

**DETERMINATION OF PHYSICAL PROPERTIES OF
HIGH-ENERGY HADRONIC INTERACTIONS FROM
THE X_{max} - N_{μ} ANTCORRELATION**

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)
und der
Universidad Nacional de San Martín (UNSAM)

angenommene
DISSERTATION

von

Mgr. Isabel Astrid Goos

Referent: **Dr. X. Bertou**

Korreferent: **Prof. Dr. J. Blümer**

Betreuer: **Dr. T. Pierog**

Tag der mündlichen Prüfung: **11. Februar 2022**

A todas las personas que
me acompañaron en esta
etapa

Contents

Contents	v
Figure Index	vii
Table Index	xi
Abstract	xiii
Resumen	xv
Zusammenfassung	xvii
Symbol Index	xiii
1. Cosmic rays	1
1.1. Historical development	1
1.2. Applications to astrophysics	5
1.2.1. Energy spectrum	5
1.2.2. Origin of cosmic rays	8
1.2.3. Anisotropy	9
1.2.4. Mass composition	10
1.3. Application to particle physics	11
2. The Pierre Auger Observatory	13
2.1. The surface detector	14
2.2. The fluorescence detector	16
2.3. AugerPrime	18
2.3.1. The Surface scintillator detector	18
2.3.2. The AMIGA muon detector	19
3. Air shower physics	21
3.1. Extensive air showers	21
3.2. Objective of this work	23

3.3. Monte Carlo simulations	25
3.4. Semi-empirical model	28
3.4.1. Electromagnetic showers	28
3.4.2. Proton initiated showers	30
3.4.3. Hadronic showers	40
3.4.4. Leading particle effect	46
3.5. The Muon Puzzle	47
3.6. Fluctuations of the muon content	53
4. Understanding the $X_{max} - N_{\mu}$ anticorrelation	57
4.1. Initial approach	57
4.2. Improvement of the semi-empirical model	60
5. Construction of the model	71
5.1. Neural network architecture	71
5.2. Feature selection	83
5.3. Description of the features	90
5.4. Performance of the model	97
6. Application to Auger data	103
6.1. Dataset	103
6.2. Mixed composition	105
7. Conclusions	113
Bibliography	117
Acknowledgments	131

Figure Index

1.1. Discovery of cosmic rays and the positron	2
1.2. Simplified sketch of an extensive air shower	3
1.3. Cosmic ray differential energy spectrum	6
1.4. Hillas and energy budget diagrams	7
1.5. Large-scale anisotropy	9
1.6. Mass composition	10
1.7. Mean and standard deviation of the X_{max} distribution for composition studies	11
1.8. Measurement of the proton-air inelastic cross-section	12
2.1. Pierre Auger Observatory	13
2.2. Water-Cherenkov detector	14
2.3. Reconstruction of shower geometry and primary energy with the SD . .	15
2.4. Fluorescence detector	16
2.5. Reconstruction of shower geometry and primary energy with the FD . .	17
2.6. Surface scintillator detector	19
2.7. AMIGA muon detector	20
3.1. Longitudinal profiles of the different components of an extensive air shower	22
3.2. X_{max} - N_μ anticorrelation for different high-energy interaction models . .	24
3.3. X_{max} - N_μ anticorrelation for different primary masses	25
3.4. Scheme of the Heitler splitting model	28
3.5. Scheme of the model for a proton initiated shower	30
3.6. Fractions of the principal groups of particles	32
3.7. Distributions of the geometric mean of N_0 and N_{ch} for proton primaries	33
3.8. Mean charged particle multiplicity in p-air and π -air collisions	34
3.9. Correlation between the geometric means of N_0 and N_{ch}	35
3.10. Distributions of the factor c_0^{ch} for proton primaries	35
3.11. Distributions of the depth of first interaction for proton primaries and inelastic cross-sections for p-Air and π -Air as a function of energy . . .	36
3.12. Critical energy as calculated with the approaches from [1] and [2] . . .	38

3.13. Distributions of the geometric mean of the fractions $f_{ch,en}$ for proton primaries	39
3.14. Distributions of the depth of first interaction for iron primaries	40
3.15. Distributions of the geometric mean of N_0 and N_{ch} for iron primaries	42
3.16. Distributions of the factor c_0^{ch} for iron primaries	43
3.17. Distributions of the geometric mean of the fractions $f_{ch,en}$ for iron primaries	43
3.18. Comparison of X_{max} from simulations with calculated values	44
3.19. Comparison of N_μ from simulations with calculated values	45
3.20. Comparison of N_{max} from simulations with calculated values	45
3.21. Mean elasticity in p-air and π -air collisions	46
3.22. Assessment of the muon deficit in modern hadronic interaction models	49
3.23. Impact of changing individual hadronic interaction features on the means and standard deviations of the logarithm of N_μ and X_{max}	51
3.24. Comparison of figure 3.23 with Auger data	52
3.25. Correlation between N_μ and α_1	56
4.1. Simple model for N_μ and ΔX	58
4.2. β as a function of charged multiplicity N_{ch} and inelasticity κ	59
4.3. Distributions of mode(κ) and N_{tot}^{FI} for proton primaries	60
4.4. Distributions of $\langle \kappa \rangle_w$ and $\langle \tilde{f}_{ch,en} \rangle_w$ for proton primaries	61
4.5. Distributions of N_{ch}^{FI} and $\langle \tilde{f}_{ch,en} \rangle_{geom}$ for proton primaries	65
4.6. ΔX - N_μ distributions as calculated with our model	67
4.7. Behaviour of our parameters on the ΔX - N_μ distributions	68
5.1. Distributions of the most influential parameters for the prediction of X_{max} and N_μ through neural networks.	74
5.2. Outline of a neural network structure	75
5.3. Learning curves of the neural networks for EPOS-LHC	78
5.4. Individual neural networks predicting X_{max} for each high-energy interaction model separately	79
5.5. Individual neural networks predicting N_μ for each high-energy interaction model separately	80
5.6. Comparison of the X_{max} - N_μ distributions from simulations with the distributions obtained through our individual models	81
5.7. Universal neural network for X_{max} prediction	82
5.8. Universal neural network for N_μ prediction	83
5.9. Comparison of the X_{max} - N_μ distributions from simulations with the distributions obtained through our universal model	84
5.10. Typical result for N_μ prediction with too few parameters	86
5.11. Universal neural network predicting X_{max} with reduced number of features	87

5.12. Universal neural network predicting N_μ with reduced number of features	88
5.13. Comparison between the original distributions of X_{max} and N_μ and the distributions obtained through our model	89
5.14. Depiction of our method based on distributions	91
5.15. Distributions of X_0 and M_0	93
5.16. Distributions of F_0 and K_{eff}	94
5.17. Impact of changing basic parameters on the observables for EPOS-LHC	95
5.18. Impact of changing basic parameters on the observables for QGSJETII-04	96
5.19. Impact of changing basic parameters on the observables for SIBYLL-2.3d	97
5.20. χ^2 -analysis to infer the parameters from EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d	98
5.21. X_{max} resolution as a function of energy	99
5.22. Minimum χ^2 fit and 1 σ and 2 σ regions	101
5.23. Minimum χ^2 fit and 1 σ and 2 σ regions after including the effect of the measurement resolution	102
6.1. Mean total number of muons at ground N_μ as a function of primary energy for different zenith angles for QGSJETII-03	104
6.2. X_{max} and N_μ distributions from the Auger dataset	105
6.3. X_{max} - N_μ distributions after taking into account measurement resolution and mass composition	106
6.4. X_{max} distributions before and after performing the cut	107
6.5. N_μ distributions before and after performing the cut	108
6.6. Minimum χ^2 fit for the calculation of M_0^{loc} and K_{eff}^{loc}	109
6.7. Minimum χ^2 fit for the calculation of F_0^{loc} and F_0^{scale}	110
6.8. Predicted distributions of M_0 , N_{ch}^{FI} , F_0 and K_{eff} for the Auger dataset	111

Table Index

4.1.	Current models and our model are tested on ranges of the parameters N_{tot}^{FI} , N_{tot} , κ^{FI} , κ and $\tilde{f}_{ch,en}^{FI}$ instead of using constant mean values. The used ranges are presented here for the different energies and are motivated by previously presented plots. We also summarize in the second column how we interpret each parameter.	62
4.2.	The ranges of parameters presented in table 4.1 are inserted into the expressions for ΔX and N_μ . The resulting ranges of calculated ΔX and N_μ are compared to the corresponding ranges obtained from simulations done with CONEX using EPOS-LHC with vertical protons of different energies as primaries. N_μ is calculated combining the approaches from Matthews, Kampert and Unger. ΔX is calculated in two ways, using Matthews' approach (M.) and using Kampert and Ungers' approach (K., U.).	63
4.3.	In the following we test our model which can be summarized in expressions (4.1), (4.4) and (4.5) which depend on the six parameters N_{ch}^{FI} , N_{ch} , κ^{FI} , κ , $\tilde{f}_{ch,en}^{FI}$ and $\tilde{f}_{ch,en}$. In this table we summarize the interpretations we assign them (second column), the values we obtain from simulations of 10^{20} eV proton initiated showers (third column) and the new simpler name we assign them for the next chapter on neural networks (fourth column). For N_{ch}^{FI} and $\tilde{f}_{ch,en}^{FI}$ we are interested in ranges of values. For κ^{FI} we will only consider the upper and lower extreme values. For the other three effective parameters of the rest of the shower we consider mean values because their distributions are quite narrow. Furthermore, we redefine M_0 as $\ln(N_{ch}^{FI} + 1)$, which is the feature we will end up using when training neural networks. Finally, we give a short description of each parameter (fifth column).	66

Abstract

Since the discovery of cosmic rays during the first decade of the 20th century many experiments were designed to study them directly or by means of the extensive air showers they generate when entering the Earth's atmosphere. The largest observatory designed to detect air showers is the Pierre Auger Observatory situated in Malargüe, Argentina. Many questions related to astrophysics and particle physics are tackled here. In particular, since cosmic rays cover energies far above those achievable in man-made accelerators, they represent excellent and unique probes to study physical properties at the highest energies.

As an extensive air shower develops in the atmosphere the number of particles that constitute it grows. At the same time individual energies decrease until it is more probable for particles to decay or be absorbed in the atmosphere, which diminishes the number of particles again. This means that there is a position of maximum development X_{max} . Hadronically interacting particles ultimately yield muons which can be measured at ground. This number N_μ together with X_{max} are observables measured at the Pierre Auger Observatory that present a meaningful anticorrelation.

In the present work this anticorrelation is studied. An analytical model is developed, which explains the X_{max} - N_μ anticorrelation as a function of parameters describing the multiplicity of hadronically interacting particles, the fraction of energy that is taken by these particles and the inelasticity of the first interaction and corresponding effective macro-parameters representative for the whole shower. This model is then further improved using neural networks trained with the values of the parameters and observables obtained from simulations performed with CONEX. The resulting model is universal in the sense that the performance does not depend on the high-energy interaction model used during simulation. Finally, a model with a reduced set of parameters is applied to a preliminary dataset from Auger. The distributions of the hadronic multiplicity of the first interaction, the hadronic energy fraction of the first interaction and the effective inelasticity of the rest of the shower for the dataset are inferred. They reveal that the hadronic multiplicity and the hadronic energy fraction of the first interaction are generally too low in current high-energy interaction models used for simulations.

Keywords: EXTENSIVE AIR SHOWER, HIGH-ENERGY HADRONIC INTERACTIONS, NEURAL NETWORKS

Resumen

Desde el descubrimiento de los rayos cósmicos a principios del siglo XX se diseñaron numerosos experimentos para estudiarlos de forma directa o mediante las lluvias atmosféricas que generan al entrar en la atmósfera terrestre. El observatorio más grande diseñado para detectar estas lluvias es el Observatorio Pierre Auger ubicado en Malargüe, Argentina. Aquí se abordan muchas cuestiones relacionadas con la astrofísica y la física de partículas. En particular, dado que los rayos cósmicos alcanzan energías superiores a las que son posibles en los aceleradores artificiales, representan objetos excelentes y únicos para estudiar las propiedades físicas a las más altas energías.

A medida que una lluvia avanza en la atmósfera, crece el número de partículas que la constituyen. Al mismo tiempo, las energías individuales disminuyen hasta que es más probable que las partículas decaigan o sean absorbidas por la atmósfera, lo que vuelve a disminuir el número de partículas. Esto significa que existe una posición de máximo desarrollo X_{max} . Las partículas que interactúan hadrónicamente producen muones que se pueden medir a nivel del suelo. N_μ junto con X_{max} son observables medidos en el Observatorio Pierre Auger que presentan una anticorrelación significativa.

En el presente trabajo se estudia esta anticorrelación. Se desarrolla un modelo analítico que expresa la anticorrelación $X_{max}-N_\mu$ como función de parámetros que describen la multiplicidad de partículas que interactúan hadrónicamente, la fracción de energía que estas partículas se llevan y la inelasticidad de la primera interacción y correspondientes macroparámetros efectivos representativos de toda la lluvia. Luego se mejora este modelo utilizando redes neuronales entrenadas con los valores de los parámetros y observables obtenidos de simulaciones realizadas con CONEX. El modelo resultante es universal en el sentido que no depende del modelo de interacciones a altas energías utilizado durante las simulaciones. Finalmente, se aplica un modelo con un conjunto reducido de parámetros a un conjunto preliminar de datos de Auger. Se infieren para el conjunto de datos las distribuciones de la multiplicidad hadrónica de la primera interacción, la fracción de energía hadrónica de la primera interacción y la inelasticidad efectiva del resto de la lluvia. Estas revelan que la multiplicidad y la fracción de energía hadrónica de la primera interacción son generalmente demasiado bajas en los modelos actuales de interacciones a altas energías utilizados en las simulaciones.

Palabras clave: LLUVIA ATMOSFÉRICA EXTENDIDA, INTERACCIONES HADRÓNICAS A ALTAS ENERGÍAS, REDES NEURONALES

Zusammenfassung

Seit der Entdeckung der kosmischen Strahlung Anfang des 20. Jahrhunderts wurden viele Experimente entwickelt, um sie direkt oder anhand der Luftschauer zu studieren, die sie beim Eintritt in die Erdatmosphäre erzeugen. Das größte Observatorium zur Erfassung von Luftschauern ist das Pierre-Auger-Observatorium in Malargüe, Argentinien. Hier werden viele Fragen der Astro- und Teilchenphysik behandelt. Da kosmische Strahlung Energien abdeckt, die weit über denen liegen, die in künstlichen Beschleunigern erreichbar sind, stellen sie hervorragende und einzigartige Objekte dar, an denen man physikalische Eigenschaften bei höchsten Energien studieren kann.

Wenn ein Luftschauer die Atmosphäre durchquert, wächst die Anzahl der in ihm enthaltenen Teilchen. Gleichzeitig nehmen ihre Energien ab. Wenn diese tief genug sind, zerfallen die Teilchen oder werden in der Atmosphäre absorbiert, was die Zahl der Teilchen wieder verringert. Dies ergibt eine Position maximaler Entwicklung X_{max} . Hadronisch wechselwirkende Teilchen bringen Myonen hervor, die am Boden gemessen werden können. Diese Zahl N_μ zusammen mit X_{max} sind Observablen, die in Auger gemessen werden und eine aussagekräftige Antikorrelation aufweisen.

In der vorliegenden Arbeit wird diese Antikorrelation studiert. Es wird ein analytisches Modell entwickelt, das diese als Funktion von Parametern wiedergibt, die die hadronische Multiplizität, den hadronischen Energieanteil und die Inelastizität der ersten Interaktion beschreiben, zusammen mit entsprechenden effektiven Parametern, die für den gesamten Schauer repräsentativ sind. Dieses Modell wird dann anhand künstlicher neuronaler Netze weiter verbessert. Hierfür werden Werte der Parameter und Observablen von Simulationen benutzt, die mit CONEX durchgeführt wurden. Das resultierende Modell hängt nicht von dem während der Simulation verwendeten Hochenergie-Wechselwirkungsmodell ab. Schließlich wird ein Modell mit reduziertem Parametersatz auf einen vorläufigen Datensatz von Auger angewendet. Die Verteilungen der hadronischen Multiplizität und des hadronischen Energieanteils der ersten Wechselwirkung und die der effektiven Inelastizität werden diesem Datensatz entnommen. Sie zeigen, dass die Multiplizität und der Energieanteil der ersten Wechselwirkung in aktuellen Modellen, die für Simulationen verwendet werden, generell zu niedrig sind.

Stichwörter: LUFTSCHAUER, HADRONISCHE INTERAKTIONEN BEI HOHEN ENERGIEN, KÜNSTLICHE NEURONALE NETZWERKE

Chapter 1

Cosmic rays

“Yes, the cosmic rays, the shortest wavelength and most highly penetrating of all vibratory forces. It has been known that they beat unceasingly upon the Earth from outer space, cast forth by the huge generators of the stars (...).”

— Edmond Hamilton, *The man who evolved*

At the beginning of the 20th century it became clear that there is more to the universe than large structures like stars and smaller components like gas and dust: *cosmic rays* were discovered. In contrast to gas and dust, cosmic rays are relativistic charged particles impinging on the Earth’s atmosphere. Until the 1950s they represented the only probes of high energy particles. This resulted in many discoveries in the area of particle physics. With the advent of the particle accelerator era the interest shifted into more astrophysical questions concerning cosmic rays. Nonetheless, cosmic rays remain to date the only possibility to understand aspects of particle physics in energy ranges beyond those achievable in particle accelerators. In this chapter we summarize discoveries and scientific results in cosmic ray physics, some of which are rather important to the present work.

1.1. Historical development

Between 1911 and 1913, Victor Hess undertook the work that won him the Nobel Prize in Physics in 1936. The assumption at the time was that radiation would decrease with increasing distance from the Earth, the then assumed source of radiation. On the contrary, Hess discovered, using radiation detectors on balloon flights (see figure 1.1, left), that ionizing radiation increases with altitude [3]. He interpreted these results as *“likely to be explained by the assumption that radiation of very high penetrating power enters from above into our atmosphere, and even in its lowest layers causes part of the*

ionization observed in closed vessels.” He early ruled out the Sun as a possible source since there was no decrease in the radiation either at night or during a solar eclipse.

In the late 20s, Skobeltsyn constructed a cloud chamber which was in principle destined for studies of electrons emitted in radioactive decays. He observed tracks that were hardly deflected at all and didn’t fit the picture. These turned out to be the first pictures of tracks left by high energy cosmic rays [5]. At the same time the Geiger-Müller detector was invented. It enabled individual cosmic rays to be detected and their arrival times determined very precisely. Bothe and Kolhörster coupled the coincidence technique with the Geiger-Müller counter to study cosmic rays, marking so the start of cosmic ray research as a branch of physics [6].

By that time, the first major achievement in the area of relativistic quantum mechanics was Dirac’s discovery of the equation that bears his name [7]. It was intended to describe free electrons but had a problematic feature. For every positive-energy solution it admitted a solution with negative energy. To rescue his equation, Dirac postulated that the negative-energy states are filled by an infinite “sea” of electrons. A “hole” in the sea would be experienced as an ordinary particle with positive energy and positive charge. This cumbersome explanation turned out to be fruitful. A particle with the required properties was indeed discovered by Anderson in 1933 [4]: the positron. Its track was observed on a cloud chamber picture of cosmic radiation (see figure 1.1, right).

This set the beginning of a series of detections of new particles in the products of cosmic rays. Some of them were searched for, others came unexpected. For example, Yukawa postulated some properties a mediator of the strong force that binds protons and neutrons together in atomic nuclei should possess in order to account for known features [8]. In 1937 Neddermeyer and Anderson discovered in cloud-chamber photographs the muon [9] which, because of its appropriate mass, was at first wrongly identified as



Figure 1.1: Left: Victor Hess, at the center, departing from Vienna about 1911, soon discovering the penetrating ionizing radiation from outer space. Right: Anderson’s cloud chamber picture of cosmic radiation from 1932 showing for the first time the existence of the positron (picture taken from [4]). The particle enters from the bottom, strikes the lead plate in the middle and loses energy as can be seen from the greater curvature of the upper part of the track. The curvature was induced intentionally by a magnetic field.

this mediator. Later, the muon was correctly assigned as a member of the lepton group of subatomic particles. Today we know that the exchange particle for the strong force between quarks is the gluon, discovered much later at the electron-positron collider PETRA of DESY in 1979 [10]. When looking into the hierarchy of scales governing the nucleon-nucleon interaction the long distance region is dominated by one-pion exchange whereas the intermediate distance region is dominated by two-pion exchange [11]. The first experimental hint of the neutral pion was observed in the mixed cosmic ray showers observed by Chao [12] and Fretter [13] using the same experimental technique as Neddermeyer and Anderson. On the other hand, charged pions were found by the collaboration lead by Cecil Powell in 1947 [14]. They used photographic emulsions that were exposed to cosmic rays for long periods of time in sites located at high-altitude mountains.

Rochester and Butler reported in 1947 the discovery of forked particle tracks “*of a very striking character*” in the form of “V”s in cloud-chamber photographs of cosmic ray showers [15]. The neutral particle which decays into the two particles rendering the V-formed tracks is now known as the K^0 . In a similar fashion the Λ particle was discovered by Anderson’s group in 1950 [16]. Being these particles and others unexpectedly heavy, they came to be known collectively as “strange” particles, which would give later the strange quark its name.

The discovery of the phenomenon that will be the main protagonist of the present work was not straightforward. At the beginning in 1934 Rossi observed correlated arri-

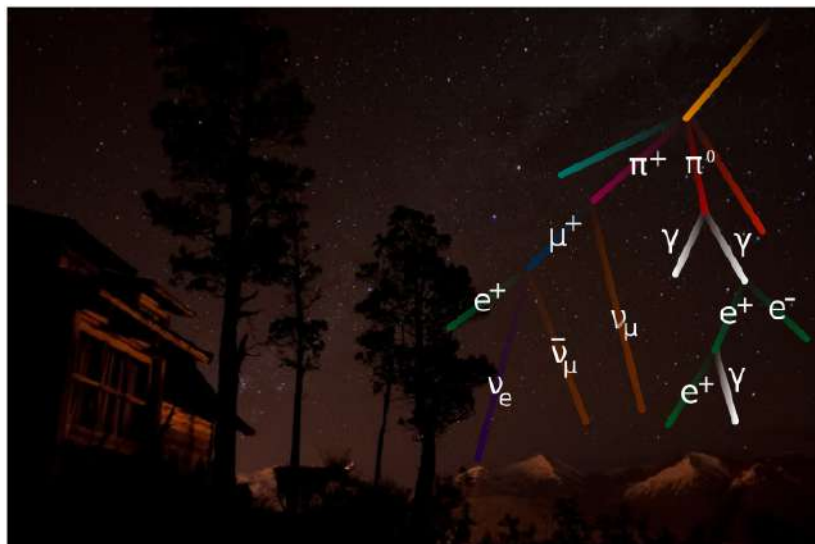


Figure 1.2: Simplified sketch of an extensive air shower generated by a cosmic ray. It can be appreciated that the generation of new particles and subsequent subshowers quickly results in a large number of particles. As the number of particles increases, the average energy carried by them decreases. At some point this energy is so low that these particles decay or are absorbed instead of contributing to multiplicative processes, i.e. creating new particles. Thus, there is a depth in the atmosphere where the number of particles is maximum.

vals of particles at widely-separated detectors [17]. We understand today that cosmic rays interact with molecules in the atmosphere setting of a succession of interactions between the products of the cosmic ray and other molecules in the atmosphere. When the cosmic ray, which we will call the *primary particle* in the present work, interacts with some molecule in the atmosphere, new particles are created which themselves undergo this process after traveling a certain distance in the atmosphere. Today we call this cascading process an *extensive air shower* (see figure 1.2 for a simplified schematic representation). In each interaction, energy is distributed among the newly created particles. At some point this energy is so low that these particles decay or are absorbed instead of contributing to more multiplicative processes, i.e. creating new particles. Thus, there is a depth in the atmosphere where the number of particles is maximum. What Rossi observed for the first time were the product particles of one such event arriving at the site of his measurements. This phenomenon was understood and studied by Schmeiser and Bothe who analyzed independently transition curves published by Rossi himself [18]. After detailed studies they already gave a first description of different components of particles present in air showers.

On the other hand, Kolhörster and his group described in 1938 how the rate at which coincidences between a pair of Geiger-Müller counters falls as a function of separation distance [19]. In addition, they already hypothesized that it should be possible to use information about the arrival points of product particles at ground to infer the point of origin of the air shower that generated them, what we will call the *depth of first interaction* in this work. Even though we know much more today and have collected huge amounts of data in different experiments we have not yet worked out how to do this inference on a shower-to-shower basis.

Pierre Auger and his collaborators calculated in 1939, using these coincidences, the first estimates of energies carried by primary cosmic rays [20]. Their method was based on the number of particles in the observed showers and on the assumption that each particle at ground would carry, on average, the *critical energy*. The concept of critical energy will be used frequently in this work and properly defined in section 3.4. The obtained values for the primary energies went up to 10^{15} eV. Due to these high values Auger theorized already back then that cosmic rays should “*acquire their energy along electric fields of a very great extension*”. Both, the energy spectrum of cosmic rays and processes of acceleration of cosmic rays, are present objects of study.

Luckily, by the time the idea of extensive air showers had settled, great advances in quantum electrodynamics had already been done, not so much in fundamentals of hadronic interactions. But since a huge portion of cascading processes is electromagnetic, advances in air shower physics could come into view which itself motivated more investigation in this area. Bethe and Heitler calculated in 1934 the probabilities for pair-production and bremsstrahlung, accounting here for the screening of the atomic

field [21]. Carlson and Oppenheimer extended the theory in 1937 taking into consideration the energy losses of electrons through ionization of the medium [22]. They also introduced diffusion equations to deal with the calculations. Using these they observed the following key features of extensive air showers:

- the total number of electromagnetic particles (electrons, positrons, and photons) at the shower maximum is proportional to the energy carried by the primary cosmic ray that generated the shower,
- the position of the shower maximum develops proportional to the logarithm of the primary energy.

