeful proof of concept for muon energy reconstruction.

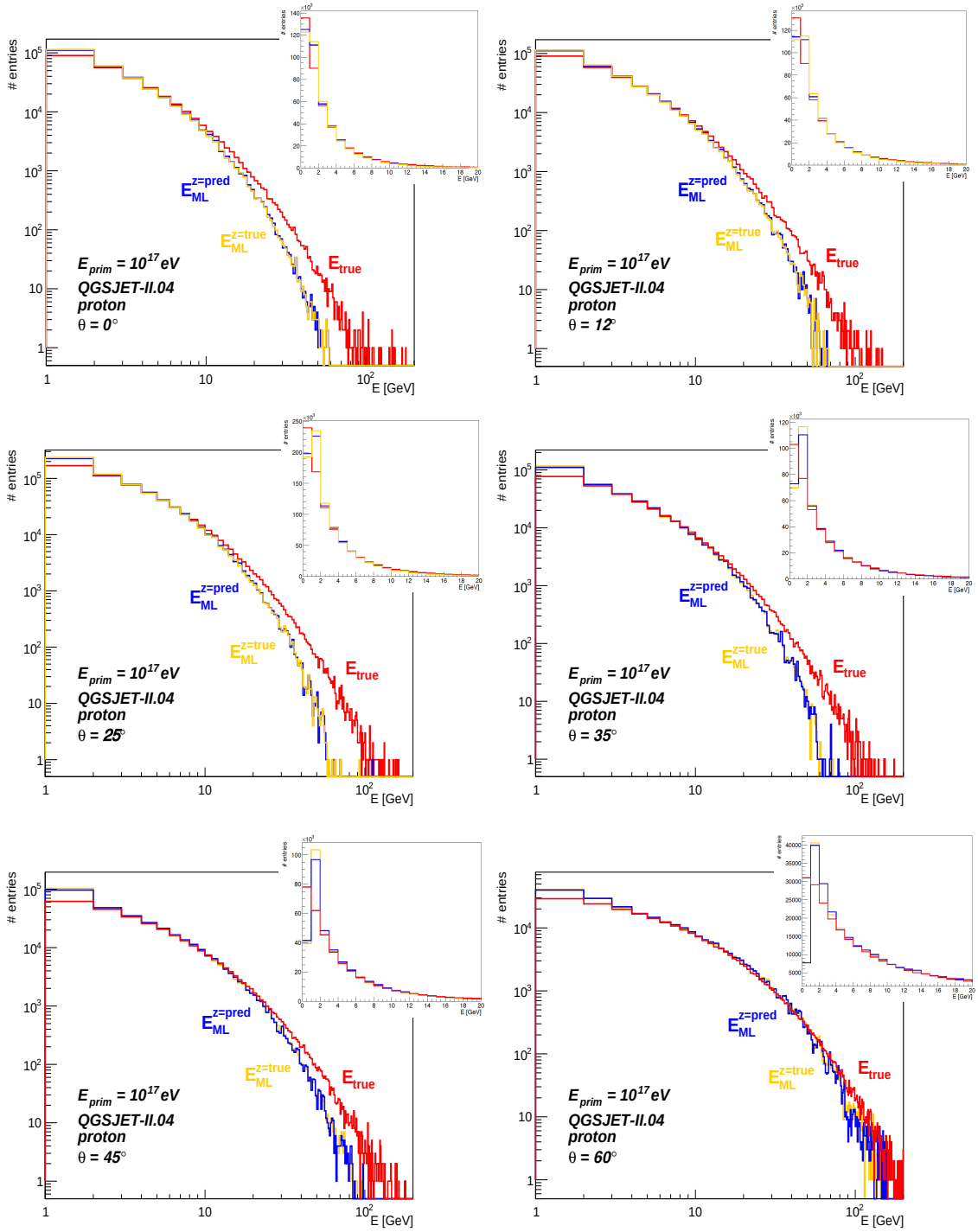

Figure 5.1: Superimposed E_{true}, $E_{ML}^{z=true}$ and $E_{ML}^{z=pred}$ distributions for a sample from the continuous zenith library. Both the whole muon energy spectrum and a close-up to first energy bins are depicted.