The diffusive equations were expanded throughout the years to include more physical aspects (such as multiple scattering) and variables of showers (such as lateral position) [23, 24]. A good three dimensional description of electromagnetic subshowers of extensive air showers was converged upon. Before going into more details on the description of electromagnetic subshowers and also that of hadronic processes in chapter 3, we will go through the major applications of cosmic ray physics to astrophysics and particle physics in the next few sections.

All these topics are investigated at the *Pierre Auger Observatory* located at a vast plain known as the Pampa Amarilla near Malargüe in Mendoza, Argentina [40, 43, 45]. This is the largest observatory up to date for detection of extensive air showers generated by cosmic rays. This observatory also stands out because it combines two main independent detection techniques making it a so-called *hybrid detector*. An array of *surface detector* stations and a collection of air *fluorescence telescopes* provide two independent ways to measure air showers. The first method detects charged particles through their interaction with water placed in surface detector tanks, while the second one tracks the development of air showers by observing ultraviolet light emitted high in the Earth's atmosphere.

1.2. Applications to astrophysics

1.2.1. Energy spectrum

Measuring the energy spectrum of cosmic rays at high precision is relevant for understanding their origin and mechanisms of cosmic ray acceleration and propagation. Figure 1.3 shows the “all-particle” spectrum combining the results of several experiments. In order to be able to compare results from different observatories the differential energy spectrum is displayed. Furthermore, it has been multiplied by a factor of $E^{2.6}$ in order to highlight features of the spectrum that because of its steepness are difficult to

discern. The spectrum can be described very well by a power law with varying spectral index for the different regions between these characteristic kinks. The so-called *knee* between 10^{15} eV and 10^{16} eV, the *second knee* near 10^{17} eV, the *ankle* around $10^{18.5}$ eV and the flux suppression at the higher end of the spectrum stand out and will be commented upon in the next paragraphs of this section.

There are two viable candidate theories to explain the physical origin of the knee feature [25]. The origin could be related to the way cosmic rays propagate. PeV cosmic ray protons possess in galactic magnetic fields of a strength of around $1 \mu\text{G}$ Larmor radii of around 1 pc. The *Larmor radius* is the radius of the circular motion of a charged particle in the presence of a uniform magnetic field. This means that the confinement of cosmic rays of higher energies in our galactic disk becomes significantly less efficient causing thus a steepening of the cosmic ray spectrum. This scenario also requires the existence of galactic cosmic ray sources which are capable of accelerating particles to energies well beyond the knee energy. An alternative explanation to the knee feature is that the PeV energy scale actually is the maximum energy achievable in galactic cosmic ray sources, which are thought to be supernova remnants. The *Hillas criterion*

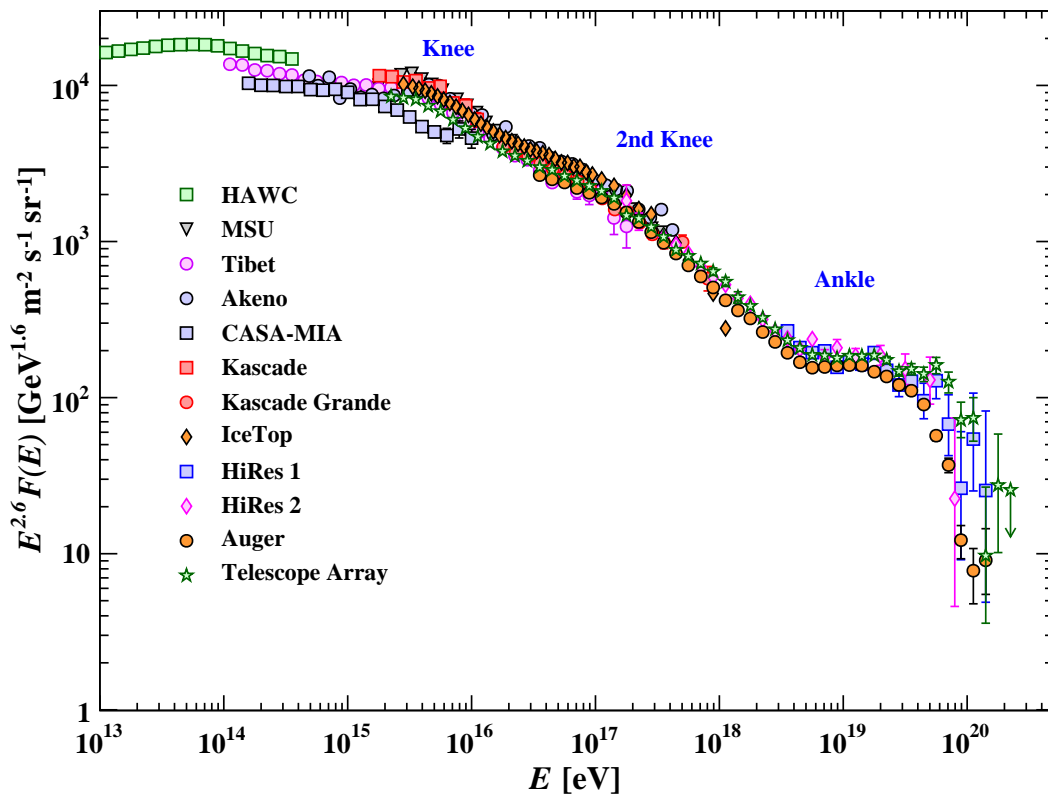


Figure 1.3: Cosmic ray differential energy spectrum reconstructed from air showers observed by a variety of experiments. The spectrum has been multiplied by $E^{2.6}$ in order to enhance the features that are present when the spectral index of the power law describing the spectrum changes. These features are the so-called *knee* between 10^{15} and 10^{16} eV, the *second knee* near 10^{17} eV, the *ankle* around $10^{18.5}$ eV and the flux suppression at the higher end of the spectrum.

can be used to study this possibility. It states that

$$E_{max} = \eta^{-1} \beta_{sh} q B R \quad (1.1)$$

is the maximum energy achievable by a particle of charge q at a source of characteristic size $R = l \cdot \Gamma$ with a magnetic field of strength B . l is the comoving size of the source and Γ the Lorentz factor of the motion. Assuming a shock acceleration mechanism with velocity β_{sh} in units of the speed of light, η represents the efficiency of the acceleration. Otherwise these two parameters do not enter the equation. Equation (1.1) is obtained demanding that the Larmor radius of the particle not exceed the size of the acceleration region. It shows that a considerable magnetic field is needed in order to see in supernova remnants feasible candidates (see figure 1.4, left). If this situation is met remains unknown.

Since expression (1.1) depends on the charge of the accelerated particle it is actually an open question whether the energy region closely below the knee corresponds to protons or if it is dominated by other masses. This is valid for the argument given above for either possible explanation of the knee feature. Assuming that cosmic rays with energy closely below the knee are mainly by protons implies that a similar feature should happen approximately 100 PeV above for iron cosmic rays. This could indeed be the explanation of the second knee [26].

At energies just above that of the second knee a steepening in the flux of heavy

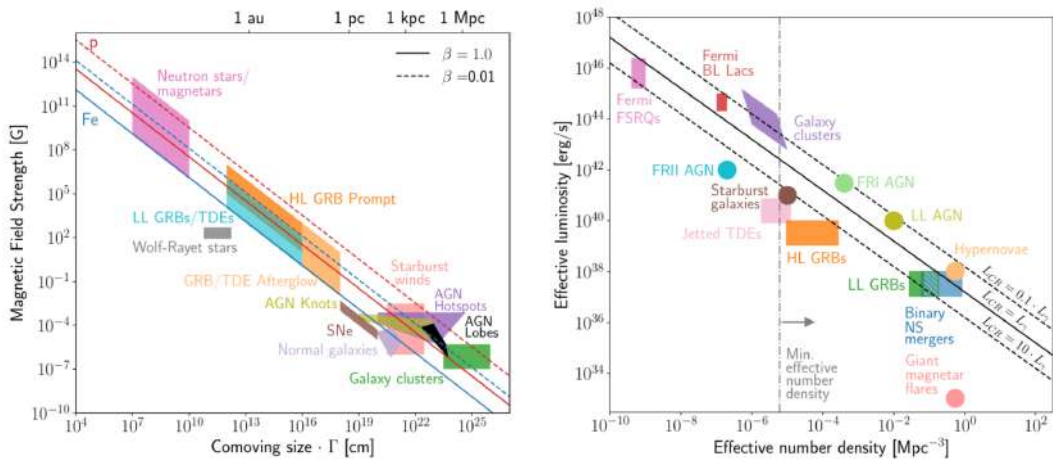


Figure 1.4: Left: Hillas diagram which shows typical characteristic sizes R and magnetic fields B of different astrophysical sources [27]. Solid and dashed lines show for which values of B and R confinement of protons (red) and iron nuclei (blue) with energy 10^{20} eV are possible. Right: Source luminosity and number density of different astrophysical objects. Different black lines correspond to various differences between the cosmic ray luminosity L_{CR} and the luminosity L_{γ} of the sources in the wavelength studied. The vertical dashed gray line represents the lower limit for the number density of sources necessary to be compatible with the analysis of arrival directions of the cosmic rays of the highest energies detected at the Pierre Auger Observatory [28]. Both diagrams are taken from [29].

primary particles was measured by the KASCADE-Grande experiment [30]. The subsequent flattening of the spectrum means an onset of a new light component which needs to be of extragalactic origin. The flux suppression at the highest energies in the cosmic ray spectrum happens close to the energy threshold of about $6 \cdot 10^{19}$ eV predicted by the *GZK-effect*. This theory states that cosmic rays with energies above this limit lose energy when interacting with photons of the cosmic microwave background to produce pions through the Λ resonance:

$$\gamma_{CMB} + p \rightarrow \Lambda^+ \rightarrow p + \pi^0$$

or

$$\gamma_{CMB} + p \rightarrow \Lambda^+ \rightarrow n + \pi^+.$$

However, a scenario solely based on the GZK-effect does not fit the flux measured by the Pierre Auger Observatory [31]. The cut-off could instead be explained by a combination of propagation effects and the maximum energy achievable at the source [32].

1.2.2. Origin of cosmic rays

Making use of the Hillas criterion given by equation (1.1) one can deduce which combination of magnetic fields and source sizes are necessary in order to obtain certain cosmic ray energies. The *Hillas diagram* in figure 1.4 (left, taken from [29]) summarizes these values for different astrophysical sources. Solid and dashed lines show there for which values of B and R confinement of protons (red) and iron nuclei (blue) with energy 10^{20} eV are possible. Objects left of the lines do not satisfy the Hillas criterion to be able to accelerate cosmic rays up to this energy [27].

In addition to certain sizes and magnetic fields, cosmic ray accelerators need to possess a specific energy budget in order to produce the ultra-high energy cosmic ray flux we observe at Earth. The budget of a particular source type can be estimated from the source luminosity and the number density of this source type. A summary of these values similar to the Hillas diagram is shown in figure 1.4 (right, taken from [29]). The energy budget corresponding to each combination of values is compared to the ultra-high energy cosmic ray rate estimated in [34]. The region at the left of the black lines does not satisfy the energy budget condition, while each black line represents various differences between the cosmic ray luminosity L_{CR} and the luminosity L_γ of the sources in the wavelength studied. The vertical dashed gray line represents the lower limit for the number density of sources necessary to be compatible with the analysis of arrival directions of the cosmic rays of the highest energies detected in the Pierre Auger Observatory [28].

1.2.3. Anisotropy

The search for concrete sources of cosmic rays of the highest energies motivates the study of excesses in the flux of these cosmic rays. Considering the highest energies is necessary in this pursuit in order to deal with the fact that the propagation of charged particles is affected by the extragalactic and galactic magnetic fields. The *rigidity*

$$R = \frac{pc}{q}$$

measures the resistance of a particle of charge q and momentum p to deflection by a magnetic field. The higher its energy, the higher its resistance and the more accurately this particle points to the source where it was created. It is therefore realistic to focus on large angular scales for cosmic rays of 10^{19} eV and look for a cumulative flux from multiple closeby objects. An anisotropy in the arrival directions was indeed detected at a more than 5.2σ level of significance by the Pierre Auger Collaboration [33]. This anisotropy can be described by a dipole which can be observed in figure 1.5. The direction of the dipole indicates an extragalactic origin of ultra-high energy particles.

The Pierre Auger Collaboration also performed searches at smaller angular scales and correspondingly higher energies [35]. No statistically significant evidence of anisotropy was discovered but some departure from isotropy was obtained for cosmic rays with energies above 58 EeV in a region around the Swift active galactic nucleus located at less than 130 Mpc and around the direction of Centaurus A.

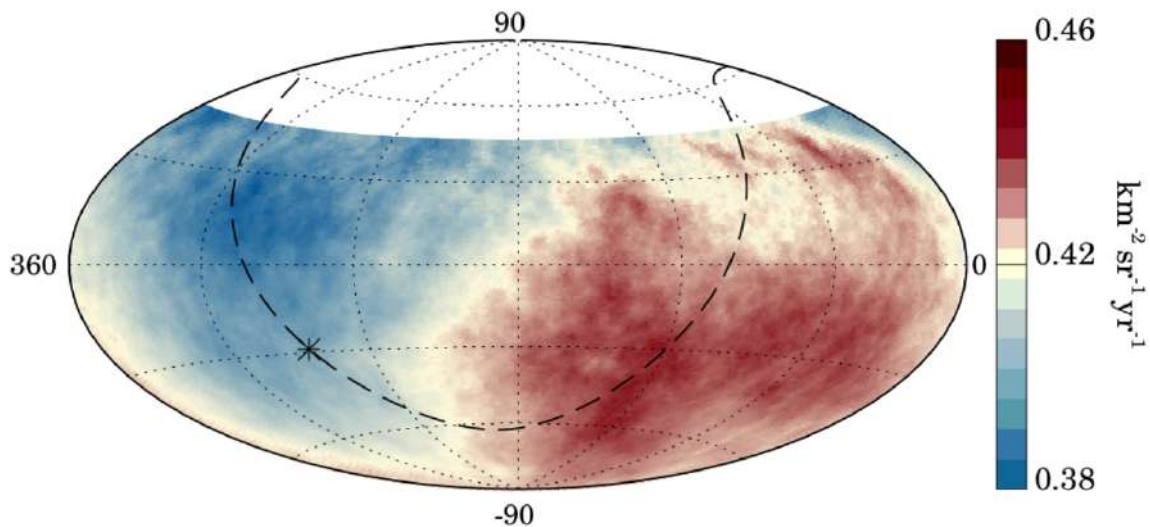


Figure 1.5: Smoothed cosmic ray flux for energies above 3 EeV in equatorial coordinates as measured at the Pierre Auger Observatory (figure taken from [33]). The dashed line follows the galactic plane while the star indicates the center of the galaxy. The region with no values is not covered by the measurements from the Pierre Auger Observatory.

1.2.4. Mass composition

We have already seen that in order to understand the features in the energy spectrum, the mass of the cosmic ray needs to be known. One of the most robust mass sensitive observables is the position X_{max} in the atmosphere at which an air shower attains the maximum number of particles (see explanation to figure 1.2). In [36], X_{max} distributions of simulations of proton, helium, nitrogen and iron initiated showers were parameterized in different energy bins. They used simulations done with different hadronic interaction models (QGSJETII-04, EPOS-LHC and SIBYLL-2.3). These models are implemented in high-energy interaction generators used in frameworks for Monte Carlo simulation of extensive air showers and will be explained in section 3.3. For different combinations of primary fractions they obtain different total X_{max} distributions. Comparing these to the true X_{max} distribution measured at the Pierre Auger Observatory they obtain a fit for the values of the different fractions. For each high-energy interaction model used when simulating, they obtain different fit values. Figure 1.6 shows the obtained fractions as a function of energy for the three high-energy interaction scenarios considered. The helium and nitrogen fractions as a function of energy have a strong dependence on the particular hadronic interaction model used. However,

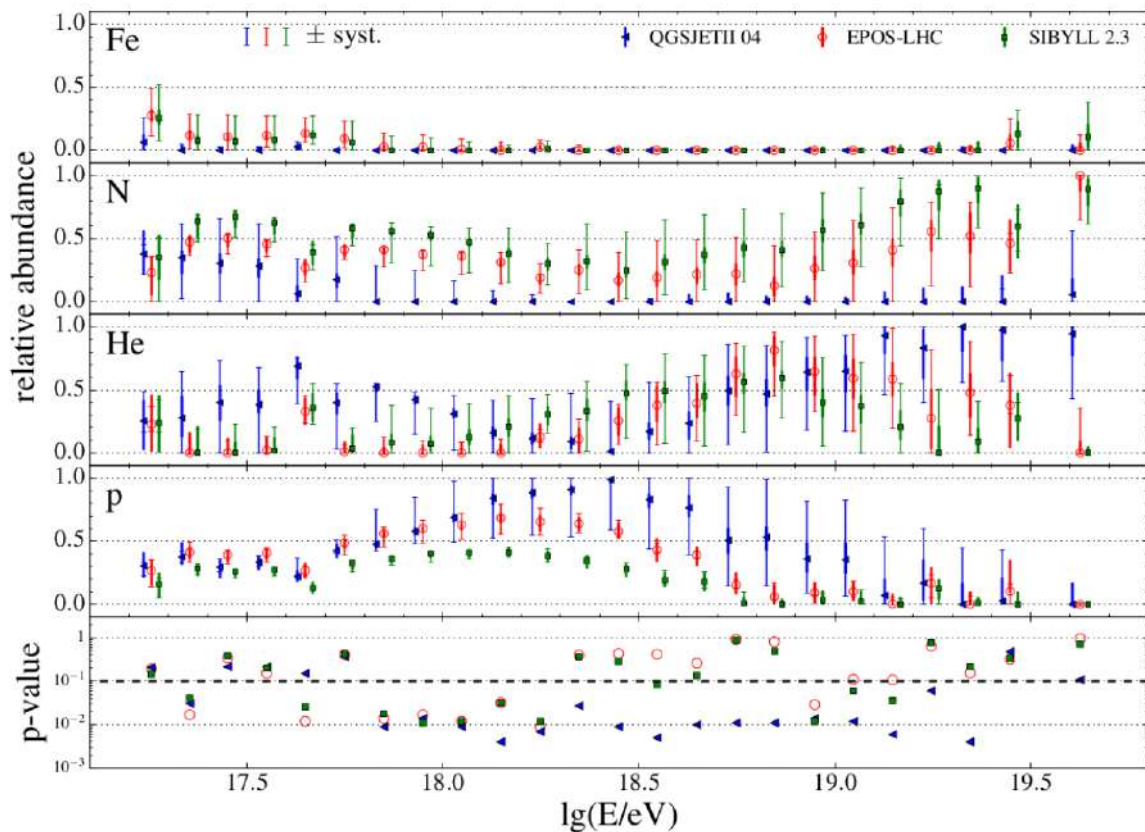


Figure 1.6: Mass fraction fits obtained by the Pierre Auger Observatory (figure taken from [36]). The error bars indicate the statistic (smaller cap) and the systematic uncertainties (larger cap). The bottom panel indicates the goodness of the fits (p-values).

the three models agree in the vanishing iron component between $10^{18.3}$ eV and $10^{19.4}$ eV.

Another way of studying the mass composition is comparing the mean and the standard deviation of the distribution of X_{max} at different energies with the corresponding values from simulated pure proton and pure iron showers [36]. The results obtained by the Pierre Auger Collaboration are summarized in figure 1.7. The red and blue lines correspond to the values obtained from simulations considering pure proton and pure iron fluxes, respectively. Full, dashed and dashed-dotted lines correspond to the high-energy interaction models EPOS-LHC, SIBYLL-2.3 and QGSJETII-04, respectively. The evolution of the data points as a function of energy is for energies below $10^{18.33}$ eV steeper than expected for a constant mass composition (see figure 1.7, left) and indicates that the mean primary mass becomes lighter as the primary energy increases. At higher energies the slope is smaller than expected for a constant mass composition and suggests a development to heavier masses when energy increases. This is confirmed by the development of the standard deviation of X_{max} (see figure 1.7, right).

1.3. Application to particle physics

The *Large Hadron Collider* is a superconducting hadron accelerator designed to accelerate proton beams up to center-of-mass energies of almost 14 TeV, being this the highest energy achieved by man-made accelerators. Since cosmic rays arrive at our atmosphere with energies some orders of magnitude above this value, they represent excellent probes to study physical properties of interactions at ultra-high energies. One such study was carried out at the Pierre Auger Observatory in order to measure the proton-air cross-section [37]. This cross-section is of particular interest because it is not

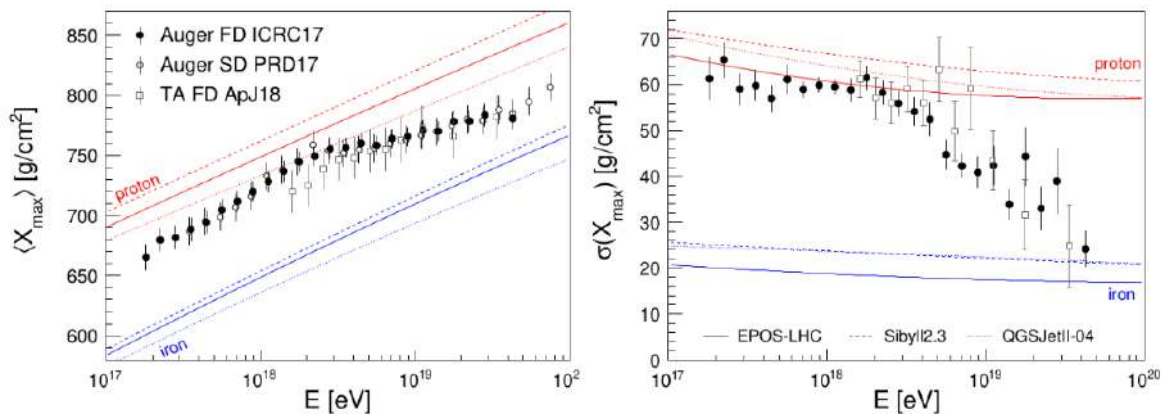


Figure 1.7: Measurements of the mean (left) and the standard deviation (right) of the distribution of X_{max} at different energies. The red and blue lines correspond to the values obtained from simulations considering pure proton and pure iron fluxes, respectively. Full, dashed and dashed-dotted lines correspond to the high-energy interaction models EPOS-LHC, SIBYLL-2.3 and QGSJETII-04, respectively. These models are implemented in the simulation frameworks used and will be discussed in section 3.3. Both figures are taken from [36].

measured for high energies at any accelerator and because it is needed to improve the understanding of extensive air showers where protons and heavier nuclei interact with air molecules. This measurement is performed in two basic steps.

In a first step an air shower observable sensitive to the cross-section is determined. An analysis [38] with air shower simulations revealed that the slope Λ_η of the tail of X_{max} distributions is closely related to the hadronic inelastic cross-section used to simulate them (see figure 1.8, left). This cross-section accounts for all interactions that produce particles and thus contribute to the air shower development. It implicitly includes diffractive interactions, which are inelastic collisions with small energy transfer between the interacting particles. The observable Λ_η was measured at the Pierre Auger Observatory using an X_{max} distribution which included a fraction $\eta\%$ of the most deeply penetrating air showers. This is done to enhance the proton fraction in the dataset used. We will use a similar method which is presented in section 6.2. In a second step this observable is converted into inelastic cross-section. The result of this measurement is shown in figure 1.8 (right) together with direct measurements and predictions obtained from simulations performed using QGSJET-01c, QGSJETII-3, SIBYLL-2.1 and EPOS-1.99 [37]. Knowing this cross-section is crucial for the present work.

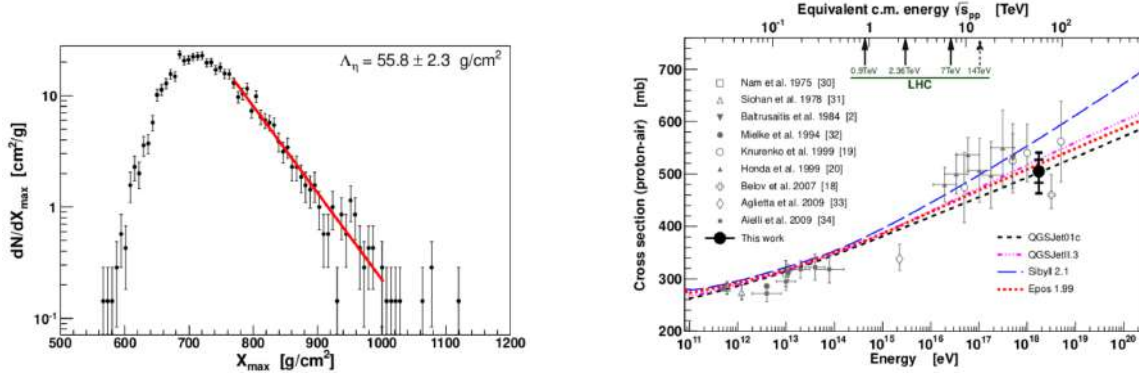


Figure 1.8: Left: X_{max} distribution together with the likelihood fit to obtain Λ_η , the slope of the tail of this distribution. This observable can be converted into inelastic cross-section σ_{inel} . Right: Resulting cross-section σ_{inel}^{p-Air} for proton-air interactions compared to measurements and model predictions of the aforementioned high-energy interaction models included in simulations. Both figures are taken from [37].

Chapter 2

The Pierre Auger Observatory

“Holmes and Watson are on a camping trip. In the middle of the night Holmes wakes up and gives Dr. Watson a nudge. Watson, he says, look up in the sky and tell me what you see.”

— Sherlock Holmes, *The Case-Book of Sherlock Holmes*

At energies higher than 10^{15} eV the flux of cosmic rays drops below one particle per square meter per year [39]. At these energies only indirect measurements can be performed. Cosmic rays of these energies are studied by means of the extensive air showers they originate in the Earth’s atmosphere. The Pierre Auger Observatory was originally designed to study extensive air showers generated by cosmic rays of energies above 10^{18} eV [40] by employing a hybrid detection technique which was briefly introduced in section 1.1.

On the one hand, the *Surface Detector* (SD) [41] consists of an array of more than 1660 water-Cherenkov Detectors (WCDs), which measure the energy deposited by particles from air showers at ground. In figure 2.1 each dot corresponds to one such

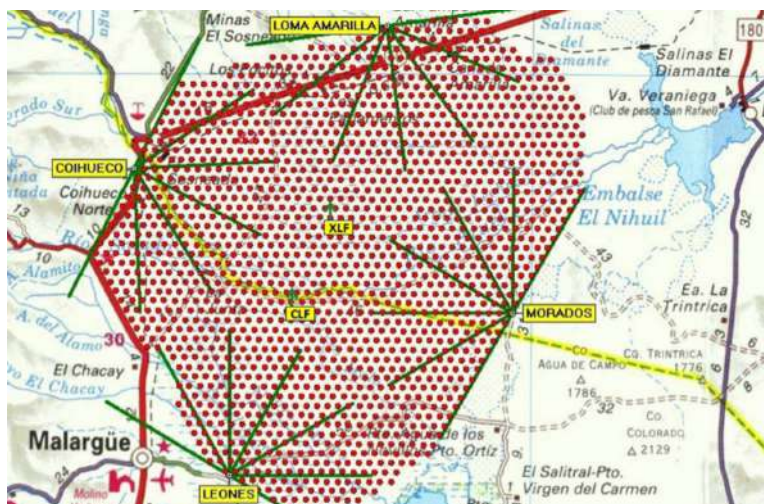


Figure 2.1: Layout of the Pierre Auger Observatory in the Pampa Amarilla near Malargüe in Mendoza, Argentina. Each dot corresponds to one of the 1660 surface detector stations. The four fluorescence detector enclosures are shown, each with the 30° field of view of its six telescopes. Also shown are the two laser facilities, CLF and XLF, near the Observatory center which are used for the calibration of the FD. This figure is obtained from [40].

detector station. On the other hand, the *Fluorescence Detector* (FD) [42] is composed of 27 telescopes deployed at four different locations overlooking the SD. These are symbolized in figure 2.1 by green lines. The telescopes measure the fluorescence light emitted by particles as the shower develops in the atmosphere.

The low event rate of the highest energy cosmic rays requires an area large enough to accumulate good statistics in a reasonable amount of time. The Pierre Auger Observatory covers a surface of around 3000 km² making it thus an ideal detector for the scarce ultra-high energy cosmic rays. Throughout the years the detector was modified to be able to measure lower energy cosmic rays as well.

2.1. The surface detector

The SD array follows a triangular spacing with 1500 m distance between stations. Due to this distance between stations the energy threshold at which the array becomes full efficient is 3×10^{18} eV [44]. Each surface detector station consists of a cylindrical polyethylene tank of about 3.6 m in diameter, which is filled with 12 tons of purified water up to a height of 1.2 m. A picture (left) and a schematic representation (right) are shown in figure 2.2 (taken from [43]). The inner surface is covered by a Tyvek bag which is a highly reflective material. When relativistic charged particles from extensive air showers pass through the water, they produce Cherenkov photons which are then collected by three 9-inch photo-multiplier tubes (PMTs) placed in transparent windows at the inner top of the tank. These are symmetrically distributed at a distance of 1.20m from the center of the tank and look downwards into the water. The inner surface reflects diffusively which enhances the probability for photons to arrive at the photo-multiplier. Each surface detector station is autonomous because it is powered by a solar power system and it contains an electronic package consisting of a processor,

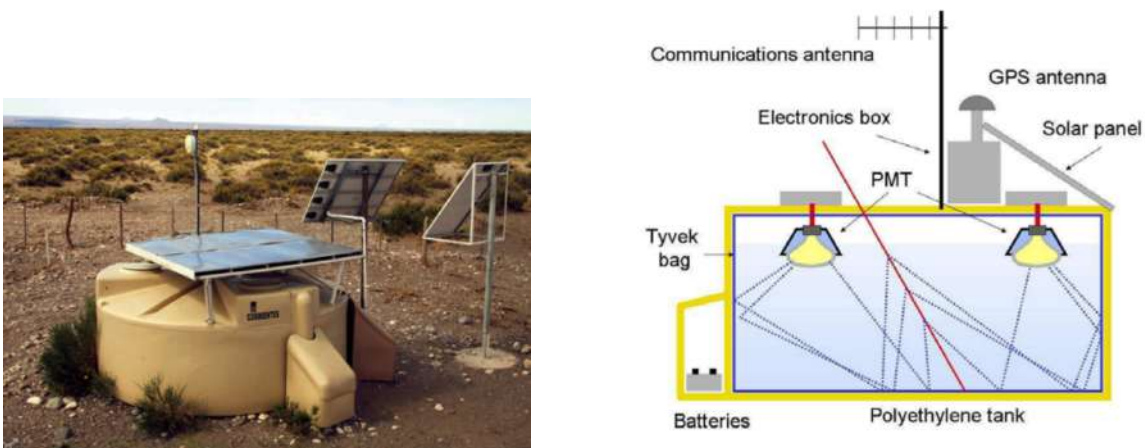


Figure 2.2: Left: Picture of a WCD. Right: Schematic representation of a WCD with its different components (taken from [43]).

GPS receiver, radio transceiver and power controller.

An approximate estimation of the arrival direction of a shower is obtained by fitting the start times of the signals in individual SD stations to a plane front (see figure 2.3, left). In case enough stations are triggered this fit can be improved replacing the plane front by a sphere which grows with the speed of light.

From this sample also an observable to estimate the shower size can be defined. To avoid the large fluctuations in the signal integrated over all distances, caused by fluctuations in the shower development, Hillas proposed to use the signal $S(r)$ at a fixed distance to determine the shower size [27]. In figure 2.3 (right) the lateral distribution of the signals at ground is depicted for an event produced by a cosmic ray of energy (104 ± 11) EeV and zenith angle (25.1 ± 0.1) degrees [45]. The function employed to fit the lateral distribution of signals at ground is a modified *Nishimura-Kamata-Greisen* function [23, 46]:

$$S(r) = S(r_{opt}) \left(\frac{r}{r_{opt}} \right)^\beta \left(\frac{r + r_1}{r_{opt} + r_1} \right)^{\beta+\gamma}$$

where r_{opt} is the optimal distance from the shower core for this calculation, $r_1 = 700$ m and $S(r_{opt})$ is an estimator of the shower size at r_{opt} . At this optimal core distance the fluctuations in the expected signal $S(r_{opt})$, due to a lack of knowledge of the exact lateral distribution function, are minimized. The value for the optimum distance is $r_{opt} = 1000$ m [47].

The value of $S(1000)$ varies with zenith angle θ due to the attenuation of the shower particles and geometrical effects. A function $f_{CIC}(\theta)$ which describes this attenuation can be used to convert $S(1000)$ to a new expression [48]:

$$S_{38} = \frac{S(1000)}{f_{CIC}(38^\circ)}.$$

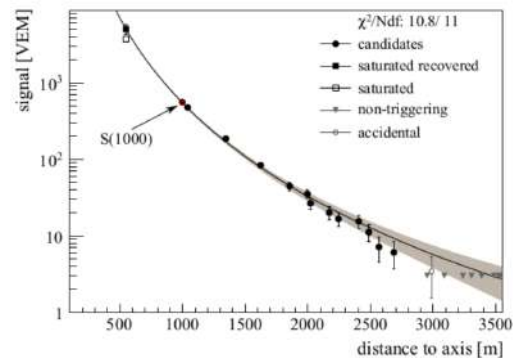
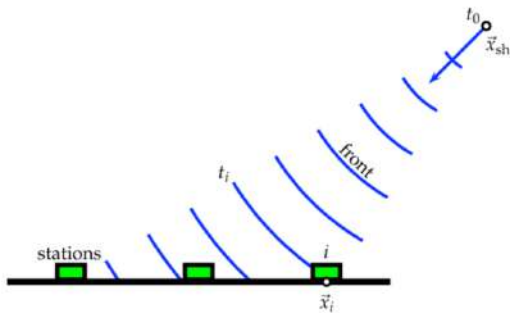


Figure 2.3: Left: Schematic representation of the evolution of the shower front taken into account for the reconstruction of the shower geometry. Right: lateral distribution of the signals at ground for an event produced by a cosmic ray of energy (104 ± 11) EeV and zenith angle (25.1 ± 0.1) degrees. Both figures are taken from [45].

CIC stands for *Constant Intensity Cut*, which is the name of this procedure of describing the attenuation with a function. This number can be regarded as the signal a particular shower with size $S(1000)$ would have produced, had it arrived at a median value of 38° . In order to obtain now an estimator of the energy of the primary particle as recorded with the SD, advantage is taken from the hybrid property of the observatory. Events that triggered independently the SD and the FD are used for cross-calibration [49]. The relation between S_{38} and the calorimetric measurement of the shower energy E_{FD} from the FD is well described by a single power-law function:

$$E_{FD} = A (S_{38}/\text{VEM})^B.$$

The parameters from the data fit are $A = (1.9 \pm 0.05) \times 10^{17}$ eV and $B = (1.025 \pm 0.007) \times 10^{17}$. VEM stands for *Vertical Equivalent Muon*, the total deposited charge of a muon that traverses a WCD vertically and through the center. The final estimator for the energy calculated by the SD is then

$$E_{SD} = A (S(1000)/f_{CIC}(\theta)/\text{VEM})^B.$$

2.2. The fluorescence detector

The basic FD consists of 24 telescopes which are situated at four locations overlooking the SD: Los Leones, Los Morados, Loma Amarilla and Coihueco. At each site, there is an FD building containing six fluorescence telescopes which cover a total field of view of 180° in azimuth and 30° in elevation. A picture of such an FD building is shown in figure 2.4 (left). Three additional telescopes with an elevated field of view

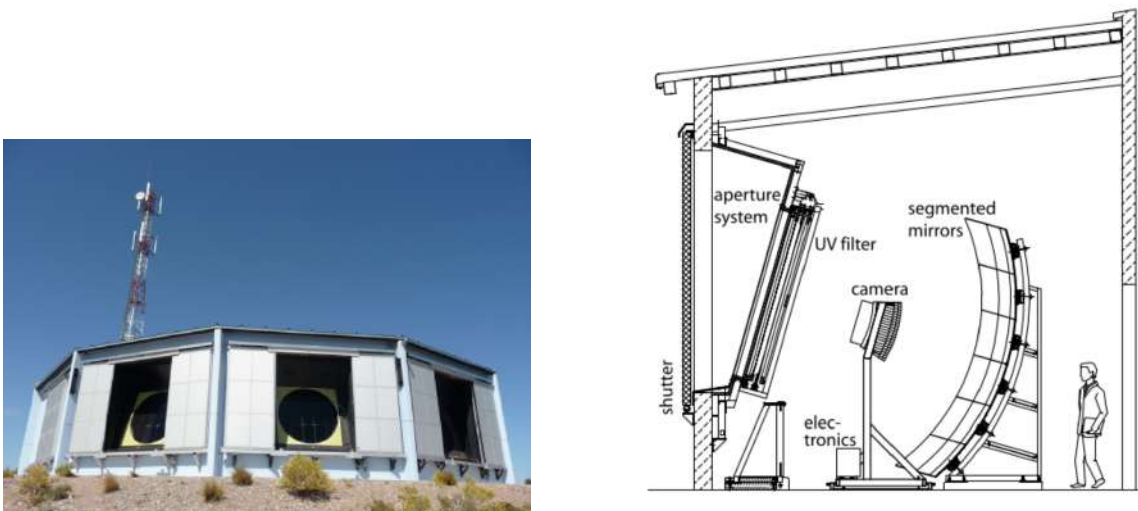


Figure 2.4: Left: Picture of the FD building at Los Leones. Right: Schematic representation of a fluorescence telescope with its different components (taken from [45]).

were deployed near the FD site at Coihueco. The so-called High Elevation Auger Telescopes (HEAT) consists of similar telescopes that can be tilted up to a maximum of 60° . This extends the applicability of the Pierre Auger Observatory to the measurement of shallower showers initiated by lower energy primaries [50]. While the standard set of fluorescence telescopes measures in combination with the SD array (described in section 2.1), which has a 1500 m spacing, the HEAT detectors work together with a denser sub-array of the SD where the detector stations follow a triangular grid with 750 m spacings, the so-called SD-750 [45].

Charged particles from air showers excite nitrogen molecules during their passage through the atmosphere. These molecules de-excite emitting fluorescence light in the range of 300 nm to 400 nm. This emission is isotropical. This light can be measured in the fluorescence telescopes (see figure 2.4). It enters the telescope passing a UV transmitting filter which reduces the background-light. This improves the signal-to-noise ratio of the measured signal and protects the area between the aperture system and the mirrors from the weather and dust. The light is then focused by a mirror (segmented mirrors in figure 2.4, right) onto a camera. This camera is composed of a grid of 22×20 hexagonal PMTs where each PMT represents a pixel. The observation of air showers with the FD is, because of the high sensitivity of the cameras, only possible at night with good weather conditions and with only a limited presence of moonlight. The duty cycle is estimated to be around 15%.

In the FD, air showers are detected as a sequence of triggered pixels in the camera. The first step is to define the plane that contains the shower axis and the detector which is closest to it, as depicted in figure 2.5 (left, taken from [41]). Next, the timing information of the pixels is used to reconstruct the shower axis. There might be some degeneracy in this calculation which can be broken using timing information of the SD stations. This is called the *hybrid reconstruction*. Since the SD is not subject to any restrictions, it has a duty cycle of 100% and, as a consequence, most events observed

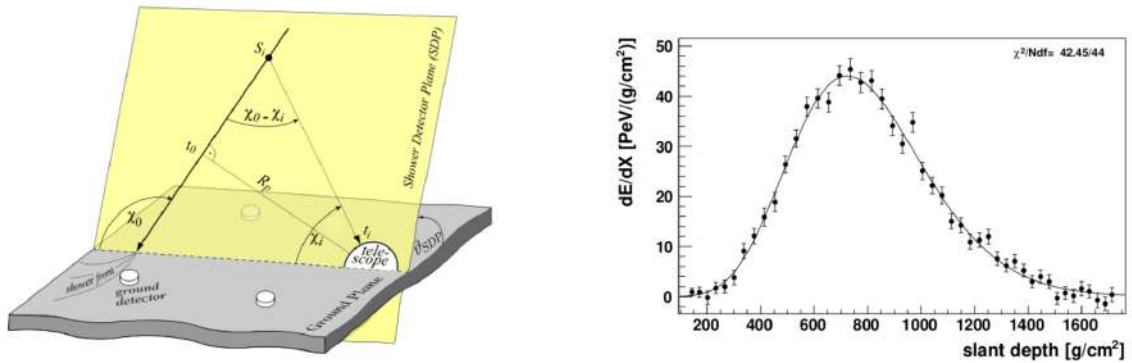


Figure 2.5: Left: Illustration of the geometrical shower reconstruction from the observables of the shower front. Right: Energy deposit profile as estimated using the FD. Both figures are taken from [41].

by the FD are actually hybrid events. A laser beam from the CLF (see figure 2.1), whose direction is known with excellent accuracy, can be used to measure the accuracy of the geometrical reconstruction.

Once the geometry of the shower has been determined, the light collected at the aperture as a function of time can be converted to energy deposit as a function of slant depth, a measure of the position in the atmosphere explained in section 3.1 (see figure 2.5, right). After estimating the attenuation from the shower to the telescope and disentangling all contributing light sources, one is left with the calorimetric energy [51]. Integrating along the shower and correcting for the “invisible” energy carried away by neutrinos and high energy muons, one obtains an estimator of the primary energy of the corresponding primary. Besides, from the energy deposit profile shown in figure 2.5 (right) an estimate of the position X_{max} of maximum development in the atmosphere can be obtained [52].

2.3. AugerPrime

As will be explained in section 3.5, there is a discrepancy in the number of muons obtained from simulations, using present hadronic interaction models, and the number of muons measured at the observatory. This means that there are uncertainties in the hadronic interaction models which introduce uncertainties in all analyses. In particular, the measurement of the mass composition of the cosmic ray flux, which is important to understand the origin of ultra-high energy cosmic rays, is affected by this. AugerPrime is a collection of upgrades to the Pierre Auger Observatory intended to improve the separation of muonic and electromagnetic components when measurements are carried out [53]. The key element of the upgrade is the installation of a plastic scintillator on top of almost all existing SD stations. Furthermore, enhanced SD electronics, addition of a small photomultiplier to the SD, installation of the underground scintillator muon detector AMIGA and new techniques to increase the duty cycle of the FD are part of this upgrade.

2.3.1. The Surface scintillator detector

Each surface scintillator detector (SSD) is a plate consisting of two modules of around 2 m^2 [54, 53] (see figure 2.6). Each module consists of 24 fibers which are 1.6 m long. These wavelength-shifting optical fibers are embedded in scintillator bars which have an outer reflective layer of TiO_2 intended to enhance reflectivity. These fibers collect the scintillation light produced by shower particles and are bundled towards a single PMT located in the central part of the module. One such detector is being deployed on top of almost every SD station.

The plastic scintillators and water-Cherenkov stations have different responses to the muonic and electromagnetic components. Muons have larger energy deposits in water than electromagnetic particles. At the same time, both components deposit on average the same amount of energy in the scintillator. This means, that the WCD is more sensitive to the muonic component, whereas the SSD is more sensitive to the electromagnetic component of the shower. Given these different sensitivities, a disentanglement of the electromagnetic and muonic components will be possible. This in turn will provide information on the muon content, on an event-by-event basis. Furthermore, the position of X_{max} can be estimated with a duty cycle of 100 %.

A so-called *matrix formalism* as developed in [55] for a layered surface detector can be adapted to be applied to this new setting. The motivation for this formalism is to relate intrinsic shower parameters at ground level, such as energy or particle fluxes, to the detector signals via a matrix whose coefficients depend only on the shower geometry but very little on the shower primary mass or on the interaction model used to describe it. After processing the detector signals, the matrix can be inverted in order to obtain the fluxes of the muonic component and the electromagnetic energy.

2.3.2. The AMIGA muon detector

The Auger Muon Detector for the Infill Ground Array (AMIGA) was designed for the direct measurement of the muonic component of extensive air showers [57]. It consists of an array of scintillators associated with the water-Cherenkov detectors from the denser SD-750 array and the SD-433 array, for which the distances between detectors are of 433 m. At each position, three 10 m² modules are buried close to the stations at a depth of around 2.3 m. The overburden of earth above the AMIGA detectors serves as a natural shielding for the electromagnetic particles of the shower and imposes a cut-off for vertical muons of 1 GeV [58]. A schematic overview of AMIGA is shown in figure 2.7. Each module consists of 64 scintillator strips. Light collected at each strip is guided, using wavelength shifting fibers, towards a 64 channel silicon photomultiplier located in the middle of the module where the signal is read.

The segmented structure of the scintillator module allows for direct counting of muons [53]. Muon counters sample scintillator signals at a frequency of 320 MHz,

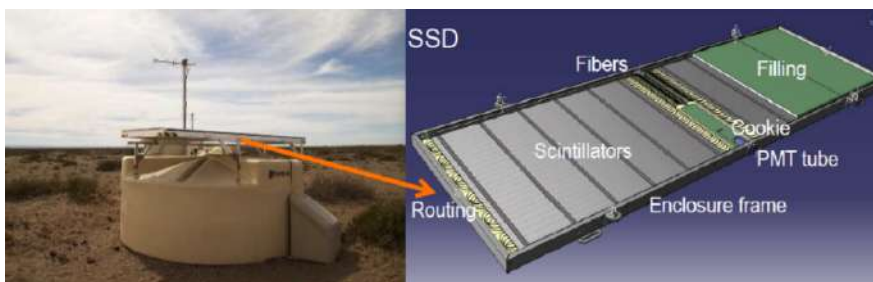


Figure 2.6: Left: Picture of one upgraded station. Right: Layout of the surface scintillator detector showing its different components. (Picture is taken from [56])

which means that every 3.125 ns 64 bits are acquired. Each bit stores the digitized value (either a “0” or “1” if the signal was below or above a predefined threshold) associated to one scintillator bar. Muon counting and the digitization of the integrated signal is implemented in the AMIGA electronics, including an FPGA with three main functional blocks: counting, data codification, and external communications.

Comparing the measurements obtained by the surface detectors with those from a small array of underground muon detectors should significantly improve the accuracy of the AugerPrime results [54]. AMIGA will provide important direct muon measurements of a sub-sample of showers. It will be possible to use these results for verification and fine-tuning of the methods used to extract muon information from the SSD and WCD measurements (see section 2.3.1).

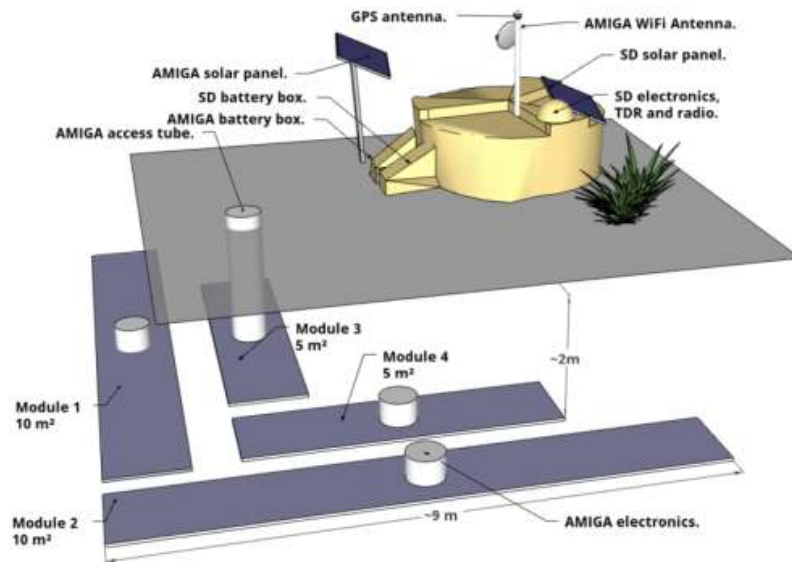


Figure 2.7: General overview of an underground muon detector from AMIGA. Both surface and underground detector, are shown in their arrangement during the prototype phase. In the final design, the 30 m² detector is splitted into three 10 m² modules. (Figure taken from [54])

Chapter 3

Air shower physics

“If you wish to make an apple pie from scratch, you must first invent the universe.”

— Carl Sagan

This chapter begins with a brief introduction to extensive air showers which are the object of study in the present thesis. This is followed by a description of the objective of this work. Subsequently, the main tool used in this work is described: the Monte Carlo simulation framework CONEX [59, 60, 61]. Afterwards, we move on to a more detailed description of extensive air showers following the semi-empirical model developed by Heitler [62] and Matthews [63, 64]. Those works concentrate on lower energy showers than the ones we study. Therefore, whenever it is suitable, we take a closer look into parameters and expressions in the context of extensive air showers generated by cosmic rays with energies above the knee region. We study the performance of current models using energies relevant to our work in order to know where improvement is needed (see chapter 4). Finally, the so-called muon puzzle [65] is presented, which is of utmost importance in the area of high-energy and air shower physics. For the present work, this issue needs to be kept in mind. Special attention is dedicated here to the related work of Cazon et al. [66, 67, 67] which inspired part of the reasoning used in our calculations.

3.1. Extensive air showers

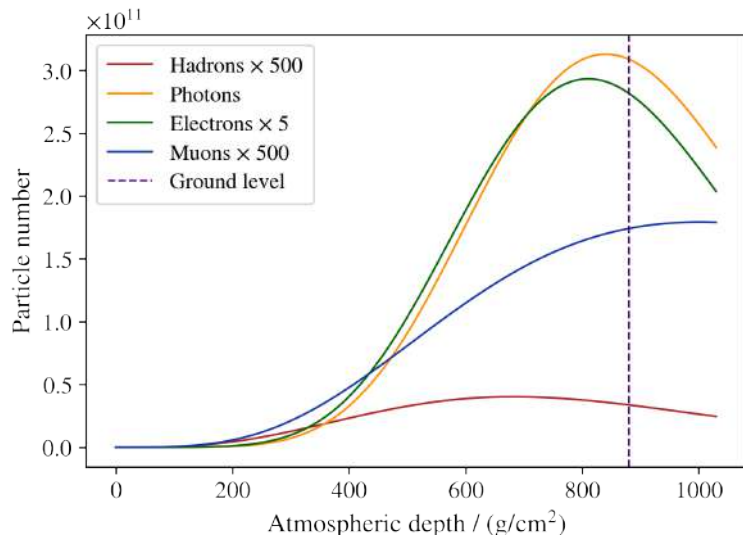
Experimental evidence so far indicates that the vast majority of cosmic rays are atomic nuclei [68, 69]. When such a nucleus reaches the Earth’s atmosphere it sets off an air shower through a hadronic interaction with a molecule or an atom in the upper atmosphere. In this interaction many new particles are created which are subject to the same process. The set of all the particles created this way in successive interactions constitutes an extensive air shower. If the showering process is observed as a function

of the atmospheric depth one obtains the longitudinal profile. The atmospheric depth, which is formally defined at the end of this section, measures the position in the atmosphere by means of amount of matter traversed. The approximate upper limit of the atmosphere corresponds to 0 g/cm^2 and the sea level to 1030 g/cm^2 . In figure 3.1 an example of such a longitudinal profile for a simulated shower is shown. The numbers of particles are displayed separated into different components which we now describe.