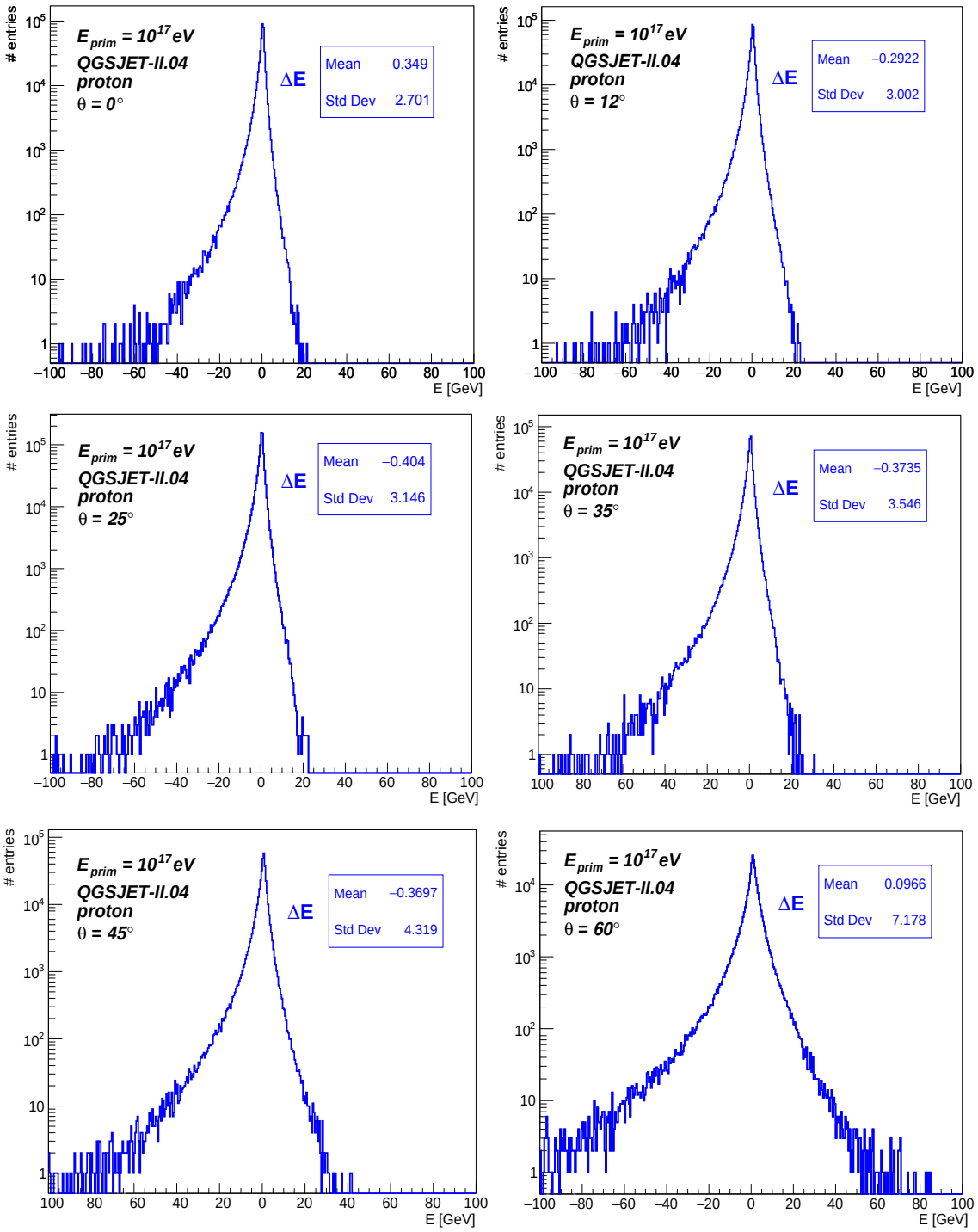

Figure 5.2: Distributions of the ΔE difference for a sample from the continuous zenith library.

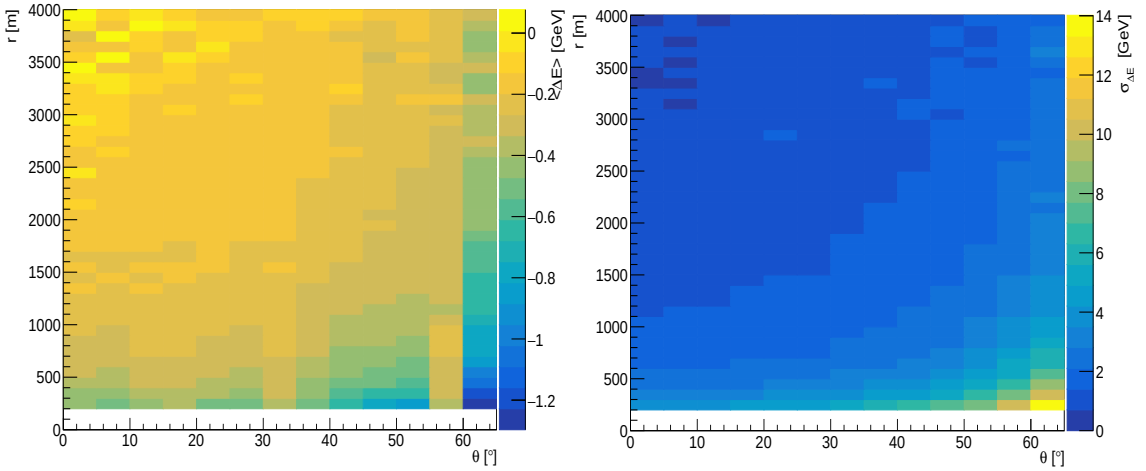

Figure 5.3: The energy reconstruction characteristics $\langle \Delta E \rangle$ and $\sigma_{\Delta E}$ as functions of the zenith angle θ and the distance from the shower core r.

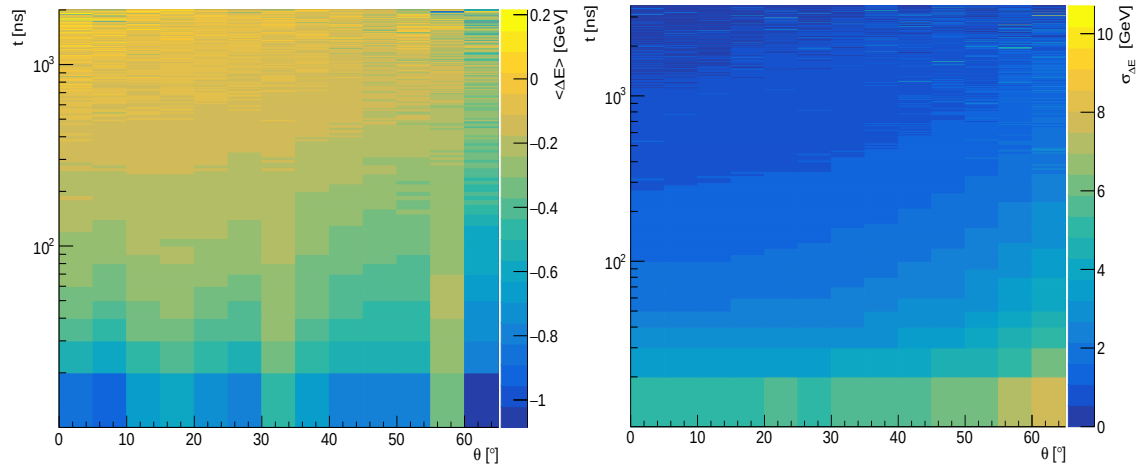

Figure 5.4: The energy reconstruction characteristics $\langle \Delta E \rangle$ and $\sigma_{\Delta E}$ as functions of the zenith angle θ and the arrival time t.