The hadronic component (full red line in figure 3.1), which mostly consists of pions, grows until the energy of individual pions falls to the level where they are more likely to decay before colliding again. The muonic component (full blue line in figure 3.1) is predominantly composed of muons that are created in these decays. This component decouples from the cascade because muons propagate to the ground with small energy loss and deflection. Furthermore, their decay length is $c\tau = 7.8 \text{ m}$ [70] which becomes of the order of a kilometer because of time dilation and the high energies involved. Consequently, they are unlikely to decay before reaching ground. For this reason, they carry information about hadronic interactions in extensive air showers. In this work the total number N_μ of muons at ground is used. In figure 3.1 the ground level of the Pierre Auger Observatory is represented by the dashed violet line at $\sim 880 \text{ g/cm}^2$. This corresponds to an altitude of around 1400 m.

In each hadronic interaction, around a third of the available energy is taken by neutral pions which instantly decay to a pair of gamma rays (their decay length is $c\tau = 25.5 \text{ nm}$ [70]). These photons originate electromagnetic subshowers. All these subshowers together build up the electromagnetic component. Due to the short interaction length of electromagnetic particles [70] any information on the distribution of the neutral pions is lost. In figure 3.1, the electromagnetic component is further separated into a distribution of photons (full yellow line) and a distribution of electrons and positrons (full green line). The latter are from now on collectively referred to as

Figure 3.1: Longitudinal profiles of the hadrons, photons, electrons and muons for a simulated shower initiated by a vertical proton of 10^{20} eV and simulated with the high-energy interaction model EPOS-LHC. Some components are scaled for visualization purposes. The ground level represents the depth at which the Pierre Auger Observatory is situated: $\sim 880 \text{ g/cm}^2$. The total number of muons at this depth will be called N_μ . At a depth of around $X_{max} = 800 \text{ g/cm}^2$ the maximum development of the shower is achieved.



electrons. Note that the vast majority of particles in a shower are of electromagnetic nature. This component grows in number until the energy per particle is low enough ($\sim 85\text{MeV}$) for these to be absorbed by the atmosphere. A maximum size of the shower is thus reached at a certain depth X_{max} in the atmosphere which is also object of study in this work.

The position of maximum development is given and the profiles in figure 3.1 are displayed as a function of *atmospheric depth*. An extensive air shower is mainly driven and determined by the successive interactions and for these what matters more than distance completed is the amount of matter traversed per unit area. This quantity is the atmospheric depth X also referred to as the *overburden*. It depends on the density of the medium ρ through:

$$X(l) = \int_l^\infty \rho(l') dl'$$

where l' describes the trajectory traversed [71]. For the case of a vertical trajectory and the *standard isothermal atmosphere* the following simple expression can be obtained:

$$X(h) = X_0 \cdot e^{-(h/h_s)}. \quad (3.1)$$

$X_0 \simeq 1030 \text{ g/cm}^2$ is the depth at sea level $X(h = 0)$, h is the height above sea level (in meters) and h_s is the mean *scale height* for the standard isothermal atmosphere. This last value depends on temperature, and thus on altitude, and is approximately 8,400 m at sea level. In this case the *barometric equation* is given by

$$\rho(h) = \rho_0 \cdot e^{-(h/h_s)}$$

where $\rho_0 = 0.00107 \text{ g/cm}^3$ is the density at sea level.

3.2. Objective of this work

An anticorrelation between the depth of maximum development X_{max} and the total number of muons at ground N_μ can be observed in simulations (see figure 3.2) and in real datasets. This anticorrelation can roughly and qualitatively be justified by the fact that if the energy is distributed along the shower in such a way that a higher fraction stays in the hadronic channel, then more muons can be created while less energy is left to build up the electromagnetic component giving a lower value of X_{max} (and vice versa). The fact that the anticorrelation spreads considerably sideways along the slope means that even having a specific repartition of the energy among the different components there is another phenomenon not captured by the argument just given.

The physical motivation of the present work is to understand the X_{max} - N_μ anti-

correlation. The work was driven by the question of how can its shape be explained quantitatively. We saw in section 1.2.4 how the first two moments of the X_{max} distribution can be used to study the cosmic ray mass composition. In section 3.6 we will show how to make use of the fluctuations in the number of muons. These and many other studies are being carried out in which characteristics of X_{max} or N_μ alone are used to extract information about extensive air showers. The purpose of the present work is to study the joint distribution of X_{max} and N_μ . The fact that these two observables are correlated means that analyzing their distributions simultaneously should render new information not accessible when considering them separately.

We begin by studying the X_{max} - N_μ anticorrelation (see figure 3.2, left) for simulations of vertical proton initiated showers with a primary energy of 10^{20} eV because the anticorrelation is most pronounced in this setting. At a later stage, however, we will consider other energies. Furthermore, at an initial phase it makes sense to study the distribution with X_{max} replaced by $\Delta X = X_{max} - X_0$ where X_0 is the depth at which the first interaction occurs (see figure 3.2, right). This removes the variability introduced by the randomness of the depth of first interaction X_0 . This way, the ΔX - N_μ distribution depends only on the processes inherent to the shower development. The goal is to find parameters that are important for the development of the shower and to develop a model which predicts X_{max} and N_μ as a function of these parameters in a way that their anticorrelation is reproduced.

We also investigate the anticorrelation for different primary masses. As can be seen

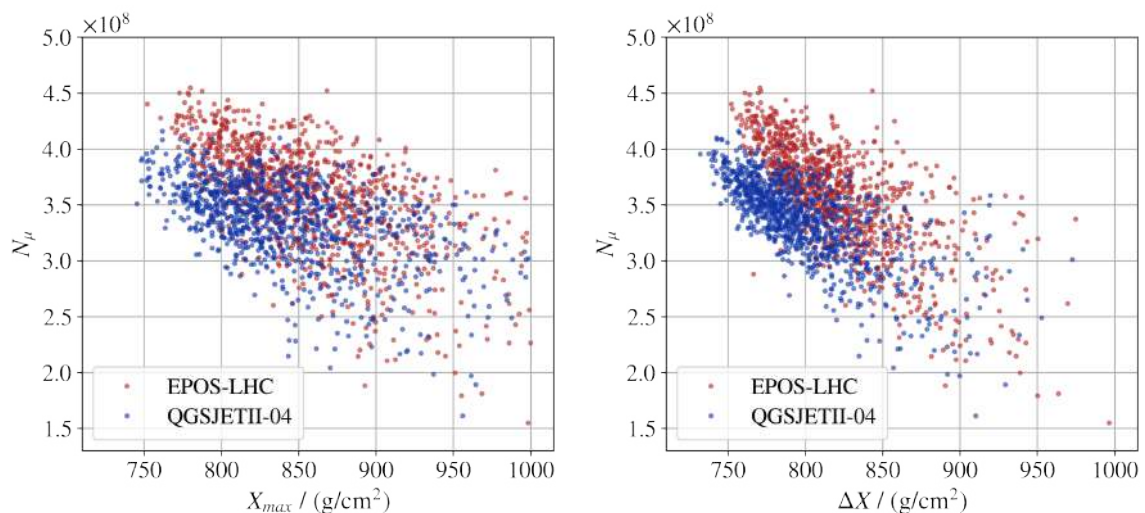


Figure 3.2: Anticorrelation between the depth X_{max} of maximum shower development and the number N_μ of muons at ground (left). Sets of 1000 showers generated by proton primaries of 10^{20} eV are used. One set is generated using EPOS-LHC (red points) and the other one using QGSJETII-04 (blue points). The anticorrelation is more pronounced when replacing X_{max} by $\Delta X = X_{max} - X_0$ (right) where X_0 is the depth at which the first interaction occurs. Furthermore, it is helpful to begin analyzing this last distribution keeping out the variability introduced by the point of first interaction.

in figure 3.3 (using X_{max} on the left and ΔX on the right) the distributions move to the upper left corner as the primary mass increases. The anticorrelation decreases also for increasing primary mass. At a later stage of the work, when we are interested in applying the newly gained knowledge to a dataset consisting of events detected at the Pierre Auger Observatory, we will need to take care of the fact that the cosmic ray flux comprises primaries ranging from protons to irons.

Every analysis of extensive air showers requires a detailed theoretical understanding and modeling of the cascade which develops in the atmosphere after the primary cosmic ray sets it off. This can be achieved combining knowledge on high-energy interactions with Monte Carlo simulations. Before applying our model to data, we operate with simulations carried out with the Monte Carlo framework CONEX [59, 60, 61]. Unless stated otherwise, all simulations used and mentioned in this work are carried out by us. As will be explained in section 3.3, these simulations can be performed using different interaction models. We use the high-energy interaction models EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d.

3.3. Monte Carlo simulations

We use the framework CONEX for Monte Carlo simulations of extensive air showers. Full Monte Carlo simulations as the ones that can be performed with the program CORSIKA are not a viable option. Unreasonably large computing time would be required for the primary energies we are interested in. Even applying weighted sampling

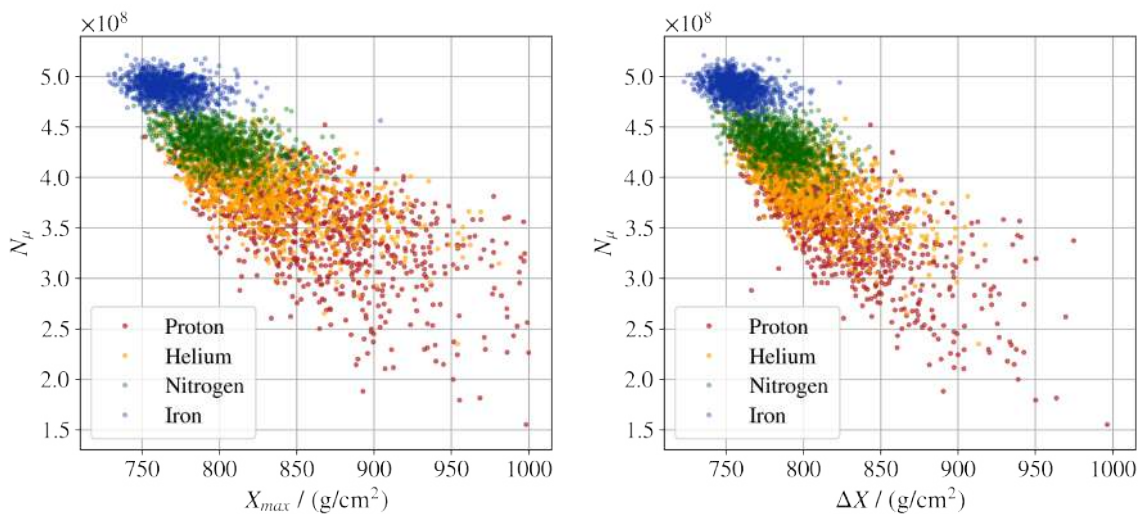


Figure 3.3: Anticorrelation between the depth X_{max} of maximum shower development and the number N_μ of muons at ground (left). Sets of 1000 vertical showers generated by different primaries of 10^{20} eV are used. These simulations are carried out using the high-energy interaction model EPOS-LHC. On the right, X_{max} is replaced by $\Delta X = X_{max} - X_0$. As the mass of the primary increases the anticorrelation is less pronounced.

algorithms included in the software, like the *thinning method* [72], are no solution. The computing time is reduced here by treating explicitly only a small portion of representative particles assigning them weights to account for the particles that are not tracked explicitly. The issue is that at the energies of interest for us the required level of thinning would introduce artificial fluctuations in the observables we need for our study [73].

In CONEX this problem is solved combining explicit Monte Carlo simulation, of the highest energy portion of the air shower (first interactions), with numerical expressions for the lower energy part [74, 75, 76, 77, 78]. An energy-threshold E_{thr} needs to be defined prior to simulation below which the particles are organized in bins of energy to be fed subsequently into the cascade equations which are solved numerically. Including the fluctuations introduced through the highest energy interactions ensures that one obtains accurate results for the fluctuations in the extensive air shower characteristics. Since the bulk of lower energy particles is large, particular characteristics are averaged out and there is no loss when dealing with these particles in a deterministic way using cascade equations [61] and thus, average extensive air shower parameters are accurate as well [79].

CONEX does not simulate showers in four dimensions as CORSIKA does. Instead it simulates their longitudinal development along the shower axis. As an output one gets the longitudinal profiles of the number of charged particles, muons, electrons, photons and hadrons above predefined energy thresholds and the energy deposit along the shower axis. From the muon profile we extract the total number of muons at ground N_μ , which we need for the present work. The depth of maximum development X_{max} , that we need for our study, is calculated and output by CONEX performing a fit of the Gaisser-Hillas function [80]

$$f_{GH}(X) = N_{max} \left(\frac{X - X_1}{X_{max} - X_1} \right)^{\frac{X_{max} - X_1}{\lambda}} \exp \left(- \frac{X_{max} - X_1}{\lambda} \right), \quad (3.2)$$

where $\lambda = p_1 + p_2 \cdot X + p_3 \cdot X^2$, to the longitudinal distribution of charged particles. The feature that makes CONEX ideal for studying the effect of hadronic interactions on the longitudinal shower features is an extension which enables the output of ROOT files [81] containing detailed information pertaining interactions and particles with energy above E_{thr} , such as particle identities, their energy and momentum, etc.

From these files we extract the information we need for our study. We are particularly interested in the first interaction, accessible via an *interaction counter*, which associates to each interaction an increasing number, starting from 1 for the first interaction. All particle identities, their energy and other properties can be accessed this way. It is also interesting to analyze the first leading interaction which follows the leading

particle from the first interaction and can also be extracted from this list. The *leading particle* is the secondary particle from the first interaction that holds the highest amount of energy. The value of the point of first interaction X_0 can also be obtained from CONEX simulations directly and should not be confused with the X_1 -parameter from the Gaisser-Hillas function (equation (3.2)) which in the literature is sometimes interpreted as the point of first interaction. We will also calculate effective parameters for the whole shower in the course of this work. The information necessary for this calculation can also be obtained scanning over all interactions with energy above E_{thr} . Since it is expected that the interactions with higher energy are the ones which contribute the most to the shower development it is sufficient to use only interactions above this energy threshold (as long as it is chosen appropriately).

The treatment of hadronic interactions below and above a threshold energy (typically around 100 GeV) is handled by separate models. For low energies we use URQMD [82]. The need to describe interactions at energies beyond the reach of colliders and the need to handle a variety of projectiles (nuclei, protons, charged pions and kaons) and targets (nitrogen, oxygen, argon molecules in the atmosphere) is specific to generators designed for high energies. We will analyze in the first part of the work showers generated using EPOS-LHC [83, 84, 85] and QGSJETII-04 [86, 87, 88]. It is interesting to use these two models because, as we shall see in later sections, distributions of certain physical characteristics are different between these two models and, as a consequence, the X_{max} - N_μ distributions are visibly different as well (see figure 3.2). At a later stage we also include simulations using SIBYLL-2.3d [89, 90, 91]. Even though its results are quite similar to those obtained with EPOS-LHC the small differences make it an interesting alternative to use.

The generators QGSJET and SIBYLL only implement phenomena that are important for shower development and need thus only a limited set of parameters (of the order of tens of parameters). EPOS includes much more processes and has therefore a broader applicability responding also to the needs of the high-energy physics and the heavy-ion collision communities. Its parameter set is larger (of the order of 100 parameters) and more datasets from collider experiments are used to constrain them. The mean values of X_{max} and the distributions of shower size and hadronic energy distribution have no significant difference between EPOS and QGSJET. On the counterpart, the muon number is considerably higher for EPOS, which can be explained by its higher (anti-) baryon production [92]. It is not clear as of today which model represents physical processes at the highest energies more truthfully. Therefore, it makes sense to use the three high-energy interaction models. Furthermore, having different scenarios from which to learn about the X_{max} - N_μ anticorrelation helps covering different possibilities and so understanding better the real relationships. For these reasons we incorporate all three scenarios into our more advanced stage of the work.

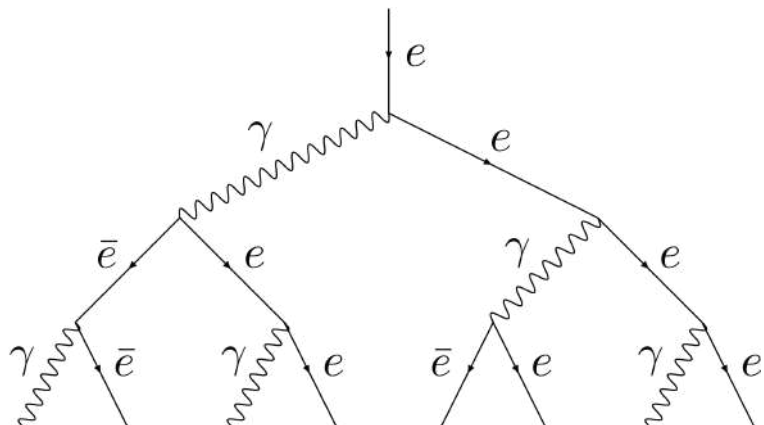
3.4. Semi-empirical model

When pioneering work on cosmic rays and their resulting showers was carried out, the fact itself of performing simulations was problematic. Computers were for a long time not ready to deal with the computing time and memory that are necessary to simulate extensive air showers. For this reason, already at an early stage, effort was put in the development of simpler but predictive analytical models of extensive air showers. Even today it is still imperative to have good models with different degrees of complexity in order to understand the essence of the phenomenon. Monte Carlo programs such as CORSIKA and CONEX take into account so many different stochastic processes that it is almost impossible to see at first sight the effect a given parameter or particular process has on the final resulting air shower. Furthermore, when designing a structure for a neural network special care needs to be taken. Since neural networks mostly behave in unpredictable ways, it is convenient to have a good previous understanding of which information is important and should be used as a *feature* from which to learn and which information is interesting and one wants to predict, the so-called *target*. It is also favourable to know which properties of the target variables are important and should be properly captured by the model. In this section we present the construction of a semi-empirical model which was designed for these purposes by Heitler [62] and Matthews [63, 64]. The calculations of formulas presented in this section are reproduced from [63, 64, 1, 2].

3.4.1. Electromagnetic showers

We begin with the Heitler splitting approximation of cascades initiated by an electromagnetic particle (see figure 3.4 for a representative diagram) before adding continuously more complexity to the model [62]. The initiating particle can be a photon or an electron. In this model particles are assumed to travel one splitting length $X_r = \lambda_r \ln 2$, where λ_r is the *radiation length*, and creating then new particles. Photons produce an

Figure 3.4: Schematic representation of the Heitler splitting model for electromagnetic cascades [62]. e represents an electron or positron and \bar{e} the corresponding anti-particle. An electromagnetic subshower is mainly driven by electron-positron pair-production induced by photons and bremsstrahlung experienced by electrons.



electron-positron pair and electrons undergo bremsstrahlung (see figure 3.4). Since the energy loss through radiation is described for electrons by $E(X) = E_0 e^{-X/\lambda_r}$, X_r is the distance over which they lose half their energy through radiation. This way the number of particles in the shower grows, while the primary energy E_0 is assumed to be distributed equally among all particles. As soon as individual energies drop below the *electron critical energy* ξ_c^e , electrons are more likely to lose energy through collisions with the atmosphere than radiating photons. From this stage on, particles start being absorbed by the atmosphere so that at this depth the total number of particles reaches a maximum. The values for the radiation length and the critical energy depend on the medium where the electrons of interest propagate. In this context, the values for air are needed, which are approximately $\lambda_r = 37 \text{ g/cm}^2$ and $\xi_c^e = 85 \text{ MeV}$.

If n_γ is the total number of splitting lengths traversed until the critical energy is reached, we obtain the maximum number of particles

$$N_{max}^\gamma = 2^{n_\gamma} = \frac{E_0}{\xi_c^e}.$$

Consequently, the maximum number of generations can be expressed as

$$n_\gamma = \frac{\ln(E_0/\xi_c^e)}{\ln 2}$$

and the depth at which the maximum number of particles is obtained is

$$X_{max}^\gamma = n_\gamma \lambda_r \ln 2 = \lambda_r \ln \left(\frac{E_0}{\xi_c^e} \right). \quad (3.3)$$

The *elongation rate* Λ is defined as the rate at which X_{max} changes with respect to the primary energy E_0 :

$$\Lambda = \frac{dX_{max}}{d \log_{10} E_0}.$$

Inserting the concrete values for this setting one obtains:

$$\Lambda_\gamma = \lambda_r \ln 10.$$

Here we recover the aforementioned qualitative properties that the total number of electromagnetic particles at the shower maximum is proportional to the energy carried by the primary and that the position of this shower maximum develops proportional to the logarithm of the primary energy. On the counterpart, a considerable defect this model has is that it predicts less photons than electrons while in truth the number of photons in any shower is much higher (as can be seen, for example, in figure 3.1). In practice this model for showers initiated by photons or electrons are mainly interesting as a description of the electromagnetic subshowers that build up the electromagnetic

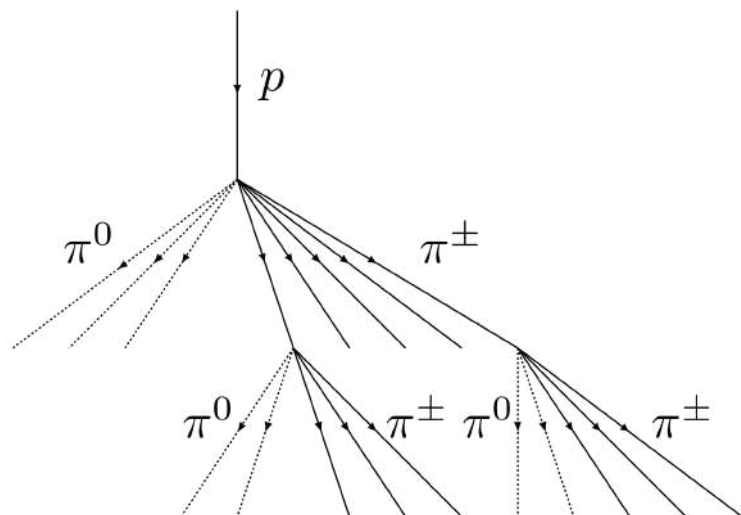
component of showers generated by protons or heavier nuclei. These are tackled next.

3.4.2. Proton initiated showers

A semi-empirical model for hadronic showers was presented by Matthews in 2001 [63, 64] (a schematic representation can be seen in figure 3.5). The case for primary protons has some similarities with photon induced showers. The interacting primary proton is assumed to generate a set of charged and neutral pions which traverse a splitting length $X_I = \lambda_I \ln 2$ where λ_I is the *interaction length* of strongly interacting particles (it is discussed in [1] that it makes more sense to take $X_I = \lambda_I$ without the factor $\ln 2$ since λ_I already is the interaction length).

In this model the primary proton and the subsequent charged pions generate new generations of N_{ch} charged and N_0 neutral pions. The neutral pions almost immediately decay to two photons which themselves originate electromagnetic subshowers as described in previous paragraphs. Again, the number of particles increases at the same time as individual energies, which are assumed to be distributed equally, decrease. When the *critical energy* of the charged pions ξ_c^π is reached, it is more probable for them to decay than to interact. Then they are all assumed to decay instantly to muons which are detected at ground. In this model the values $\lambda_I = 120 \text{ g/cm}^2$ and $\xi_c^\pi = 20 \text{ GeV}$ are used for interactions in air. Besides, multiplicities are reduced to the constant values $N_{ch} = 10$ and $N_0 = 5$, which means that the fraction of charged particles out of the total multiplicity N_{tot} is $f_{ch} = 2/3$. These constant values serve the purpose for a range of primary energies covering the knee region of the energy spectrum (10^{14} eV to 10^{17} eV). Whenever it is suitable, special attention will be dedicated to parameters that in the corresponding literature are described for the knee region. In those cases we present an analysis in the context of the present work, which deals with the region of the energy spectrum above 10^{17} eV .

Figure 3.5: Representative scheme of the model for an extensive air shower generated by a proton [63, 64]. Hadronic interactions are simplified and considered to generate only neutral and charged pions. The former decay almost immediately to a pair of photons which generate subsequently electromagnetic subshowers. The charged pions interact again producing a new set of pions. Full fermionic lines represent charged pions and dotted fermionic lines represent neutral pions. Only a few evocative lines are shown.