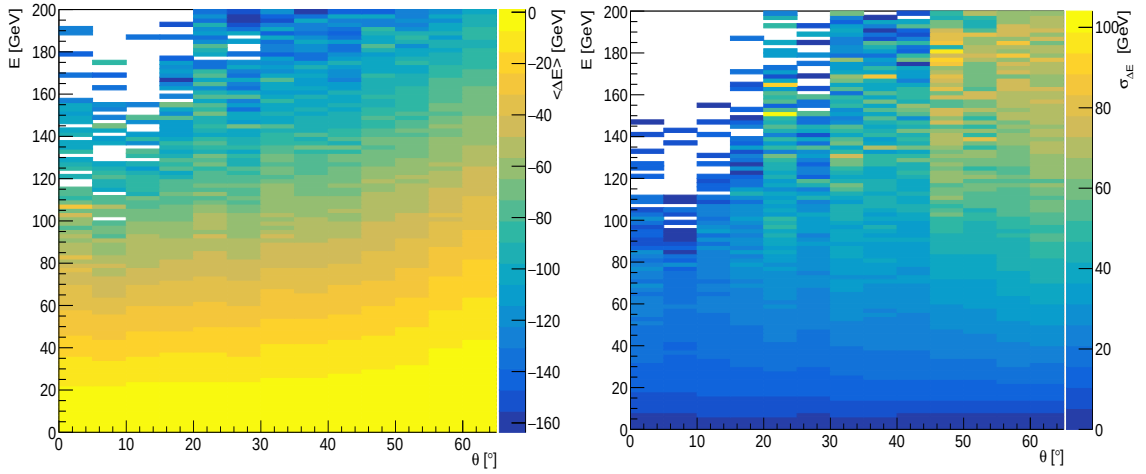

Figure 5.5: The energy reconstruction characteristics $\langle \Delta E \rangle$ and $\sigma_{\Delta E}$ as functions of the zenith angle θ and the muon energy E.

Conclusions

In this work, a new model of reconstructing the Muon Production Depth, based on the Gradient-Boosted Decision Trees algorithm, was introduced. With the aim of improving the current reconstruction method described in [6, 7], the proposed machine learning model was designed and analyzed with the use of Monte Carlo simulations of Extensive Air Showers. The desired areas of improvement included a drastic lowering of the applied radial cut and an application of the method to lower zenith angles, while boosting the overall reconstruction quality. With these improvements in mind, the proposed model was designed to reconstruct the MPD of muons arriving at $r \geq 200\,\mathrm{m}$ within EAS propagating under a wide range of zenith angles $\theta \in (0°, 65°)$. In order to simulate a more realistic setup of buried detectors measuring a pure muonic signal, we imposed an energy cut to all muons, with the form $E_{th} \simeq 1\,\mathrm{GeV}/\cos\theta$, and discarded all muons not fulfilling this condition. The minimum vertical energy of 1 GeV is derived from the soil density and the depth of the AMIGA muon detectors at the Pierre Auger Observatory. The newly introduced model of MPD reconstruction was applied to EAS simulations, assuming different hadronic interaction models, types of primary particles, energies, and zenith angles, allowing us to make relevant conclusions about the method's performance and generalization capabilities. As a proof of concept, a second machine learning model was built to reconstruct the energies of muons in EAS, utilizing the predictions of the MPD model in the training process. We summarize the performance results of the models below.

In the first part of the analysis, we explained our reasoning behind the design of the proposed MPD model, which was based on a combination of the Arrival Time Model structure and standard machine learning procedures. We argued that for the model to reconstruct the MPD correctly, both muon-by-muon and the MPD distributions, we should tune the model's parameters to minimize the objective function in the form of (4.7). This way, we prevented the model from overpredicting values close to the mean value of the MPD distributions, which is a common problem for many machine learning models and which also happened multiple times in our attempts to design a competitive model. We also created new features for our model through feature engineering, which helped to boost the model's performance. Lastly, we performed a specific undersampling of our data to mitigate fundamental differences between EAS induced by various cosmic rays.

The second part of our study concerned the proposed model's performance on various EAS samples. At the lowest energies, $E_{\mathrm{prim}} = 10^{17}\,\mathrm{eV}$, we observed that the reconstruction performed well for zenith angles below 50°. We recorded an almost

unbiased muon-by-muon reconstruction of $\left(|\langle\Delta X\rangle| < 15\,\mathrm{g\,cm^{-2}}\right)$. Regarding the reconstruction of the mass-composition sensitive observable X^{μ}_{max}, we observed a reconstruction bias of $|\langle\Delta X^{\mu}_{\mathrm{max}}\rangle| \lesssim 30\,\mathrm{g\,cm^{-2}}$. On the other hand, the reconstruction quality generally decreases for $\theta > 50°$, both muon-by-muon and distribution-wise. However, if we compare our proposed model to the current MPD method, we observe a clear upgrade in the reconstruction capability. Additionally, as expected, an improvement was recorded in the reconstruction performance when discarding muons close to the shower core. The dependence of the model's behavior with respect to relevant muon observables, the distance from the shower core r, the arrival time t and the muon energy E, was also investigated, revealing that the MPD reconstruction improves, with some exceptions, for higher values of the aforementioned (r, t, E) parameters. The energy dependence, however, shows that the MPD model can reasonably well reconstruct only the most-represented energies in the muon spectrum, a clear consequence of the muon energy not being a feature of the MPD model. Additionally, while the MPD model performed overall evenly for all models of hadronic interactions, it generally performed worse on the MPD reconstruction of iron-induced showers. On the other hand, the reconstructed proton and iron X^{μ}_{max} distributions seemed clearly distinguishable from each other and were satisfactorily reconstructed to allow for mass-composition analyses. At higher cosmic-ray energies $E_{\mathrm{prim}} \in (10^{18.5}, 10^{19})$ eV, while the muon-by-muon reconstruction performed similarly to the one for $E_{\mathrm{prim}} = 10^{17}$ eV, the overall performance of the reconstruction of the X^{μ}_{max} distributions was better than in the previous case. Also, we found that the reconstruction of $\langle\Delta X^{\mu}_{\mathrm{max}}\rangle$ was unbiased for the QGSJET-II.04 and Sibyll 2.3d models, while a small bias of $\sim 15\,\mathrm{g\,cm^2}$ was recorded for EPOS-LHC. The dependences on zenith angles copy the ones at the lower cosmic ray energy, while the X^{μ}_{max} reconstruction characteristics register an almost constant dependence on E_{prim} for both proton- and iron-induced EAS and all models of hadronic interactions. This is a promising result, which will be built upon in the following studies.

Lastly, a second model, designed to reconstruct the energy of muons in EAS was introduced in Chapter 5. Also based on the GBDT algorithm, it was trained using the Muon Production Distance as one of its features. We found that the best-reconstructed energies were in the range of roughly 3-20 GeV, while the behavior at the lowest and high muon energies varied with zenith angle. The reconstruction, though, is virtually unbiased ($|\langle\Delta E\rangle| < 0.4$ GeV), while its resolution $\sigma_{\Delta E}$ rises with rising zenith angles.