If n_p is the number of generations of new particles until individual pion energies $E_\pi^{ind}(n_p)$ reach the critical energy ξ_c^π then

$$\xi_c^\pi = E_\pi^{ind}(n_p) = \frac{(f_{ch})^{n_p} E_0}{(N_{ch})^{n_p}}. \quad (3.4)$$

Here $(f_{ch})^{n_p}$ is the fraction of the primary energy that stays within the group of charged pions after n_p layers. This energy needs to be distributed among $(N_{ch})^{n_p}$ particles in the last generation when all the charged pions decay to muons. As a consequence

$$n_p = \frac{\ln(E_0/\xi_c^\pi)}{\ln(N_{ch}/f_{ch})}$$

and the number of muons at ground is

$$N_\mu^p = N_\pi(n_p) = (N_{ch})^{n_p} = \left(\frac{E_0}{\xi_c^\pi}\right)^\beta \quad (3.5)$$

where

$$\beta = \frac{\ln(N_{ch})}{\ln(N_{ch}/f_{ch})}. \quad (3.6)$$

This time, the depth X_{max}^p at which the maximum number of particles is obtained needs to be calculated differently compared to electromagnetic showers, because in this scenario different types of particles are present in the shower. As can be intuited, for example from figure 3.1, the number of electromagnetic particles is much higher than the number of pions. As a consequence, X_{max}^p must depend heavily on the bulk of electromagnetic subshowers and more specifically on the most influential ones. These are the ones initiated by the highest energy neutral pions, i.e. by the group of neutral pions from the first interaction. Thus, a simple expression for the maximum depth is

$$X_{max}^p = X_0 + \lambda_r \ln\left(\frac{E_0/(3 \cdot 2N_0)}{\xi_c^e}\right) = X_0 + \lambda_r \ln\left(\frac{E_0/(3 \cdot N_{ch})}{\xi_c^e}\right) \quad (3.7)$$

where $X_0 = \lambda_I^p(E_0) \ln 2$ is the depth where the first interaction occurs and $E_0/(3 \cdot 2N_0)$ represents the fact that in the first interaction one third of the primary energy stays among the neutral pions which produce $2N_0$ photons. Furthermore, the maximum number of electromagnetic particles is incorporated in the following expression which simply represents energy conservation:

$$E_0 = \xi_c^e N_{max}^p + \xi_c^\pi N_\mu^p. \quad (3.8)$$

The first addend is the energy which ends up in the electromagnetic component and the second addend is the energy that stays in the hadronic component and is carried by

muons after full development of the shower. Lastly, the elongation rate can be expressed as a function of the one corresponding to electromagnetic showers of the same energy:

$$\Lambda_p = \Lambda_\gamma + \frac{d}{d \log_{10} E_0} (X_0 - \lambda_r \ln(3N_{ch})). \quad (3.9)$$

When calculating X_{max}^p this way, all the lower energy electromagnetic subshowers are not taken into account, which causes an underestimation of the value. Furthermore, this expression for X_{max}^p is very sensitive to the values of X_0 and N_{ch} which fluctuate from shower to shower, in particular, because the most realistic value for N_{ch} to be used here is in fact the multiplicity of charged pions created in the first interaction. In the next paragraph we show that these values are immensely variable. Still, expression (3.9) for the elongation rate captures quite well results obtained from simulations. One point to make here is that up to now X_0 and N_{ch} are treated as constant values but they actually depend on energy: X_0 decreases and N_{ch} increases with increasing energy [70]. This means that the elongation rate depends on physical properties of showers that for high energies are not available from experiments. Thus, this property is important for interpretation of experimental results. It also reflects Linsley's *elongation rate theorem* which states that the elongation rate for electromagnetic showers represents an upper limit to the elongation rate of hadronic showers [93].

We now explore the just presented parameters for energies relevant in the context of the present work using our sets of simulations. Concerning the values of the effective

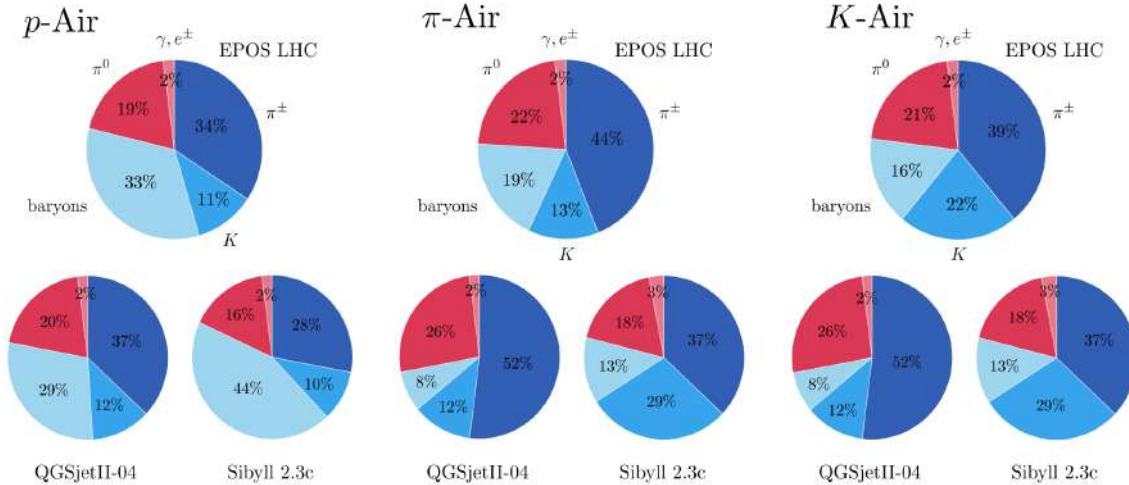


Figure 3.6: Fractions of the principal groups of particles (charged pions, kaons, baryons, neutral pions and electromagnetic particles) created in p -Air, π -Air and K -Air interactions at 10^{19} eV for different high-energy interaction models (EPOS-LHC, QGSJETII-04, SiBYLL-2.3c, reproduced from [66]). The blue slices represent groups of particles that mainly contribute to the hadronic channel and the red ones represent particles that mainly divert energy to the electromagnetic component. Consequently, it makes sense to assign the former group to the number N_{ch} and the latter group to the number N_0 .

multiplicities N_{ch} and N_0 , it is most sensible to assign them geometric mean values. The geometric mean of the numbers x_1, x_2, \dots, x_n is

$$\langle (x_1, x_2, \dots, x_n) \rangle_{geom} = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}$$

and is used, instead of the usual arithmetic mean, when the values are meant to be multiplied together or are exponential in nature. The former is the case here since the cascade process is a multiplicative one. Numbers of particles are obtained by multiplying the values from the previous generations. Furthermore, extensive air showers in truth contain much more particles than only charged and neutral pions. In figure 3.6 the fractions of charged pions, kaons, baryons, neutral pions and electromagnetic particles are shown for different interactions and high-energy interaction models (reproduced from [66]). We need to define the multiplicities N_{ch} and N_0 taking into account that there are these particles other than pions, which is not considered explicitly in the literature. Pion fractions vary from only 44 % up to 78 %. All the other particles can be assigned to the group of charged or to the group of neutral pions, depending on their most probable decay products, i. e. depending on whether they divert energy directly to the electromagnetic component or keep energy in the hadronic channel.

Eta mesons (η) decay extremely rapidly mainly to neutral pions or directly to photons [70]. Therefore, we add them to the number N_0 . Kaons behave similarly to pions. The decay lengths for charged (K^\pm), short-lived (K_S^0) and long-lived kaons (K_L^0) are $c\tau = 3.7$ m, $c\tau = 2.7$ cm and $c\tau = 15.3$ m, respectively [70], and their most numerous products are charged pions or directly muons. This means that the interplay between interaction and decay happens differently than for charged pions but their contribution is similar. For simplicity we add them to the number N_{ch} . Baryons are mostly protons and neutrons and some lambdas (Λ). Lambdas decay rather quickly

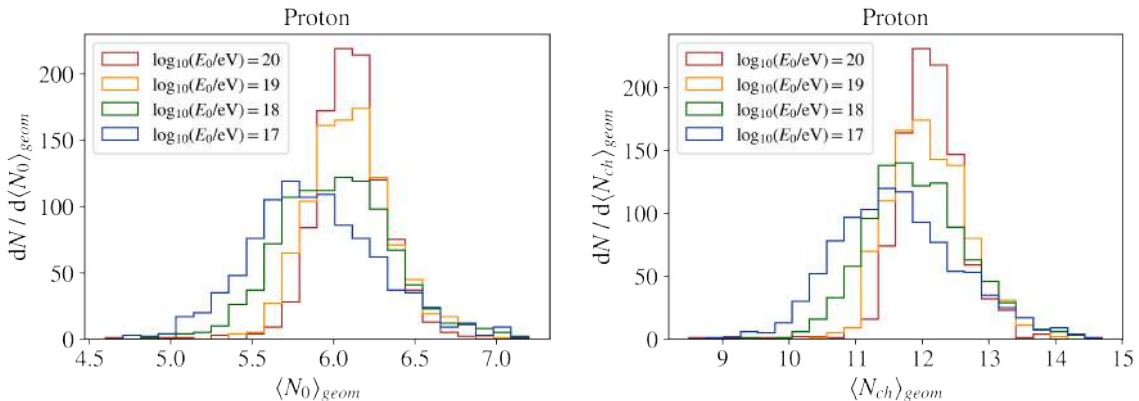


Figure 3.7: Distributions of the geometric mean of the numbers N_0 (left) and N_{ch} (right) for sets of 1000 showers initiated by vertical protons and simulated with the high-energy interaction model EPOS-LHC for different primary energies.

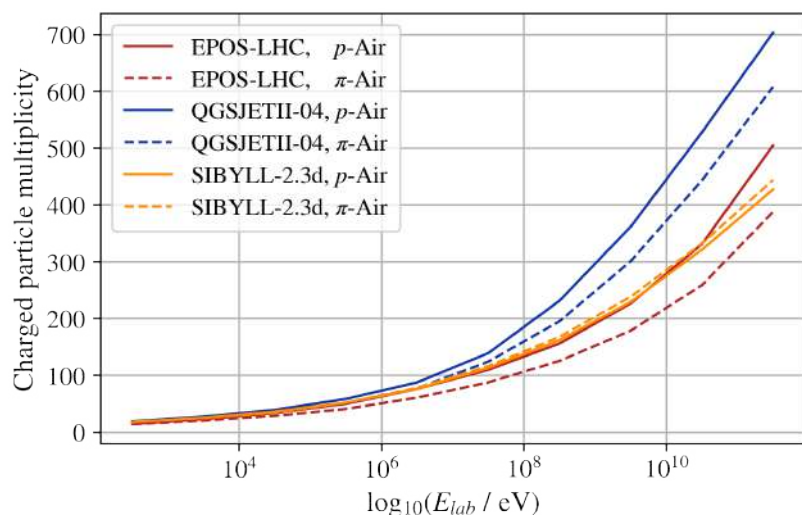
($c\tau = 7.9$ cm [70]) and mainly into protons, neutrons and pions. Nucleons in turn interact creating all the types of particles already mentioned. As a result, we add baryons to the number N_{ch} .

In figure 3.7 distributions of the geometric means of the numbers N_0 (left) and N_{ch} (right) for different energies of interest in the present work are shown. In each case 1000 simulations for vertical protons are used and the high-energy interaction model was EPOS-LHC. The distribution's mean tends to grow with energy (5.89, 6.02, 6.09, 6.07 for $\langle N_0 \rangle_{geom}$ and 11.64, 11.93, 12.08, 12.03 for $\langle N_{ch} \rangle_{geom}$) and the peaks get narrower (standard deviations 0.4, 0.37, 0.33, 0.41 for $\langle N_0 \rangle_{geom}$ and 0.94, 0.82, 0.77, 0.94 for $\langle N_{ch} \rangle_{geom}$). Consequently, the values are a bit higher than those used for the knee region (constant values 5 and 10 are used in [63]). This increase is a consequence of the mean charged particle multiplicities in pion-air collisions. These are shown together with the ones from proton-air collisions in figure 3.8 as calculated with EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d (reproduced from [94]). In addition, it becomes clear that within the shower the multiplicity varies heavily depending on the energy of the interaction. As a consequence, multiplicity varies noticeably along the shower. A possible explanation for the slight decrease at the highest energy might be that having higher multiplicities in the first interactions causes particles quickly to have very low energy and to lead then to very low multiplicities very soon.

The correlation between $\langle N_0 \rangle_{geom}$ and $\langle N_{ch} \rangle_{geom}$ for the same sets of simulations is shown in figure 3.9 where the black line corresponds to $y = 1.98x$. Here the factor $c_0^{ch} = 1.98$ is the common best fitting value for each individual energy group. In this sense we can carry over the factor 2 from Heitler and Matthews to higher energies.

It is worth noting that when comparing the multiplicities within individual showers, like the example shower in figure 3.10 (left), this factor c_0^{ch} can be different from 2. It is well known that, under the assumption of isospin invariance, the quark-parton model

Figure 3.8: Mean charged particle multiplicity in p-air and π -air collisions as calculated with EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d as a function of energy in the laboratory system (reproduced from [94]). The difference in the predicted secondary particle multiplicity increases with energy. The multiplicity of neutral pions is closely linked to that of charged particles and hence shows qualitatively the same behavior.



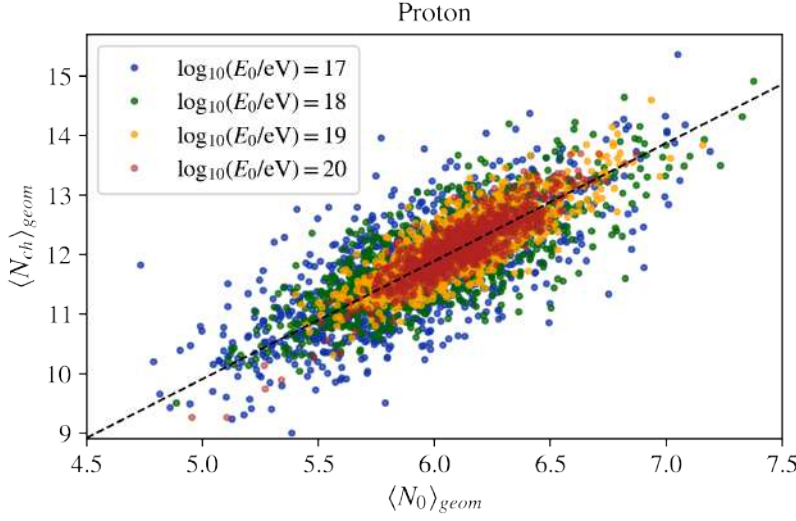


Figure 3.9: Correlation between the geometric means of N_0 and N_{ch} from figure 3.7. The black line corresponds to $y = 1.98x$. Therefore, in certain contexts, it is safe to assume that there are 2 times more hadronically interacting secondaries than secondaries delivering electromagnetic particles.

predicts that the multiplicity for neutral pions is equal to the average of those for positively and negatively charged pions [95]. This is a consequence of the assumption that the probability for a certain quark to fragment into a π^0 is equal to the average of the probabilities for it to fragment into a π^+ or a π^- , because the quark content of the π^0 is the same as the average of the contents of the π^+ and the π^- :

$$\pi^0 = \frac{1}{\sqrt{2}} (u\bar{u} - d\bar{d}),$$

$$\pi^+ = u\bar{d}, \quad \pi^- = d\bar{u}.$$

However, we are adding kaons and baryons to the charged pions in order to obtain the number N_{ch} , while N_0 only includes η particles in addition to the neutral pions. Consequently, considering that pions constitute the majority of the hadrons produced

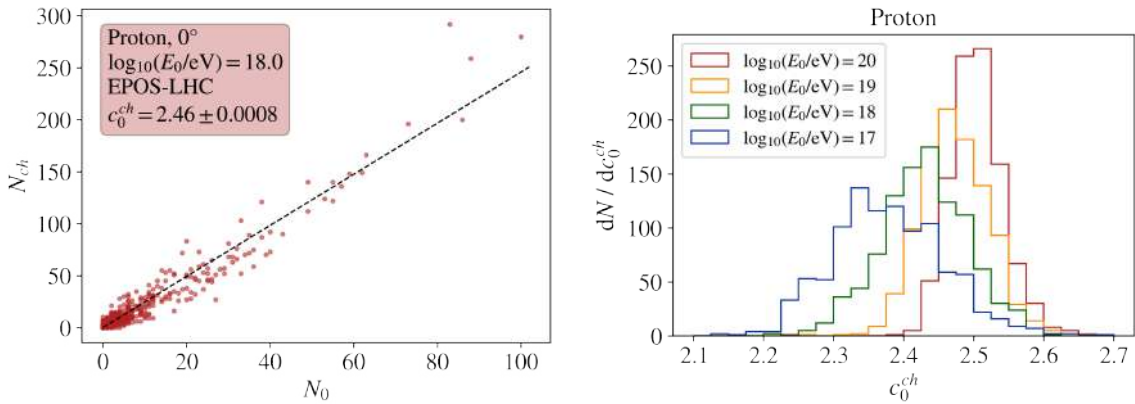


Figure 3.10: Left: Relationship between the multiplicities N_0 and N_{ch} from all the interactions (above $E_{thr} = 0.005 \times E_0$) present in a typical proton initiated shower of $E_0 = 10^{18}$ eV simulated using the high-energy interaction model EPOS-LHC. It has on average $c_0^{ch} = 2.46 \pm 0.0008$ times more particles that contribute to the hadronic channel than to the electromagnetic channel. Right: Distributions of the factor c_0^{ch} for the different groups of simulations from figure 3.7.

in air showers, it is understandable that for the shower in figure 3.10 (left), when inspecting the multiplicities N_0 and N_{ch} from all the interactions (above an energy threshold of $E_{thr} = 0.005 \times E_0$ for this set of showers), a factor of $c_0^{ch} = 2.46$ is found which is close to but higher than 2. It can be seen that when keeping only the lower energy interactions, values closer to 2 are obtained. This is a consequence of the fact that for lower collision energies the kaon-to-pion ratio is lower [96]. In figure 3.10 (right) the distributions of the factor c_0^{ch} , obtained like in figure 3.10 (left) for the showers from figure 3.7, are shown. The mean of this factor grows with energy as would be expected from the behaviour of the kaon-to-pion ratio (2.38, 2.43, 2.47, 2.50) and its standard deviation decreases (0.08, 0.06, 0.05, 0.04). Obtaining factors higher than 1.98 is in no contradiction with the value we obtained when considering the geometric means of the multiplicities because those represent values that correspond to lower energy interactions where pions are dominant. Depending on the application, one might use either the factor 1.98 from the previous paragraph or the energy-dependent factor presented here.

Concerning the interaction length of strongly interacting particles, it is too simplistic to use a constant value for it. To begin with, when looking at the depth of the first interaction $X_0 = \lambda_i^p(E_0)$ (figure 3.11, left), one sees that a large range of values is possible for fixed values of energy and primary mass. But also the interaction lengths along the development of the shower vary heavily. We can see in figure 3.11 (right) that the inelastic cross section for pion-air interactions can vary from around 200 mb to almost 600 mb, with small differences between the different high-energy interaction models (reproduced from [94]). Using that the interaction length λ_I is related to the

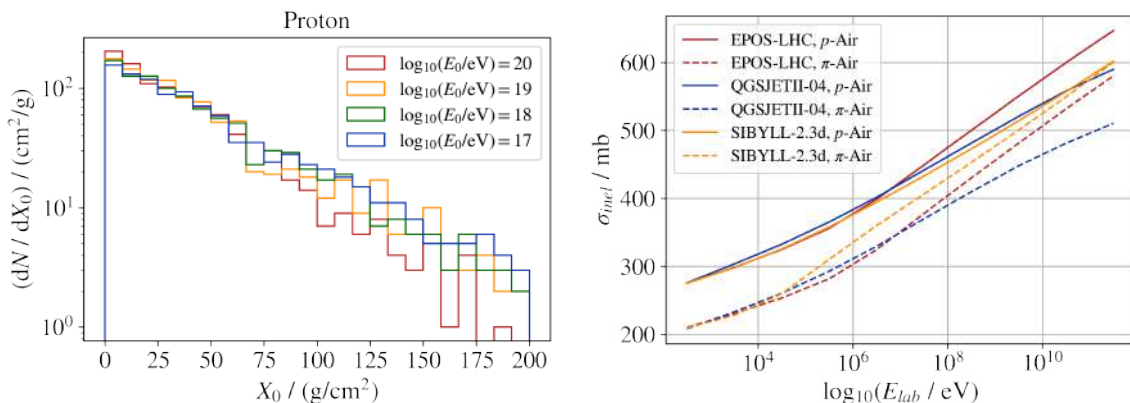


Figure 3.11: Left: Distributions of the depth of first interaction for vertical showers initiated by protons simulated with the high-energy interaction model EPOS-LHC for different energies. Right: Proton- and pion-air inelastic cross section as a function of energy in the laboratory system (reproduced from [94]). The inelastic cross section is defined as the cross section for all collisions in which at least one new particle is produced. It can be written as $\sigma_{inel} = \sigma_{tot} - \sigma_{el} - \sigma_{qel}$ where σ_{tot} is the total cross section and σ_{el} and σ_{qel} are the elastic and quasielastic cross sections, respectively.

particle inelastic cross-section σ_{inel} via the expression

$$\lambda_I = \frac{\langle m_{air} \rangle}{\sigma_{inel}}$$

[97] and that the effective mass number of air is 14.8, we obtain that the interaction length can vary from $\lambda_I = 41 \text{ g/cm}^2$ for the highest energies to $\lambda_I = 124 \text{ g/cm}^2$ for the lowest energies considered in figure 3.11. This means that the mean interaction length is noticeably different at different stages of the shower development. At the same time, for each particular energy, taking into account that the interaction length is exponentially distributed, a broad range of values can arise.

There is no unique way to estimate the critical energy ξ_c^π because different degrees of complexity can be considered. In Montanus' work energy dependent values of the particle multiplicity and the interaction length are taken into consideration [1]:

$$N_{tot}(E) = 0.15 \cdot (E/\text{eV})^{0.18},$$

$$X_0(E) = 145 \text{ g/cm}^2 - 2.3 \text{ g/cm}^2 \cdot \ln(E/\text{eV}),$$

$$\lambda_I(E) = 200 \text{ g/cm}^2 - 3.3 \text{ g/cm}^2 \cdot \ln(E/\text{eV}).$$

Considering that the energy after the i -th interaction point is

$$E_i = \frac{E_{i-1}}{N_{tot}(E_{i-1})}$$

an expression for the individual energies after the n -th generation can be obtained:

$$E_n = 6.7^{\alpha_n} \cdot E_0^{\beta_n}$$

where

$$\alpha_n = \frac{1 - 0.82^n}{1 - 0.82}, \quad \beta_n = 0.82^n.$$

Then the depth

$$X_n(E_0) = X_0(E_0) + \sum_{i=1}^n \lambda_I(E_i), \quad n \geq 1$$

is where the $(n + 1)$ -th generation occurs.

Inverting expression (3.1) we get that the difference in height (in meters) between the n -th and the $(n + 1)$ -th interaction is

$$\Delta h = h_s \ln \left(\frac{X_n(E_0)}{X_{n-1}(E_0)} \right).$$

Assuming that the pions decay when the decay length is half the layer thickness leads

to

$$\gamma c\tau_{\pi^\pm} = \frac{1}{2}\Delta h.$$

Remembering that $\gamma = E_n/m_{\pi^\pm}$ where γ is the Lorentz-boost, an expression solely depending on E_0 and n is obtained here. For each possible number of generations \hat{n} , after which the critical energy is obtained by individual pions, the corresponding value of E_0 can be calculated from this expression numerically. The critical energy is then

$$\xi_c^\pi = \frac{E_0}{(N_{tot}(E_0))^{\hat{n}}}.$$

The results from [1] are shown in figure 3.12 (left). From left to right the dots correspond to $\hat{n} = 1, 2, 3, 4, 5$. The decrease with energy makes sense because more energetic primaries induce deeper showers. This means that particular individual pion energies are obtained deeper in the atmosphere for higher E_0 . Deeper positions in turn imply denser atmospheric layers which means that the interaction probability gets higher. Thus, in order to decay, more energy needs to be “lost” first.

A less detailed approach is possible where no numerical calculations are needed. In the work of Kampert and Unger [2] the interaction and decay length are equalized:

$$\lambda_I = \lambda_{dec}. \quad (3.10)$$

The decay length is

$$\lambda_{dec} = \rho(h)\gamma c\tau_{\pi^\pm}, \quad (3.11)$$

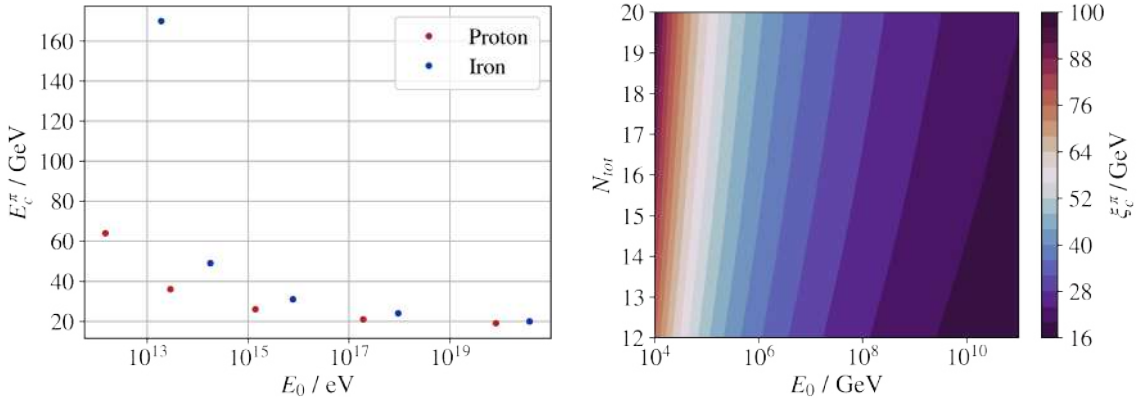


Figure 3.12: Critical energy as a function of primary energy as calculated in [1] for proton and iron primaries (left) and as a function of primary energy and effective total multiplicity as calculated in [2] for proton primaries (right). The first approach is more detailed and complex than the second one but the results are quite similar when considering a fixed value of N_{tot} around 15.

where ρ is the height-dependent density of air and

$$\gamma = \frac{E_0 / (N_{tot})^n}{m_{\pi^\pm}}.$$

Here the primary energy is merely equally divided among N_{tot} particles in each interaction. If θ is the angle of incidence, λ_I can be obtained from

$$\cos(\theta) = \frac{\rho(h)h_s}{X} = \frac{\rho(h)h_s}{n\lambda_I}. \quad (3.12)$$

Combining (3.10), (3.11) and (3.12) leads to

$$n_d(N_{tot})^{-n_d} = \frac{h_s}{c\tau_{\pi^\pm}} \frac{m_{\pi^\pm}}{E_0} \frac{1}{\cos(\theta)},$$

where n_d is the number of generations necessary for the individual pions to reach their critical energy. Equivalently we can write

$$-n_d \ln(N_{tot}) e^{-n_d \ln(N_{tot})} = -\frac{h_s}{c\tau_{\pi^\pm}} \frac{m_{\pi^\pm}}{E_0} \frac{\ln(N_{tot})}{\cos(\theta)}.$$

Finally, the number of generations n_d can be expressed as:

$$n_d = -\frac{W_{-1}\left(-\frac{h_s}{c\tau_{\pi^\pm}} \frac{m_{\pi^\pm}}{E_0} \frac{\ln(N_{tot})}{\cos(\theta)}\right)}{\ln(N_{tot})}$$

where W_{-1} denotes the lower branch of the *Lambert-W* function [98].

The Lambert-W function is defined as the inverse function of $f(x) = xe^x$. This mapping is not injective and has therefore no unique inverse. It can be seen that the Lambert-W function has two real branches. For the calculation of n_d the bottom branch W_{-1} is chosen because it is defined in $[-e^{-1}, 0]$ and $-n_d \ln(N_{tot}) e^{-n_d \ln(N_{tot})} \in [-e^{-1}, 0]$

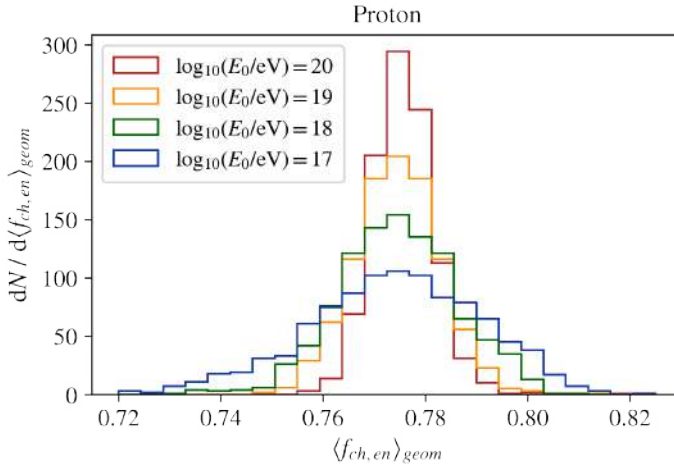


Figure 3.13: Distributions of the geometric mean of the fractions of energy $f_{ch,en}$ taken by hadronically interacting particles for the different groups of simulations from figure 3.7. More than two thirds of the available energy is taken by these particles in every energy group. This parameter is needed in order to improve the calculation of the critical energy in the formalism given in [2].

(since $n_d \ln(N_{tot}) < e^{n_d \ln(N_{tot})-1} = 1 + (n_d \ln(N_{tot}) - 1) + \dots$). The critical energy is obtained as before via

$$\xi_c^\pi = \frac{E_0}{(N_{tot})^{n_d}}$$

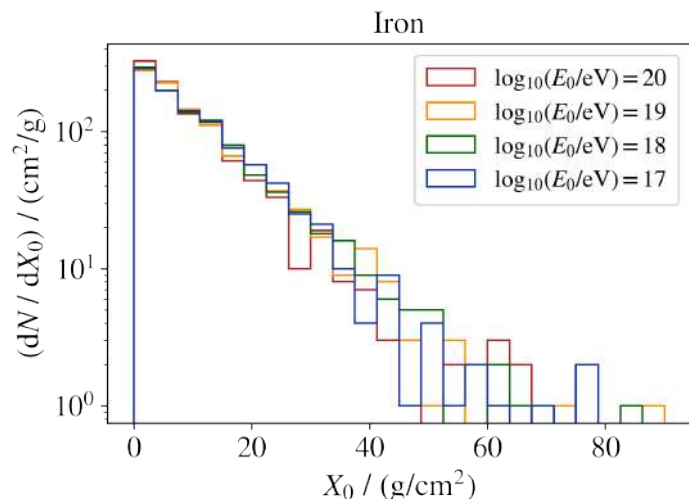
and shown in figure 3.12 (right) as a function of primary energy and effective total multiplicity.

In this semi-empirical model it is assumed that in each interaction the energy is divided equally among all secondaries. But, as is shown in figure 3.13 for extensive air showers generated by vertical protons of different energies, this is not true. Since these fractions are intended to be multiplied, in order to get the fraction $f_{ch,en}$ that stays in the hadronic channel, we calculate again a geometric mean. We have already seen that approximately two thirds of the particles created in an interaction are hadronically interacting particles. But more than two thirds of the available energy are taken by these particles (close to 0.77 for all primary energies with standard deviations of 0.01, 0.01, 0.012 and 0.017 for decreasing energy). This fact will be taken into account in our calculations in chapter 4 of the critical energy with a similar approach to that of Kampert and Unger.

3.4.3. Hadronic showers

When a heavy nucleus enters the atmosphere, it interacts rather quickly. The median of the depth of first interaction for the showers initiated by iron primaries is close to 7 g/cm² for all energy groups (see figure 3.14, for proton showers this value is around 30 g/cm², see figure 3.11, left). In this first collision only some of the nucleons interact inelastically with some nucleus in the air molecule producing subshowers involving pions. Several other nucleons and light nuclear fragments may be released while there will mostly be one heavy fragment. In order to generalize the model from section 3.4.2

Figure 3.14: Distributions of the depth of first interaction for sets of 1000 showers initiated by vertical iron primaries and simulated with the high-energy interaction model EPOS-LHC for different primary energies. Even though showers generated by lower energy particles have on average a depth of first interaction deeper in the atmosphere, having 56 of them counteracts this fact to the point that iron induced showers start earlier in the atmosphere than proton initiated ones (of the same energy).



to nuclear primaries with atomic number $A > 1$ the *superposition model* is applied where an even distribution of energy is assumed [63, 64]. As a result, the distribution of the depth of first interaction is supposed to be the same as if the nucleons had entered the atmosphere separately. Even though this is a considerable simplification, the model is adequate for many purposes. The assumption made is that a nucleus with atomic number A and primary energy E_0 behaves like A individual and independent nucleons of energy E_0/A . Hence, inserting in former formulas the energy E_0/A and summing up A showers where it is suitable, the corresponding formulas for a nuclear shower can be obtained:

$$X_{max}^A(E_0) = X_{max}^p(E_0) - \lambda_r \ln(A), \quad (3.13)$$

$$N_{max}^A(E_0) = N_{max}^p(E_0), \quad (3.14)$$

$$N_{\mu}^A(E_0) = N_{\mu}^p(E_0) \cdot A^{1-\beta} \quad (3.15)$$

$$\Lambda_A(E_0) = \Lambda_p(E_0). \quad (3.16)$$

We examine now the same parameters as in the previous section but for the case of iron primaries using our sets of simulations and our interpretation of the parameters. The distributions of the geometric means of the multiplicities N_0 and N_{ch} are shown in figure 3.15 (top and bottom). The mean values increase with energy as happens for proton primaries: 5.08, 5.61, 5.84, 5.94 for $\langle N_0 \rangle_{geom}$ and 10.02, 11.04, 11.53, 11.73 for $\langle N_{ch} \rangle_{geom}$. It is to be expected that the peak corresponding to 10^{20} eV is close to the one corresponding to 10^{18} eV for proton primaries (see figure 3.7) because the shower generated by the iron primary can be taken as 56 showers generated by nucleons of $(10^{20}/56)$ eV. The same argument applies to the 10^{19} eV iron showers. In addition, these distributions are narrower than the ones corresponding to proton showers because having 56 lower energy showers induces an averaging effect on parameters. The relation between the geometric means of the multiplicities N_0 and N_{ch} gives for all energy groups a factor of 1.97 times more hadronically interacting particles. This means that also here the factor 2 from Heitler and Matthews can be carried over. The distributions of the factor c_0^{ch} for iron showers are presented in figure 3.16. If a value more representative of individual showers is needed, the factor 1.97 should be replaced by the corresponding median value of c_0^{ch} : 2.29, 2.36, 2.41 and 2.46 for $\log_{10}(E_0/\text{eV}) = 17, 18, 19, 20$ iron showers. We suggest median values because of the skewness of the distributions. Also here are the distributions narrower than for proton primaries because of the averaging effect of having 56 lower energy subshowers.

In figure 3.14 the distributions of the depth of first interaction for the showers initiated by iron primaries of different energies are shown. The range of possible values is smaller than for proton primaries but still considerable. Even though showers generated

by lower energy particles have on average a depth of first interaction deeper in the atmosphere, having 56 of them counteracts this fact to the point that iron induced showers start earlier in the atmosphere than proton initiated ones.

The geometric means of the fractions of energy that stay in the hadronic channel are very similar to the ones corresponding to proton primaries (see figure 3.17, all mean values are close to 0.775). This means that also for iron primaries more than two thirds of the energy stay in the hadronic channel. Finally, the critical energy for iron induced showers, as calculated in [1], is shown in figure 3.12 (left). It makes sense that the values are higher than for proton primaries because of the lower energy nucleons being involved. In order to obtain values like in [2] one can use figure 3.12 reading the value corresponding to $E_0/56$.

We see in equation (3.13) that according to Matthews' model X_{max} is shifted upwards in the atmosphere by an amount of $\lambda_r \ln(A)$. The multiplicity of the first in-

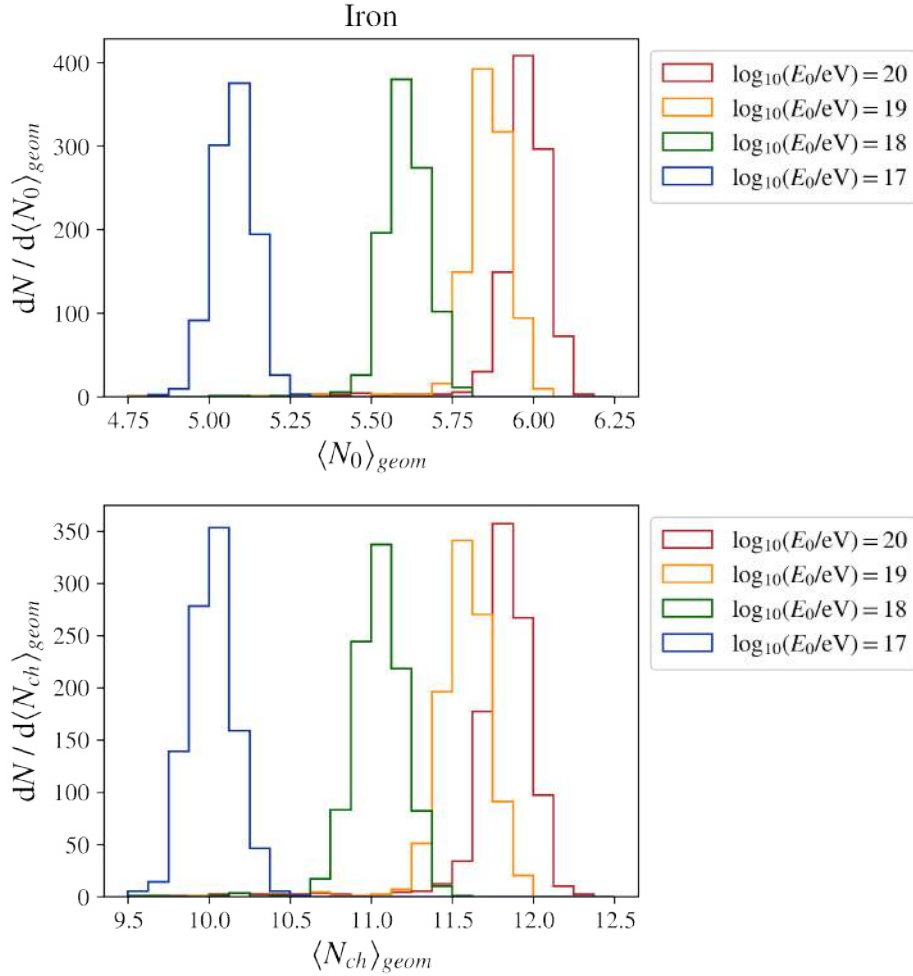


Figure 3.15: Distributions of the geometric mean of N_0 (top) and N_{ch} (bottom) for the showers from figure 3.14. These values are calculated the same way as for figure 3.7. It is as for proton primaries, in certain contexts, safe to assume that in iron showers there are two times more hadronically interacting secondaries than secondaries contributing to the electromagnetic component.

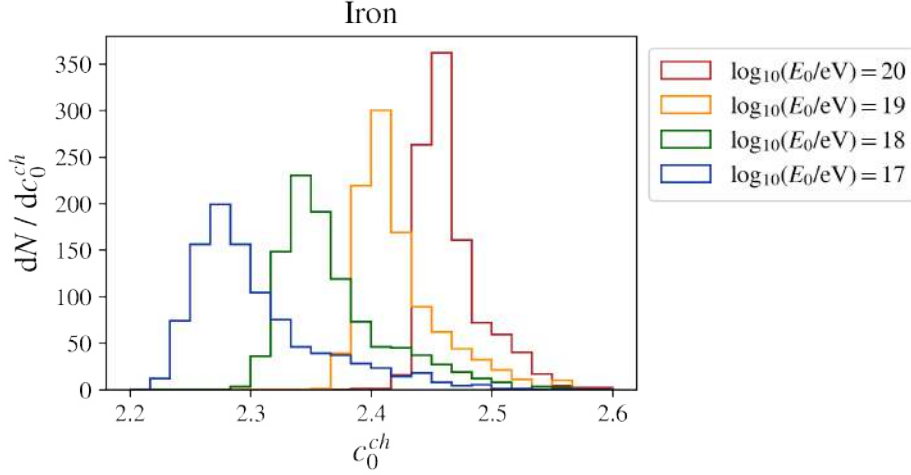


Figure 3.16: Distributions of the factor c_0^{ch} for the different groups of simulations from figure 3.14 as calculated for figure 3.10. As for proton primaries these values are slightly above 2 and can be used when values representative of individual showers are needed.

teraction is higher for iron primaries than for proton primaries. This can roughly be justified using following expression from [1]:

$$N_{tot}(E_0) = A \cdot 0.15 \cdot \left(\frac{E_0}{A}\right)^{0.18} = A^{0.82} \cdot 0.15 \cdot E_0^{0.18}$$

and considering that $A^{0.82} \geq 1$. Qualitatively expressed, both the smaller interaction length and the larger multiplicity reduce the depth of maximum shower size with respect to a proton initiated shower. In figure 3.18 the calculated values of X_{max} for proton and iron primaries of ultra-high energy (filled circles) are compared to the values of X_{max} obtained from our simulations done with CONEX (stars). We use expression (3.7) for the calculation of $X_{max}^p(E_0)$. For each of the 1000 simulated showers in each energy group the corresponding geometric mean of the multiplicities in that shower and the corresponding X_0 value are inserted. The markers are the median values of

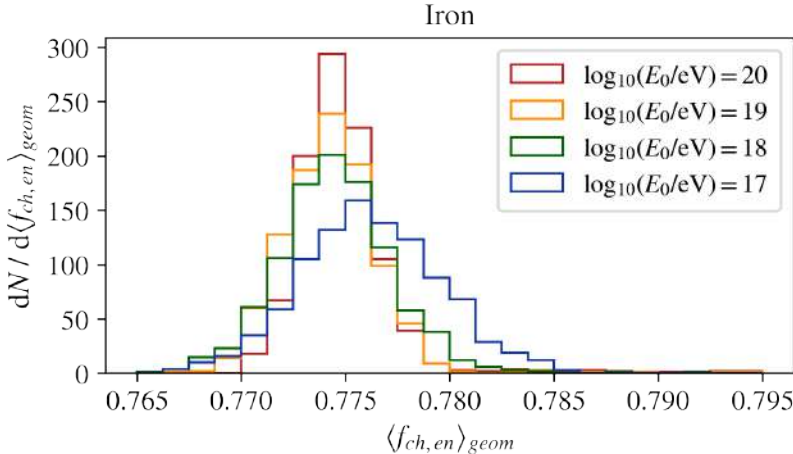


Figure 3.17: Distributions of the geometric mean of the fractions of energy $f_{ch,en}$ taken by hadronically interacting particles for the simulations from figure 3.14. Also for iron primaries more than two thirds of the available energy are taken by these particles.

the respective distributions and the bars the standard deviations for the calculated distributions. The agreement is for some combinations of energy and primary mass quite good but for others there is room for improvement. Furthermore, the elongation rate differs between simulations and calculations but is equal when comparing proton and iron simulations, as predicted by expression (3.16). For visibility of this feature, fitted lines are plotted as well. For proton primaries the slope is 57.8 g/cm^2 per decade of energy and for iron primaries it is 58.8 g/cm^2 per decade of energy.

From equation (3.6) and since $\ln(f_{ch}) \leq 0$ we can conclude that $0 < \beta \leq 1$. So, following equation (3.15), the number of muons is higher for higher atomic number (for fixed primary energy E_0). This can also be observed in figure 3.19 where the muon numbers at ground for our simulated showers and calculated with equation (3.5) are shown. In addition to the individual effective multiplicities this time also the critical energy is needed. We calculate it with the method from [2]. Qualitatively, the observed behaviour makes sense because, if the showers generated by each nucleon start with less energy, the shower is fully developed after less generations and thus less energy is “lost” to the electromagnetic component and more energy is available for muon production. Quantitatively, the predictions are quite good but, as shall be seen, there is room for improvement.

According to Matthews’ model, N_{max} is equal for different primaries of equal energy. In figure 3.20 values calculated through expression 3.8 are presented together with

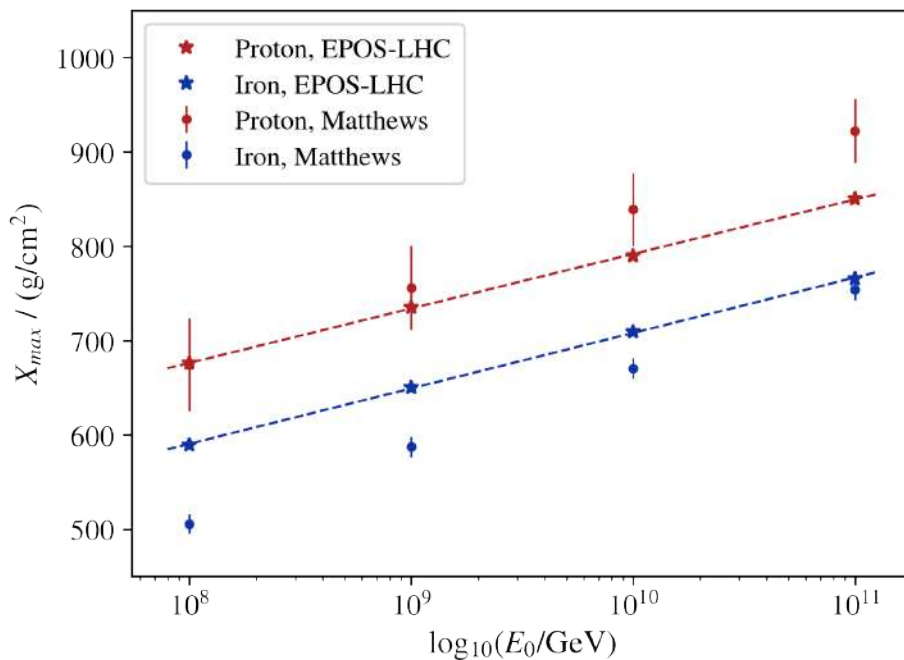


Figure 3.18: Comparison between X_{max} calculated using Matthews’ model (filled circles) and its values from our simulations (stars). Sets of 1000 simulations for vertical protons are used. The dashed lines are fitted to the simulated median values. They make the elongation rate, which is very similar when comparing different primary types, visible.

values from our simulations. In this expression ξ_c^π is calculated again according to [2] and the previously calculated values of N_μ (figure 3.18) are used. Effectively, values

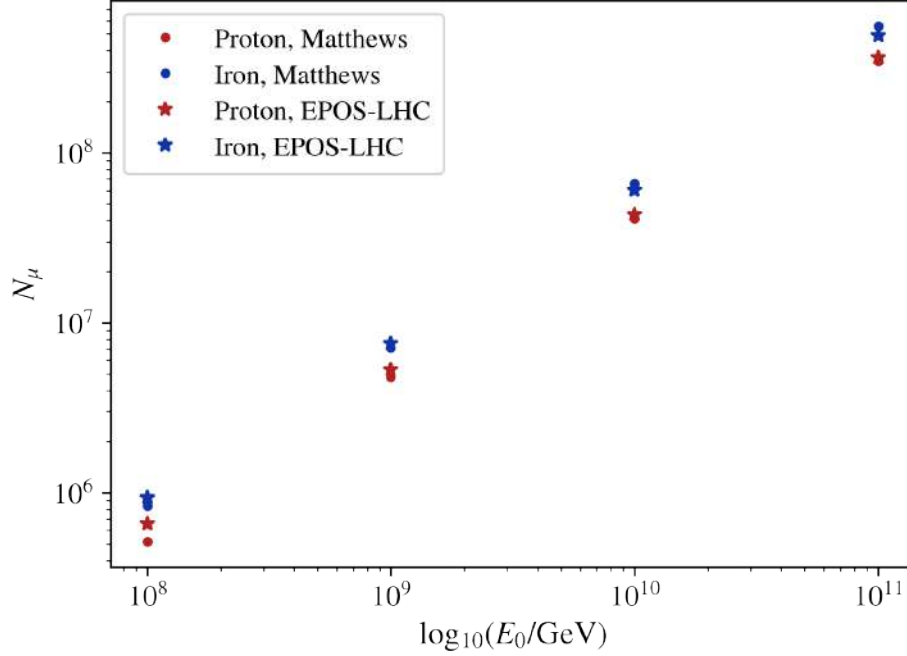


Figure 3.19: Comparison between N_μ calculated using Matthews' model (filled circles) and its values from our simulations (stars). The same simulations as in figure 3.18 are used. According to calculations, there are $A^{1-\beta}$ times more muons in showers generated by primaries with atomic number A than in proton showers.

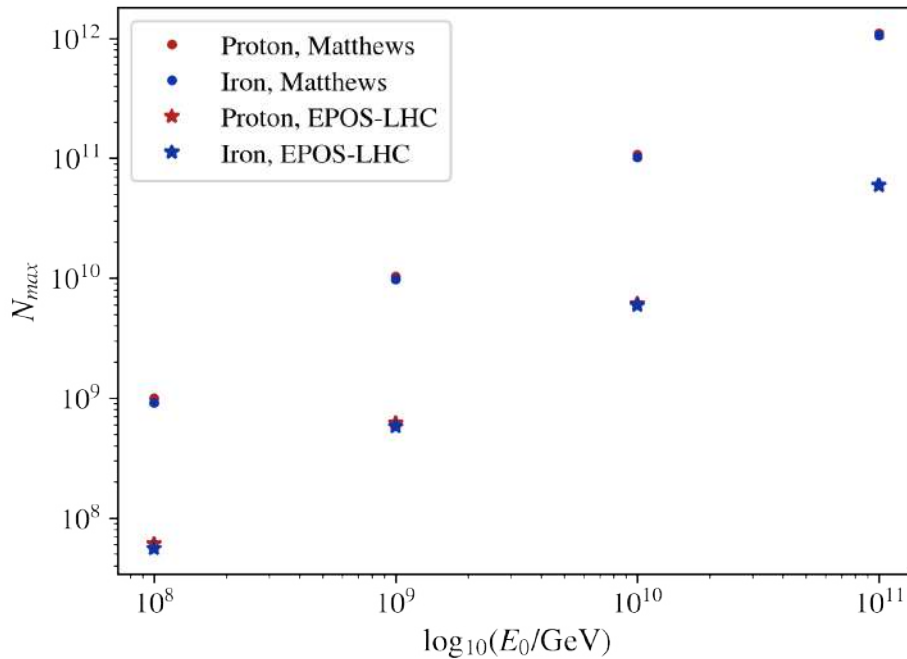


Figure 3.20: Comparison between N_{max} calculated using Matthews' model (filled circles) and its values from our simulations (stars). The same simulations as in figure 3.18 are used. According to calculations, the maximum shower size does not depend on the primary's mass.

of different primaries of same energy are very close to one another for simulated and calculated values. However, there is a gap of more than an order of magnitude between values from simulations and calculated ones.

3.4.4. Leading particle effect

An important phenomenon that is not taken into account up to now is the *leading particle effect*. In each interaction one *leading particle* keeps a considerable amount of the total energy which is then not available for production of new particles. Instead, this particle travels an interaction length, which means that its energy is at disposal just at a later stage. The fraction of energy that is indeed directed into production of new particles (in this case charged and neutral pions) is the *inelasticity* κ , while the energy that is carried by the leading particle is the *elasticity*. In this scenario, the energy is not evenly distributed among particles belonging to the same generation but there are groups of particles that are created by different combinations of leading particles and “generic” ones. The mean elasticities for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d are summarized in figure 3.21 (reproduced from [94]) for a wide range of energies.