All-in-all, both models represent proofs that machine learning algorithms can be beneficial in the MPD field. That being said, there is still considerable room for improving the proposed MPD reconstruction model in this regard. Specifically, in the data undersampling part of the process, there are other methods than Random Undersampling, more computationally expensive but also more powerful. There is also a possibility of using data undersampling for the energy reconstruction model, which will be a topic of the future works. Considering this, the proposed MPD reconstruction model is ready, with the necessary modifications, to be implemented in further studies, including an application of the model to advanced simulations with simulated detector responses and, subsequently, real data from buried particle detectors. This will allow us to gain additional insight into the mass composition of

cosmic rays and, with the help of the muon energy reconstruction, will let us understand the muon energy spectrum in EAS better. Additionally, by reconstructing the MPD for a large portion of muons in an EAS, we might be able to refine the current models of hadronic interactions, which will allow us to learn relevant information about the hadronic interactions governing the EAS evolution.

Bibliography

[1] D. J. Bird *et al.*, "Detection of a cosmic ray with measured energy well beyond the expected spectral cutoff due to cosmic microwave radiation," *The Astrophysical Journal*, vol. 441, p. 144, 1995. DOI: 10.1086/175344.

[2] A. Aab *et al.*, "The Pierre Auger Cosmic Ray Observatory," *Nucl. Instrum. Meth. A*, vol. 798, pp. 172–213, 2015. DOI: 10.1016/j.nima.2015.06.058.

[3] H. Tokuno *et al.*, "The telescope array experiment: Status and prospects," *Journal of Physics: Conference Series*, vol. 120, p. 062 027, 2008. DOI: 10.1088/1742-6596/120/6/062027.

[4] A. Aab *et al.*, "Inferences on mass composition and tests of hadronic interactions from 0.3 to 100 eev using the water-cherenkov detectors of the pierre auger observatory," *Phys. Rev. D*, vol. 96, p. 122 003, 12 2017. DOI: 10.1103/PhysRevD.96.122003.

[5] A. Aab *et al.*, "Deep-learning based reconstruction of the shower maximum X_{max} using the water-cherenkov detectors of the pierre auger observatory," *Journal of Instrumentation*, vol. 16, no. 07, P07019, 2021. DOI: 10.1088/1748-0221/16/07/p07019.

[6] L. Cazón, R. Vázquez, A. Watson, and E. Zas, "Time structure of muonic showers," *Astroparticle Physics*, vol. 21, no. 1, pp. 71–86, 2004. DOI: 10.1016/j.astropartphys.2003.12.009.

[7] L. Cazón, R. Vázquez, and E. Zas, "Depth development of extensive air showers from muon time distributions," *Astroparticle Physics*, vol. 23, no. 4, pp. 393–409, 2005. DOI: 10.1016/j.astropartphys.2005.01.009.

[8] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. DOI: 10.1214/aos/1013203451.

[9] E. G. Zweibel, "The microphysics and macrophysics of cosmic rays," *Physics of Plasmas*, vol. 20, no. 5, p. 055 501, 2013. DOI: 10.1063/1.4807033.

[10] B. Peters, "Primary cosmic radiation and extensive air showers," *Il Nuovo Cimento*, vol. 22, no. 4, pp. 800–819, 1961. DOI: 10.1007/BF02783106.

[11] C. Grupen, G. Cowan, S. Eidelman, and T. Stroh, *Astroparticle Physics*. Springer Berlin Heidelberg, 2010.

[12] K. Greisen, "End to the cosmic-ray spectrum?" *Phys. Rev. Lett.*, vol. 16, pp. 748–750, 17 1966. DOI: 10.1103/PhysRevLett.16.748.

[13] G. T. Zatsepin and V. A. Kuz'min, "Upper Limit of the Spectrum of Cosmic Rays," *Soviet Journal of Experimental and Theoretical Physics Letters*, vol. 4, p. 78, 1966.

[14] J. Abraham *et al.*, "Measurement of the Energy Spectrum of Cosmic Rays above 10^{18} eV Using the Pierre Auger Observatory," *Phys. Lett. B*, vol. 685, pp. 239–246, 2010. DOI: 10.1016/j.physletb.2010.02.013.

[15] R. Aloisio, V. Berezinsky, and A. Gazizov, "Disappointing model for ultrahigh-energy cosmic rays," *J. Phys. Conf. Ser.*, vol. 337, V. N. Auerbach, M. Hass, and M. Paul, Eds., p. 012 042, 2012. DOI: 10.1088/1742-6596/337/1/012042.

[16] A. Haungs *et al.*, "The KASCADE cosmic-ray data centre KCDC: Granting open access to astroparticle physics research data," *The European Physical Journal C*, vol. 78, no. 9, 2018. DOI: 10.1140/epjc/s10052-018-6221-2.

[17] J. Matthews, "A Heitler model of extensive air showers," *Astropart. Phys.*, vol. 22, pp. 387–397, 2005. DOI: 10.1016/j.astropartphys.2004.09.003.

[18] K.-H. Kampert and M. Unger, "Measurements of the cosmic ray composition with air shower experiments," *Astroparticle Physics*, vol. 35, no. 10, pp. 660–678, 2012. DOI: 10.1016/j.astropartphys.2012.02.004.

[19] L. A. Anchordoqui, "Ultra-high-energy cosmic rays," *Physics Reports*, vol. 801, pp. 1–93, 2019. DOI: 10.1016/j.physrep.2019.01.002.

[20] W. Heitler, *The Quantum Theory of Radiation*. Oxford University Press, London, 1954, p. 386.

[21] S. Cecchini and M. Sioli, "Cosmic ray muon physics," 2000. arXiv: hep-ex/0002052 [hep-ex].

[22] D. E. Groom, N. V. Mokhov, and S. I. Striganov, "Muon stopping power and range tables 10 mev–100 tev," *Atomic Data and Nuclear Data Tables*, vol. 78, no. 2, pp. 183–356, 2001. DOI: https://doi.org/10.1006/adnd.2001.0861.

[23] P. Zyla *et al.*, "Review of Particle Physics," *PTEP*, vol. 2020, no. 8, p. 083C01, 2020. DOI: 10.1093/ptep/ptaa104.

[24] T. Stanev, *High Energy Cosmic Rays* (Springer Praxis Books). Springer Berlin Heidelberg, 2010.

[25] P. A. Collaboration and A. Etchegoyen, "Amiga, auger muons and infill for the ground array," 2007. arXiv: 0710.1646 [astro-ph].

[26] N. Chiba *et al.*, "Akeno giant air shower array (agasa) covering 100 km2 area," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 311, no. 1, pp. 338–349, 1992. DOI: https://doi.org/10.1016/0168-9002(92)90882-5.

[27] H. Dembinski, R. Engel, A. Fedynitch, T. Gaisser, F. Riehn, and T. Stanev, "Data-driven model of the cosmic-ray flux and mass composition from 10 gev to 10^{11} gev," 2017. arXiv: 1711.11432 [astro-ph.HE].

[28] A. Coleman *et al.*, "Ultra high energy cosmic rays the intersection of the cosmic and energy frontiers," *Astroparticle Physics*, vol. 149, p. 102 819, 2023. DOI: 10.1016/j.astropartphys.2023.102819.

[29] T. Pierog, I. Karpenko, J. M. Katzy, E. Yatsenko, and K. Werner, "Epos lhc: Test of collective hadronization with data measured at the cern large hadron collider," *Physical Review C*, vol. 92, no. 3, 2015. DOI: 10.1103/physrevc.92.034906.

[30] S. Ostapchenko, "Monte carlo treatment of hadronic interactions in enhanced pomeron scheme: Qgsjet-ii model," *Phys. Rev. D*, vol. 83, p. 014018, 1 2011. DOI: 10.1103/PhysRevD.83.014018.