Considering then, that a fraction $(1 - \kappa)$ of the available energy E is kept by the leading particle, $\frac{2}{3}\kappa E$ is distributed among N_{ch} charged pions and the rest ($\frac{1}{3}\kappa E$) among N_0 neutral pions, one gets a slightly modified expression for β :

$$\beta = \frac{\ln(1 + N_{ch})}{\ln\left(\frac{1 + N_{ch}}{1 - \frac{1}{3}\kappa}\right)} \quad (3.17)$$

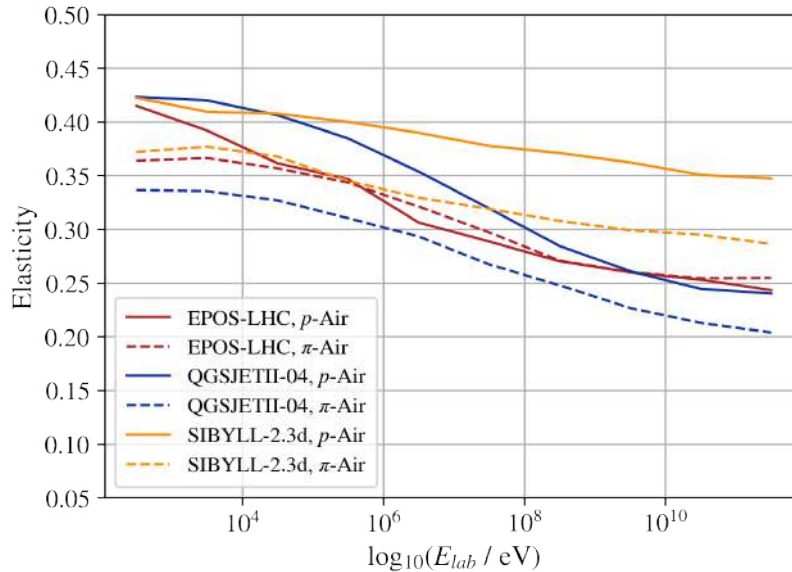


Figure 3.21: Mean elasticity in proton-air and pion-air collisions as predicted by EPOS-LHC, QGSJETII-04 and SIBYLL-2.3.d (reproduced from [94]).

which is obtained almost the same way as (3.6) and can still be inserted in expression (3.5) in order to get N_μ^p . However, the number of generations is accordingly modified:

$$n_p^{lp} = \frac{\ln(E_0/\xi_c^\pi)}{\ln\left((1 + N_{ch})/(1 - \frac{1}{3}\kappa)\right)}. \quad (3.18)$$

Here $(1 + N_{ch})$ is the number of charged pions with the addition of one leading particle. The difference with how (3.6) is obtained is the fact that here the energy is not equally distributed among the charged pions and the leading particle. But this approximation is adequate because it can be seen that a considerable fraction of the low-energy particles has similar energy. Obtaining new expressions for X_{max}^p and N_{max}^p is in this context not trivial. Now less energy is at disposal for the group of highest energy neutral pions, which would mean a smaller value of X_{max}^p as calculated earlier. But having a leading particle means that a considerable amount of energy is at disposal at a later stage, which could mean an important shift of X_{max}^p to higher depths. What can be said concretely is that considering the leading particle effect means that $\kappa < 1$ which raises both the values of β and n_p^{lp} (see equations (3.17) and (3.18)). This implies a deeper shower but the number of muons is not automatically raised because ξ_c^π is higher for iron primaries.

3.5. The Muon Puzzle

Air shower simulations with state-of-the-art QCD models show a significant muon deficit with respect to measurements. This was already reported in the year 2000 by the HiRes/MIA collaboration [99] and it was first established by the Pierre Auger Collaboration using nearly model-independent measurements of a set of air showers at ultra-high energies, initiated by cosmic rays with zenith angles between 62° and 80° [100]. This open problem is called the *Muon Puzzle* and caused a re-analysis of existing air shower data and oriented measurements. In order to compare results from measurements carried out under different conditions (zenith angle of air showers, lateral distance of the muon density measurement, energy threshold for the muons that are measured) the z -scale is introduced [65]:

$$z = \frac{\ln(N_\mu^{det}) - \ln(N_{\mu_p}^{det})}{\ln(N_{\mu_{Fe}}^{det}) - \ln(N_{\mu_p}^{det})}.$$

Here N_μ^{det} is the muon density measured by the detector, while $N_{\mu_p}^{det}$ and $N_{\mu_{Fe}}^{det}$ are the simulated muon densities for proton and iron initiated showers where detector effects are taken into account. If there was no discrepancy between real and simulated showers

the z -scale for a certain mass composition would be given by

$$z_{mass} = \frac{\langle \ln(A) \rangle}{\ln(56)} \quad [101].$$

In this case the z -scale would vary from 0 for a pure proton-flux to 1 if all the showers were initiated by iron primaries. After applying cross-calibration, in order to take any energy-scale offsets into account, the difference $\Delta z = z - z_{mass}$ is studied in [65]. Here z_{mass} is computed from the *Global Spline Fit* model [102]. In this model, the cosmic ray flux is divided into four mass groups, which cover roughly equal ranges in logarithmic mass $\ln(A)$. The differential flux of the leading element of each group is parameterized by a modified spline curve. In this process, measurements of the flux of individual elements in the low energy regime carried out by satellites and balloon experiments are combined with indirect measurements of mass groups obtained in air shower experiments. Subtracting z_{mass} from z is expected to remove the effect of the changing mass composition. The conclusion is that muon measurements seem to be consistent with simulations based on the latest hadronic interaction models up to about 10^{17} eV (see figure 3.22). However, at higher energies a growing muon deficit in the simulations is observed. It is noticeable that the results for different experiments are fairly consistent in this behaviour.

The two principal observables used to infer the mean logarithmic mass, $\langle \ln(A) \rangle$, are X_{max} and N_μ . X_{max} is mainly connected to the electromagnetic component. The most important processes that build up this component are pair production, annihilation, bremsstrahlung, Moeller, Bhabha and Compton scattering, which are well understood and can be calculated from first principles [97]. Indeed, no large discrepancies are observed in the values of X_{max} between measurements and simulations. But N_μ stems directly from the hadronic cascade which is mainly driven by relativistic heavy-ion collisions with nitrogen and oxygen atoms under low momentum transfer in the non-perturbative regime of quantum chromodynamics. Hadron production under these conditions cannot be calculated directly from first principles. Consequently, effective theories and phenomenology need to be used. The different high-energy interaction models that are used in simulation frameworks implement different approaches. These are tuned to the newest available accelerator data, which do not cover all the energy and phase-space regions needed. Hence extrapolations are necessary. These are the largest source of uncertainties. As a result, one of the dominant sources of systematics in $\langle \ln(A) \rangle$ originates from the uncertainty in the high-energy interaction models. Furthermore, different astrophysical models predict a different evolution of $\langle \ln(A) \rangle$ with energy. As long as the uncertainty for $\langle \ln(A) \rangle$ is too large many astrophysical theories can not be excluded.

Up to now it was not possible to solve the Muon Puzzle by tweaking the parameters

in the high-energy interaction models. The necessary changes would either ruin the consistency between simulations and other air shower observables, such as X_{max} , or violate constraints imposed by accelerator data. This suggests that some physical effect is missing in the generators. Rare events do not influence air shower development and are thus not interesting in this context.

On the one hand, the highest attainable energy for proton-proton collisions at the *Large Hadron Collider* (LHC) is ~ 13 TeV, which corresponds to an incident cosmic ray energy of almost 10^{17} eV. The cosmic ray energy spectrum extends at least three orders of magnitude further. This means that for ultra-high energy cosmic rays the first few interactions, which are the ones with the highest available energies, can not be understood from accelerator data. The energy range where the muon deficit is observable is only partially covered by experimental data, which suggests that the origin of the Muon Puzzle is likely to be found in the first stages of the shower, where the highest energies occur. This makes a non-exotic explanation plausible in which a relatively small change in the hadronic interactions at the highest energies propagates throughout generations causing a considerable change in the final muon number.

However, for example for primary protons with $E_0 = 10^{15}$ eV the most probable energy of the *grandmother* particle of any muon arriving at ground is within the range

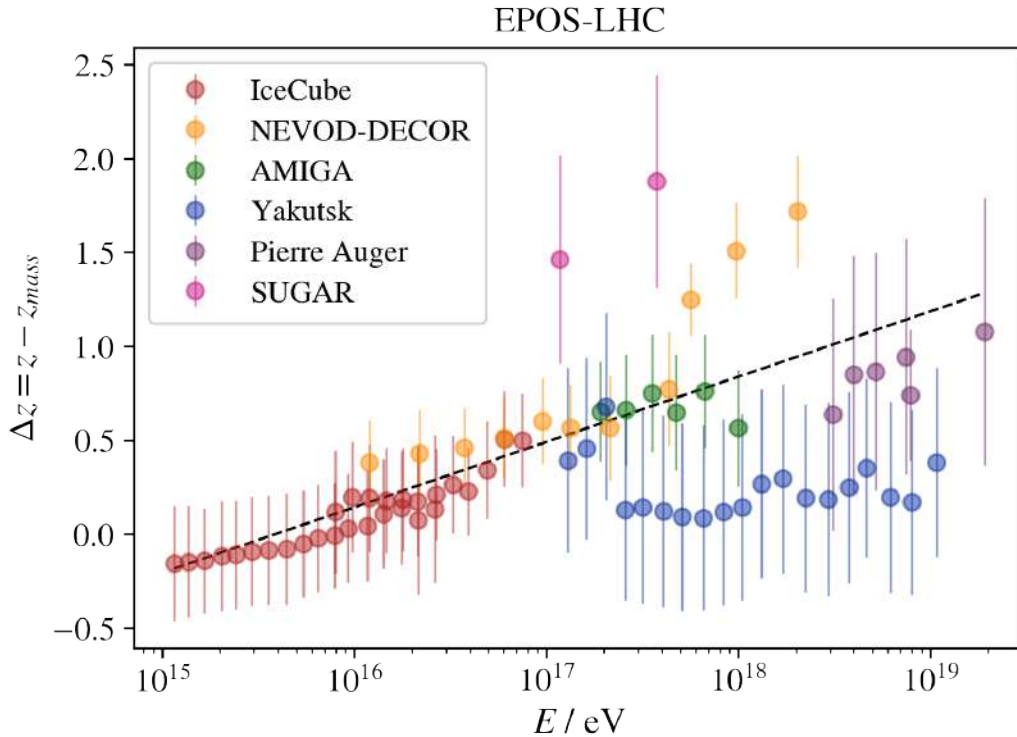


Figure 3.22: $\Delta z = z - z_{mass}$ for EPOS-LHC for different air shower experiments, after applying cross-calibration (reproduced from [65]). A consistency between measurements and simulations is observed up to about 10^{17} eV. At higher energies a growing muon deficit in the simulations is evident.

of beam energies of fixed experiments [103]: about 100 GeV for vertical showers and several 100 GeV for 60° inclined showers. The grandmother particle is the hadron inducing the last hadronic interaction that leads to a meson - called *mother* particle - which decays into the corresponding muon. This means that the number of muons in extensive air showers is sensitive to hadronic multiparticle production at low energy, as well. This is confirmed in [104], where it is shown that even at ultra-high shower energies the predictions on the lateral distribution of shower particles, including muons, depend strongly on the applied low-energy interaction model.

On the other hand, most experiments at the LHC are capable of measuring in the mid-rapidity region, where new heavy particles like the Higgs boson are best detected. However, the air shower development is dominated by the subshowers generated by the particles that carry the highest energies. Besides, muons in air showers heavily depend on the development of these hadronic subshowers. Now, the highest energies are carried by forward-produced particles. The relevant phase-space starts at pseudo-rapidities of about $\eta \geq 5 - 6$.

In addition, not all combinations of projectiles (nuclei, proton, charged pions and kaons) and targets (nitrogen, oxygen, argon) interesting for air shower simulations are represented in collider experiments. The theoretical approach that describes nuclear collisions in EPOS-LHC and QGSJETII-04 is an extension of the *Gribov-Regge Field Theory* [105]. In this approach the nucleus-nucleus scattering amplitude is defined by the sum of contributions of diagrams corresponding to multiple scattering processes between parton constituents of projectile and target nucleons. In SIBYLL-2.3d this is accomplished by means of the *Semi-superposition Model* [106]. This model refines the already mentioned superposition model by treating a collision of a nucleus with A nucleons on a nucleus with mass B as A proton- B collisions [107]. Here the primary energy is equally divided among nucleons, while the proton- B cross-section is based on a Glauber calculation [108]. This means that the depth of first interaction for each nucleon is distributed along the trajectory of the shower axis. There is still considerable theoretical uncertainty when extrapolating from proton-proton collisions to proton-air interactions [109].

A first approach to address the Muon Puzzle is to investigate the relationship between important physical parameters and extensive air shower observables. A first connection was shown in sections 3.4.1 to 3.4.4 with the support of the Heitler-Matthews Model. The inelastic cross-section σ_{inel} , the effective total multiplicity N_{tot} , the elasticity $(1 - \kappa)$ and the fraction of electromagnetically interacting particles (predominantly neutral pions) $f_{em} = N_0/N_{tot}$ turned out to be interesting parameters. Ulrich et al. introduced an *ad-hoc* model for air shower simulations in which these parameters are modified after each individual interaction and implemented it in the CONEX framework [111]. A particular high-energy interaction model is used as a baseline (for example

EPOS-LHC). The individual hadronic interaction features just mentioned are altered by the factor

$$f(E, f_{19}) = 1 + (f_{19} - 1)F(E)$$

which depends on the energy E of the colliding hadron in the frame where the target is at rest and where

$$F(E) = \begin{cases} 0 & E \leq 1 \text{ PeV} \\ \frac{\log_{10}(E/1\text{PeV})}{\log_{10}(10\text{EeV}/1\text{PeV})} & E > 1\text{PeV}. \end{cases}$$

The factor $f(E, f_{19})$ is 1 below 1 PeV, where the models are constrained by accelerator data and thus no modifications are implemented. Above 1 PeV the modification $F(E)$ increases logarithmically with energy, which reflects the increasing uncertainty of the

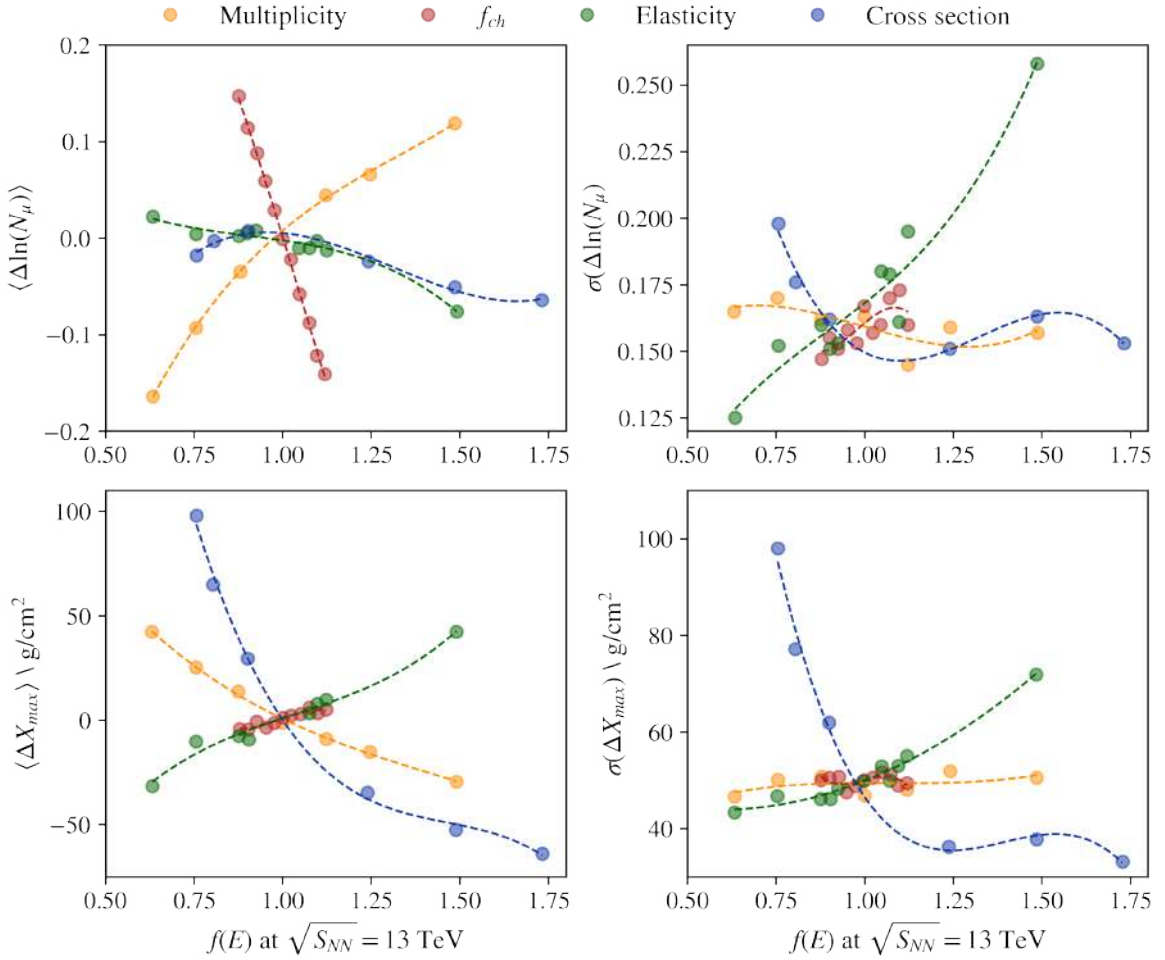


Figure 3.23: Impact of changing individual hadronic interaction features on the means and standard deviations of the logarithm of N_μ (top) and X_{max} (bottom) (reproduced from [110]). $10^{19.5}$ eV proton showers simulated with CONEX using SIBYLL-2.1 as the baseline model are used. In the left column relative shifts to the mean values are shown. The curves serve the purpose to help the eye. All values are shown as a function of $f(E, f_{19})$, where f_{19} is varied, evaluated at $E = \sqrt{S_{NN}}$ TeV. This factor is extrapolated logarithmically towards higher energies.

extrapolations with energy. Finally, at 10^{19} eV the value of f_{19} is reached, which governs the size of the modification.

The effect of these modifications on the mean and standard deviations of the muon number N_μ and the depth of shower maximum X_{max} are presented in figure 3.23 (reproduced from [110]). The values for proton showers are shown as a function of the modification factor evaluated at the LHC energy scale of nucleon-nucleon collisions at 13 TeV. It can be seen that the most efficient way to increase the number of muons is decreasing the fraction of neutral pions or increasing the total multiplicity. Reasonable agreement between data and simulations can be observed concerning the standard deviation of the muon number for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d. This constraints the changes that can be applied to the elasticity, which is the parameter with the biggest impact on $\sigma(N_\mu)$. Post-LHC models also give a consistent description of X_{max} . This means that parameter values that have a big repercussion on the mean or the standard deviation of X_{max} cannot deviate too much from those used in current models. The cross-section is the most influential parameter on X_{max} which indeed has already been measured very precisely.

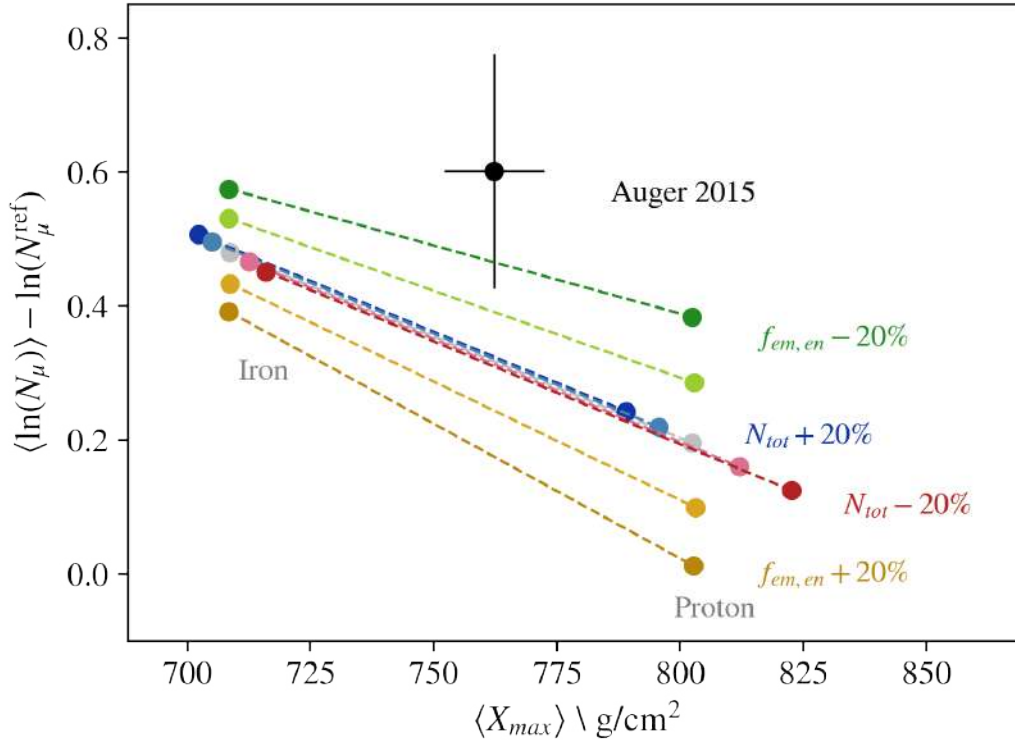


Figure 3.24: Impact of modifying the hadronic multiplicity N_{tot} (dashed lines) and the electromagnetic energy ratio $f_{em,en}$ (dotted lines) in collisions at 13 TeV on EPOS-LHC predictions of the air shower observables X_{max} and $\ln(N_\mu)$. This result is reproduced from [112]. The datum is from the Pierre Auger Observatory [113]. The model lines represent all values that can be obtained for any combination of nuclei from a pure proton scenario (bottom right) to pure iron scenario (top left).

The correlated impact of N_{tot} and the fraction of energy

$$f_{em,en} = \frac{E_{em}}{E_{tot}}$$

that goes into the electromagnetic channel on X_{max} and $\ln(N_\mu)$ in full air shower simulations is explored in [112]. The result for 10^{19} eV simulations done with EPOS-LHC in CONEX is presented in figure 3.24 (reproduced from [112]). Lines represent all possible resulting mean values of X_{max} and $\ln(N_\mu)$ for mass compositions of cosmic rays between pure proton (bottom right end of lines) and pure iron compositions (top left end of lines). *Ad-hoc* modifications on N_{tot} and $f_{em,en}$ by up to $\pm 20\%$ are also shown. These are compared to a measurement done by the Pierre Auger Observatory [100]. Changes in N_{tot} move the line almost horizontally and thus no answer regarding the Muon Puzzle can be found here. However, a change in $f_{em,en}$ has a perpendicular effect to the lines and suggests that a solution to the Muon Puzzle might be found in the modification of this energy ratio.

There are already glimpses of the solution to the puzzle. The ALICE experiment observed an enhancement of strangeness production in high-multiplicity events at mid-rapidity [114]. An increase in strangeness leads to a decrease in the pion yield which includes neutral pions. If this effect is also present in the forward region and in other collision systems important for air shower development, the Muon Puzzle might be solved. It was also shown in [115] that the forward production of ρ^0 mesons plays an important role in increasing $f_{ch,en}$. Forward ρ^0 production is namely an alternative to the charge exchange reaction $\pi^- + p \rightarrow \pi^0 + n + X$. In this process, the leading π^0 carries a considerable amount of energy away from the hadronic channel. It is furthermore known that an increased baryon production increases the number of interactions where no leading π^0 is produced [116]. As a consequence, more energy goes into hadronic subshowers, leading to more hadronic generations and in the end to more muons. Yet, there is lack of data on both identified hadron spectra and on strangeness production in the forward region.

3.6. Fluctuations of the muon content

The study presented in figure 3.24 focuses on the average muon content. Since the parameters N_{tot} and $f_{em,en}$ are modified there for the whole shower, this suggests that the average muon number depends on all processes that occur along the shower, including the low-energy interactions. In an effort to understand the phenomenology of the muon component one can also take a closer look at its shower-to-shower distributions [66, 117, 67, 118, 119]. There are many arguments in favor of a close relationship between the fluctuations of the muon content and characteristics of the first interaction

which we summarize here.

In order to understand the shower-to-shower fluctuations of N_μ , one can start taking into account that the multiplicity varies from interaction to interaction. This way one gets an expression slightly different from (3.5):

$$N_\mu = \prod_{i=1}^{n_p} N_{ch}^i$$

where N_{ch}^i is an average hadronic multiplicity of each generation. Note that the value of n_p is not calculated as for expression (3.5) anymore because allowing for a varying hadronic multiplicity changes the overall energy budget of each subshower leading to a different point where the critical energy is reached. Assuming that the hadronic multiplicities of all interactions arise from a common probability distribution with mean N_{ch} and dispersion $\sigma(N_{ch})$, the dispersions of N_{ch}^i can be calculated as

$$\sigma(N_{ch}^i) = \frac{\sigma(N_{ch})}{\sqrt{N_{i-1}}}$$

where N_{i-1} is the number of hadronically interacting particles in the previous generation. It becomes evident that the fluctuations of the average multiplicity decrease as the shower develops and the number N_{i-1} increases. Consequently, the contributions to the fluctuation of N_μ from later shower stages become smaller. This effect is amplified by the fact that $\sigma(N_{ch})$ actually decreases with energy [120].

Another fluctuation comes from the variability in how energy is shared among particles emerging in each generation. The number of muons can be calculated taking into account only those fluctuations, that arise from the first interaction, by summing the average number of muons of each of the N_{ch}^1 subshowers descending from the first interaction:

$$N_\mu^1(E_0) = \sum_{i=1}^{N_{ch}^1} \langle N_\mu(E_i^{had}) \rangle = \sum_{i=1}^{N_{ch}^1} \left(\frac{E_i^{had}}{\xi_c^\pi} \right)^\beta \quad (3.19)$$

where E_i^{had} is the energy of the i th hadronically interacting particle. The approximation done here is that a subshower generated by a hadron from the first interaction, most probably a charged pion, behaves the same way as a proton initiated shower. Defining the fraction $x_i = E_i^{had}/E_0$, equation (3.19) can be rewritten as

$$N_\mu^1(E_0) = \left(\frac{E_0}{\xi_c^\pi} \right)^\beta \cdot \sum_{i=1}^{N_{ch}^1} (x_i)^\beta = \langle N_\mu(E_0) \rangle \cdot \alpha_1.$$

In the last step expression (3.5) is used and the parameter

$$\alpha_1 = \sum_{i=1}^{N_{ch}^1} (x_i)^\beta$$

defined.