[31] F. Riehn, R. Engel, A. Fedynitch, T. K. Gaisser, and T. Stanev, "Hadronic interaction model sibyll 2.3d and extensive air showers," *Physical Review D*, vol. 102, no. 6, 2020. DOI: 10.1103/physrevd.102.063002.

[32] D. Heck, J. Knapp, J. N. Capdevielle, G. Schatz, and T. Thouw, "CORSIKA: A Monte Carlo code to simulate extensive air showers," 1998.

[33] L. Cazón, "Modelling the muon time distribution in extensive air showers," Ph.D. dissertation, 2004.

[34] S. Andringa, L. Cazon, R. Conceição, and M. Pimenta, "The muonic longitudinal shower profiles at production," *Astroparticle Physics*, vol. 35, no. 12, pp. 821–827, 2012. DOI: https://doi.org/10.1016/j.astropartphys.2012.03.010.

[35] L. Cazon, R. Conceição, M. Pimenta, and E. Santos, "A model for the transport of muons in extensive air showers," *Astroparticle Physics*, vol. 36, no. 1, pp. 211–223, 2012. DOI: 10.1016/j.astropartphys.2012.05.017.

[36] M. Ave, R. Vázquez, and E. Zas, "Modeling horizontal air showers induced by cosmic rays," *Astroparticle Physics*, vol. 14, no. 2, pp. 91–107, 2000. DOI: 10.1016/s0927-6505(00)00113-4.

[37] A. Aab *et al.*, "Muons in air showers at the pierre auger observatory: Measurement of atmospheric production depth," *Physical Review D*, vol. 90, no. 1, 2014. DOI: 10.1103/physrevd.90.012012.

[38] A. Aab *et al.*, "Extraction of the muon signals recorded with the surface detector of the pierre auger observatory using recurrent neural networks," *Journal of Instrumentation*, vol. 16, no. 07, P07016, 2021. DOI: 10.1088/1748-0221/16/07/P07016.

[39] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*. O'Reilly Media, Incorporated, 2019.

[40] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," 2022. arXiv: 2207.08815 [cs.LG].

[41] Hshan.T. "Bias-variance trade off from learning curve." (2020), [Online]. Available: https://hshan0103.medium.com/understanding-bias-variance-trade-off-from-learning-curve-a64b4223bb02.

[42] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016. DOI: 10.1145/2939672.2939785.

[43] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[44] A. Navlani. "Decision tree classification in python tutorial." (2023), [Online]. Available: https://www.datacamp.com/tutorial/decision-tree-classification-python.

[45] S. Jansen, *Hands-On Machine Learning for Algorithmic Trading: Design and Implement Investment Strategies Based on Smart Algorithms That Learn from Data Using Python*. Packt Publishing, 2018.

[46] G. Battistoni *et al.*, "Overview of the fluka code," *Annals of Nuclear Energy*, vol. 82, pp. 10–18, 2015, Joint International Conference on Supercomputing in Nuclear Applications and Monte Carlo 2013, SNA + MC 2013. Pluri- and Trans-disciplinarity, Towards New Modeling and Numerical Simulation Paradigms. DOI: https://doi.org/10.1016/j.anucene.2014.11.007.

[47] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, "Hyperopt: A python library for model selection and hyperparameter optimization," *Computational Science & Discovery*, vol. 8, no. 1, p. 014 008, 2015. DOI: 10.1088/1749-4699/8/1/014008.

[48] M. Filho. "Can gradient boosting learn simple arithmetic?" (2020), [Online]. Available: https://forecastegy.com/posts/can-gradient-boosting-learn-simple-arithmetic/.

[49] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[50] T. K. Gaisser and A. M. Hillas, "Reliability of the Method of Constant Intensity Cuts for Reconstructing the Average Development of Vertical Showers," in *International Cosmic Ray Conference*, ser. International Cosmic Ray Conference, vol. 8, 1977, p. 353.