Taking into account the fluctuations down to the second interaction results in

$$N_\mu^2(E_0) = \sum_{i=1}^{N_{ch}^1} \sum_{j=1}^{(N_{ch}^2)_i} \left(\frac{E_{ij}^{had}}{\xi_c^\pi} \right)^\beta \approx N_\mu^1(E_0) \cdot \sum_{j=1}^{N_{ch}^2} (\tilde{x}_j)^\beta = N_\mu^1(E_0) \cdot \alpha_2$$

where $E_{ij}^{had} = E_0 x_i x_{ij}$ is the energy carried by the hadronically interacting particles from the second generation and an equivalent parameter to α_1 , but for the second generation, was defined. The interactions of the second generation are approximately treated as equal in order to be able to extract the common factor $N_\mu^1(E_0)$. The interesting point here is that there is no need to assume equipartition of energy. It is only assumed that the uneven way energy is distributed among particles is in that generation equal for all interactions. If one defines α_k as the sum of all energy fractions of hadronically interacting particles from generation k (out of the energy they have at their disposal) elevated to the value of β , repeats the reasoning and approximates the final muon number N_μ with the number of muons obtained from taking into consideration all the hadronically interacting particles from generation n_p , one is left with:

$$N_\mu = \langle N_\mu(E_0) \rangle \cdot \alpha_1 \cdot \dots \cdot \alpha_{n_p}. \quad (3.20)$$

Equation (3.20) shows that the number of muons is connected with a variable of the first interaction, α_1 . The correlation is indeed strong, as can be seen in figure 3.25 for a set of 5000 showers initiated by vertical protons of 10^{20} eV (left) and sets of 1000 showers initiated by vertical protons of lower energies (right). The accumulation of events along the vertical line corresponding to $\alpha_1 = 1$ (best visible for the highest energy simulations because there are more events and the α_1 - N_μ plane is zoomed in) corresponds to quasi-elastic events. The correlation coefficient ($\rho_{X,Y} = cov(X,Y)/(\sigma_X \cdot \sigma_Y)$) is close to 0.8 for all cases.

The relative fluctuations of N_μ can be estimated through the quadratic sum of the relative fluctuations of α_k from the different generations k :

$$\left(\frac{\sigma(N_\mu)}{N_\mu} \right)^2 = \sum_{k=1}^{n_p} \left(\frac{\sigma(\alpha_k)}{\alpha_k} \right)^2.$$

The more advanced the shower is (higher value of k) the more interacting particles

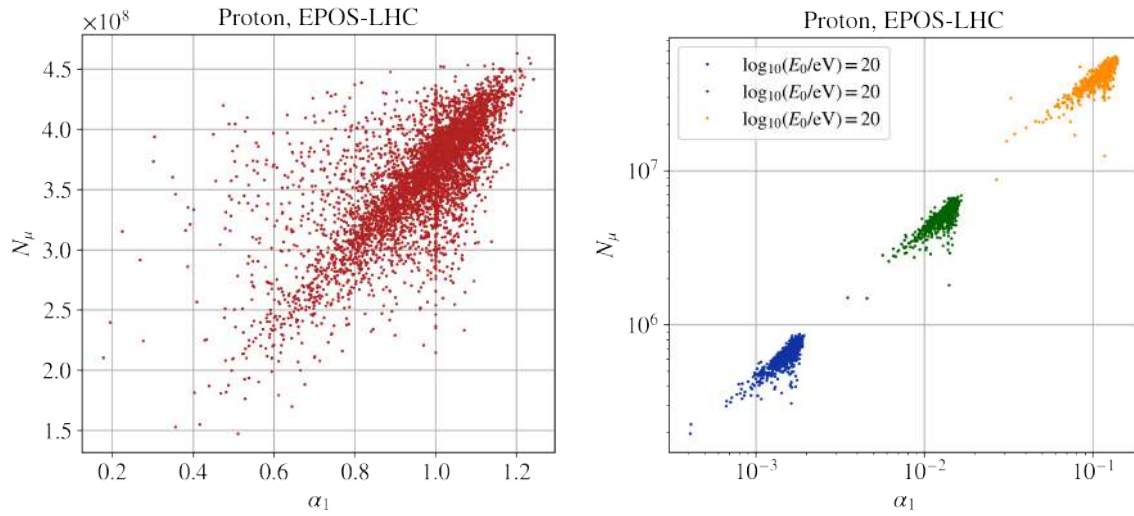


Figure 3.25: Correlation between the number of muons and α_1 for a set of 5000 showers initiated by vertical protons of 10^{20} eV (left) and sets of 1000 showers initiated by vertical protons of lower energies (right). All simulations were performed using the high-energy interaction model EPOS-LHC. The accumulation of events along the vertical line at $\alpha_1 = 1$ corresponds to quasi-elastic events. The correlation coefficient ($\rho_{X,Y} = cov(X,Y)/(\sigma_X \cdot \sigma_Y)$) is close to 0.8 for all cases.

there are. As a consequence $\sigma(\alpha_k)$ becomes smaller as k increases. This means that the highest contribution to the fluctuation in the number of muons comes from the fluctuations that occur in the first interaction.

In this chapter we have seen a wide variety of parameters that are useful to describe extensive air showers and simple expressions to relate them with different observables. We will combine and expand some of these expressions in the next chapter in order to describe the $X_{max}-N_\mu$ anticorrelation. The parameters presented so far are not enough to achieve this goal. When carrying out our calculations, new useful parameters will be defined.

Chapter 4

Understanding the $X_{max} - N_{\mu}$ anticorrelation

“Forget this world and all its troubles and if possible its multitudinous Charlatans - everything in short but the Enchantress of Numbers.”

— Ada Lovelace

In order to comprehend the X_{max} - N_{μ} anticorrelation it is reasonable to begin examining X_{max} and N_{μ} separately. In section 3.3 we explained how we obtain these two observables and other parameters of interest from simulations. By examining X_{max} and N_{μ} separately, the link between these two observables may be detected as soon as they depend on some parameter in common. In this pursuit, ideas based on the semi-empirical model and related topics presented in section 3.4 are used as a basis. The shift from constant values to shower-to-shower values is important to reproduce X_{max} - N_{μ} distributions instead of mean values. Since the achievable degree of simplicity of the future model was not foreseeable at this stage, it made sense to start incorporating as much detail as possible, prioritizing a thorough understanding. The possibility to break down the model and thus get more fundamental and simple expressions would come at a later step. We conclude with a simple model with a small set of parameters which describes the X_{max} - N_{μ} anticorrelation.

4.1. Initial approach

A first attempt to express N_{μ} as a function of physical features of the shower is shown in figure 4.1 (left). Here the number of muons is calculated as the sum of the number of muons of each subshower that originated from a particle of the first interaction that contributes to the hadronic channel, namely each particle that is not

a π^0 , η or an electromagnetic particle:

$$N_\mu = \sum_i \left(\frac{E_i^{had}}{\xi_c^\pi} \right)^\beta.$$

This is very similar to the work of Cazon et al. that we presented in section 3.6. We use expression (3.5) for the number of muons of each subshower and a constant value of $\xi_c^\pi = 20$ GeV which we discussed in section 3.4.2. E_i^{had} is the energy of the i th particle which contributes to the hadronic channel. For this part of the analysis we use simulated showers generated by vertical protons with primary energies of 10^{17} eV, 10^{18} eV, 10^{19} eV and 10^{20} eV. For each energy group, β can be chosen so that the errors in the predictions are minimized. The obtained values are 0.861, 0.868, 0.872 and 0.881, respectively, for increasing energy. They are a bit low considering the more realistic values that follow from equation (3.17), which are shown in figure 4.2, but they follow the expected increasing behaviour. One important comment to make here is that it is once again confirmed (like in the works from Cazon et al., [66, 117, 67, 118, 119]) that information about the first interaction is enough to get a good estimate of N_μ . It is also worth noting that the aim here is not to test the semi-empirical model as it is but to see if by reasonable modifications N_μ (and X_{max} in the next paragraph) can be estimated from a certain set of features of the shower.

A similar approach in order to get an estimate of X_{max} is to consider the electromagnetic subshowers generated by each particle of the first generation. Of course, in

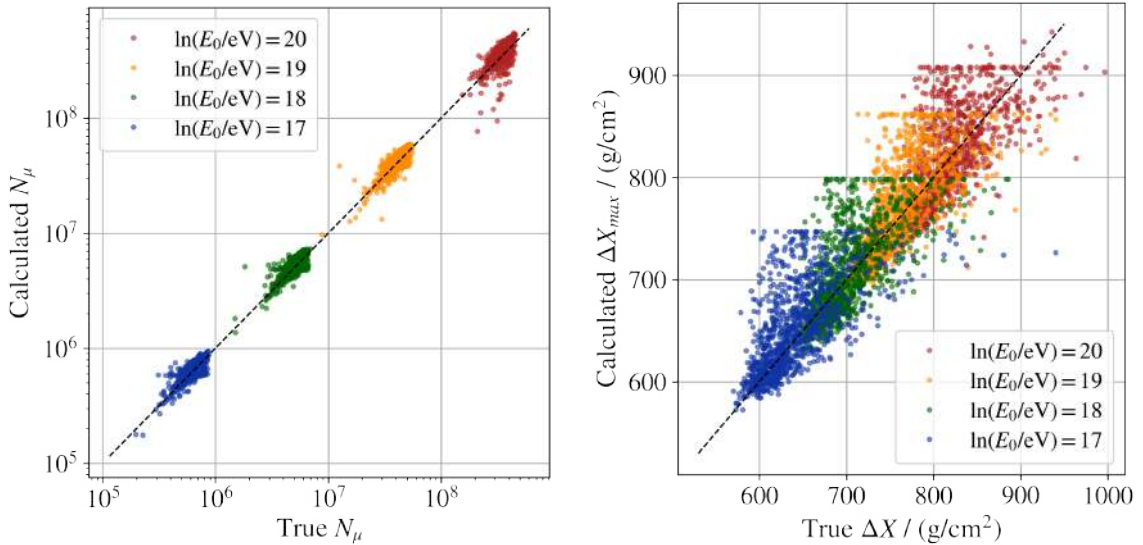


Figure 4.1: Comparison between the true number of muons at ground (left) and the true values of $\Delta X = X_{max} - X_0$ (right) with the corresponding calculated values. The *true* values are extracted from showers simulated with CONEX using EPOS LHC. Vertical protons of different energies are considered as primaries. The calculated values are computed using the corresponding parameters from these same simulations as explained in the text.

the case of a π^0 or η the assumption is that it decays immediately into 2 photons of half the secondary's energy. The value of X_{max} of these subshowers is estimated through expression (3.3) replacing E_0 by $E_j^{em}/2$. E_j^{em} is the energy of the j th π^0 or η . For all the other secondaries we use expression (3.3) for the calculation of X_{max}^γ , as well, with the exception that E_0 is replaced by $E_i^{had}/(2 \cdot N_{tot}^*)$. Here is N_{tot}^* an effective multiplicity that differs between the different energy groups. It can be thought of as the number of particles that carry enough energy to generate subshowers that have some observable impact on the final complete shower. This parameter can be chosen so as to minimize the errors of the predictions. For increasing energy the values are 1, 2.5, 4.5 and 13. There is no particular physical justification for these precise values, but again the fact that these values grow with energy is in agreement with the fact that total multiplicities grow with energy. Finally, we calculate the average of all the obtained individual X_{max} values corresponding to the electromagnetic subshowers, weighted by the energies of the corresponding initiating secondaries. Since the individual subshowers are considered to start at the point where the first interaction occurred, the calculated estimates are compared to $\Delta X = X_{max} - X_0$ in figure 4.1 (right).

In order to obtain deep showers an energetic leading particle that in many successive interactions keeps a considerable amount of energy is certainly necessary. The deeper the shower the less combinations there are to achieve that value of X_{max} . Conversely, the smaller the depth, the more combinations of repartition of energy among particles there are in order to reach that depth. Combining this reasoning with the fact that the leading interaction is dominant in the prediction of X_{max} in this approach, it makes sense that the errors in the predictions are higher for smaller depths (within each energy group). The accumulation of events on horizontal lines (in figure 4.1, right) corresponds to particularly bad predictions that arose from quasi-elastic events where the information useful for this simple model actually would be found in the second interaction.

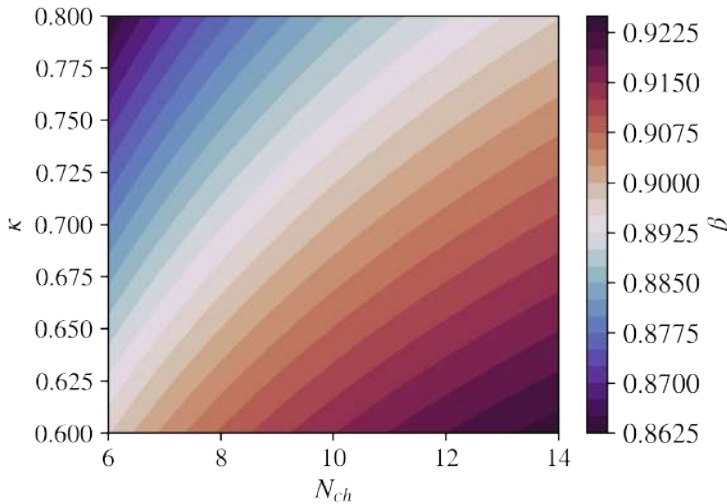


Figure 4.2: β as a function of the multiplicity N_{ch} of hadronically interacting particles and the inelasticity κ . Expression (3.17) is used which involves the leading particle effect.

These predictions are still quite rudimentary. As we shall see, there is much room for improvement. But, most importantly, the drawback in these approaches is that we use very detailed information about the first interaction, namely each individual energy and the identity of each particle in order to assign it the correct role in the model. It is crucial to be able to shift to more simple parameters that represent this complexity in a shower-to-shower manner. What can already be perceived is that the energy that goes into the hadronic channel and the multiplicity N_{tot}^* might have a major role.

4.2. Improvement of the semi-empirical model

We go back now to the expressions for X_{max} and N_μ presented in chapter 3 but instead of using mean effective parameters we insert ranges of realistic values. Thus, we obtain ranges of ΔX and N_μ that can be compared to ranges covered by distributions from simulations. This gives hints as to where improvement is needed in order to be able to describe the $\Delta X - N_\mu$ anticorrelation.

The most detailed but simple expressions for ΔX and N_μ so far include the leading particle effect in the calculation of β and use the critical energy ξ_c^π as calculated by Kampert and Unger, which only depends on the total multiplicity N_{tot} . We calculate $N_\mu = (E_0/\xi_c^\pi)^\beta$ with the corresponding expressions from chapter 3 for proton showers of different energies. For the parameters needed we use values in intervals of the effective multiplicity N_{tot} and the effective inelasticity κ motivated by the corresponding distributions shown in figures 3.7 and 4.3 (left). This means that we use realistic values of $\langle N_0 \rangle_{geom} + \langle N_{ch} \rangle_{geom}$ to replace the effective total multiplicity N_{tot} . When looking into the distribution of κ from all interactions occurring in a particular shower, a skewed distribution with a long tail can be observed. A more representative value for the

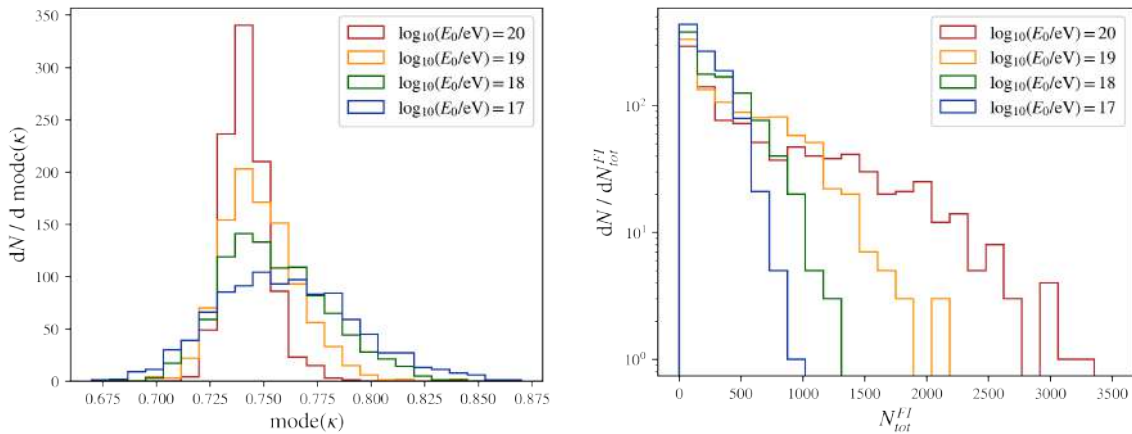


Figure 4.3: Distributions of the mode of the inelasticity κ (left) and the total multiplicity N_{tot}^{FI} of the first interaction (right) for sets of 1000 simulations done with CONEX using EPOS-LHC with vertical protons of different energies as primaries.

shower is then the mode of these values (the most probable value), rather than the ordinary mean, since it is not affected by how long the tail is. Furthermore, since the product of the inelasticities has no interpretable meaning or use in the model it does not make sense to consider geometric means here. Thus, we use values of $\text{mode}(\kappa)$ to replace the effective inelasticity. The distribution of these values is shown in figure 4.3 (left) for showers of different primary energy.

We calculate $\Delta X = X_{max} - X_0$ in two ways. First, Matthews' approach is used. No concrete expression was given by him for ΔX including the leading particle effect. So we follow the idea he uses for hadronic showers without considering this effect and calculate X_{max} for the bulk of subshowers that come from the first interaction (and are treated as if they were all equal) with the following expression:

$$\Delta X = \lambda_r \ln \left(\frac{(1 - \tilde{f}_{ch,en}^{FI}) \kappa^{FI} E_0}{2 (N_{tot}^{FI}/3) \xi_c^e} \right) \quad (4.1)$$

This expression, which requires three parameters that change from shower to shower, is quite simplified because the inelasticity is not taken into account in subsequent generations. κ^{FI} and N_{tot}^{FI} are the inelasticity and the total multiplicity of the first interaction, respectively. $\tilde{f}_{ch,en}$ is the fraction of the hadronic energy out of the energy that is available for production of new particles in contrast to $f_{ch,en}$ which was the same fraction but out of the total energy of the interaction. Ranges of realistic values for N_{tot}^{FI} , κ^{FI} , $\tilde{f}_{ch,en}$ are taken from figures 4.3 (right) and 4.4 (left and right). For N_{tot}^{FI} we count explicitly how many particles are created in the first interaction of each shower, disregarding electrons, photons and muons. They most probably will not influence the

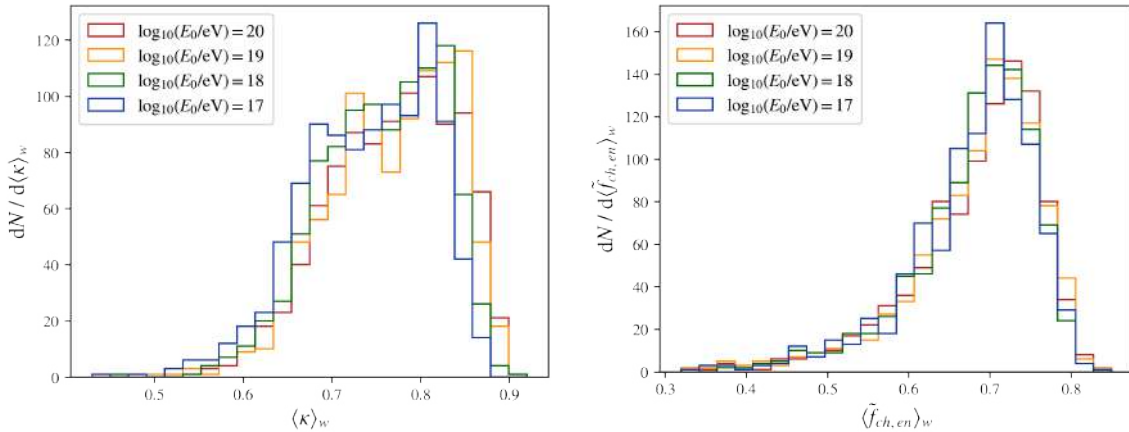


Figure 4.4: Distributions of the inelasticity $\langle \kappa \rangle_w$ (left) and the hadronic fraction of energy $\langle \tilde{f}_{ch,en} \rangle_w$ (right). These weighted averages are computed for each shower taking into account all the interactions that occur (down to the threshold E_{thr}). The respective weight is the energy available for production of new particles in that interaction, i.e. the inelasticity times the energy of the interaction. This means that the first interaction will dominate this weighted average. Distributions shown here correspond to the same sets as shown in figure 4.3.

development of the shower, considering that they typically take a negligible amount of energy. We could indeed also use the strict values from the first interaction for the inelasticity and the hadronic energy fraction. Instead we already present in this context the weighted averages of the corresponding values over all interactions present in that shower and denote it by $\langle \cdot \rangle_w$. The weight w for each interaction is the energy that is available for production of new particles, i.e. inelasticity times the energy of the interaction. These weighted averages will be used at a later point of the work and the results for this stage don't change substantially. Consequently and in order to make later sections easier to read, we introduce them now. In brief, we use realistic values of N_{tot}^{FI} , $\langle \kappa \rangle_w$ and $\langle \tilde{f}_{ch,en} \rangle_w$ to be inserted in N_{tot}^{FI} , $\tilde{f}_{ch,en}^{FI}$ and κ^{FI} , respectively. Note that, since the highest weights w happen in the first interaction(s), these weighted averages will indeed be very correlated to the corresponding values of the first interaction.

Secondly, Kampert and Unger's approach is tested. Even though their calculations are intended to follow the hadronic development of the shower and it is known that X_{max} stems from the electromagnetic component, one could argue that this component is fed by neutral pions which arise from the hadronic component. This motivates probing

$$\Delta X = n_d(N_{tot}, (1 - \kappa^{FI})E_0) \cdot 120 \text{ g/cm}^2,$$

where N_{tot} is the effective multiplicity of the whole shower and κ^{FI} the inelasticity of the first interaction. This expression means that we take the energy "left behind" by the leading particle of the first interaction, which is why κ^{FI} is used in the expression, and calculate then the ΔX value from the corresponding subshower. The effective multiplicity N_{tot} is used because, once we have kept the inelastic energy from the first interaction, the development is the same as in Kampert and Unger's approach. Being consistent with the calculations so far, we replace κ^{FI} and N_{tot} by reasonable values of the weighted average $\langle \kappa \rangle_w$ and $\langle N_0 \rangle_{geom} + \langle N_{ch} \rangle_{geom}$, respectively. The interaction length $\lambda_I = 120 \text{ g/cm}^2$ is used as in Matthews' work. We only vary the other parameters because the effect of varying a multiplicative factor is easy to interpret.

The concrete ranges we used to test the expressions from Matthews and Kampert

Parameter	Replacement	10^{17} eV	10^{18} eV	10^{19} eV	10^{20} eV
N_{tot}^{FI}	N_{tot}^{FI}	1 – 1000	1 – 1250	1 – 2000	1 – 3250
N_{tot}	$\langle N_0 \rangle_{geom} + \langle N_{ch} \rangle_{geom}$	15.0 – 20.0	15.25 – 20.0	16.5 – 20.0	16.5 – 20.0
κ^{FI}	mode(κ)	0.59 – 0.7	0.6 – 0.68	0.62 – 0.67	0.62 – 0.66
κ	$\langle \kappa \rangle_w$	0.62 – 0.88	0.62 – 0.88	0.62 – 0.88	0.62 – 0.88
$\tilde{f}_{ch,en}^{FI}$	$\langle \tilde{f}_{ch,en} \rangle_w$	0.6 – 0.8	0.6 – 0.8	0.6 – 0.8	0.6 – 0.8

Table 4.1: Current models and our model are tested on ranges of the parameters N_{tot}^{FI} , N_{tot} , κ^{FI} , κ and $\tilde{f}_{ch,en}^{FI}$ instead of using constant mean values. The used ranges are presented here for the different energies and are motivated by previously presented plots. We also summarize in the second column how we interpret each parameter.

& Unger are summarized in 4.1, while the results of our calculations are summarized in table 4.2. The “parameter” column in table 4.1 refers to the parameters in the expressions for ΔX while the “replacement” column indicates how we interpret these parameters and how we calculate them for each simulated shower. The resulting calculated values of ΔX tend to be low for both approaches. Furthermore, only small ranges of ΔX can be obtained with Kampert and Unger’s approach compared to the simulations while Matthews’ approach covers the possible values quite well. With regard to the muon number, values tend to be too high. We already got a hint in the previous section that the first interaction plays a significant role and this is not taken into account in these expressions yet. We propose now an improvement to this model by adding details concerning the first interaction.

More concretely, we will improve these expressions by incorporating more parameters related to the first interaction in Kampert and Unger’s approach. We still use that

$$\frac{\rho(h)h_s}{n \cos(\theta)} = \lambda_I = \lambda_{dec} = \rho(h)c\tau_{\pi^\pm}\gamma \quad (4.2)$$

with the difference being in how

$$\gamma = \frac{E_\pi^{ind}}{m_{\pi^\pm}c^2}$$

is calculated. E_π^{ind} is the energy of individual pions in the n -th generation. n_g will be called the number of generations necessary for the individual pions to achieve their critical energy taking into account the effective hadronic multiplicity N_{ch} , the effective fraction of energy that stays in the hadronic channel $\tilde{f}_{ch,en}$, the effective inelasticity κ and the corresponding values of the first interaction N_{ch}^{FI} , $\tilde{f}_{ch,en}^{FI}$ and κ^{FI} instead of only the total multiplicity N_{tot} :

$$\xi_c^\pi = \frac{\left(1 - (1 - \tilde{f}_{ch,en}^{FI})\kappa^{FI}\right) \left(1 - (1 - \tilde{f}_{ch,en})\kappa\right)^{n_g-1} E_0}{(1 + N_{ch}^{FI})(1 + N_{ch})^{n_g-1}}. \quad (4.3)$$

Observable	10^{17}eV	10^{18}eV	10^{19}eV	10^{20}eV
ΔX - EPOS-LHC	571 – 940	633 – 885	698 – 940	752 – 996
ΔX - M.	455 – 749	532 – 834	600 – 920	667 – 1005
ΔX - K., U.	510 – 624	610 – 729	708 – 812	805 – 915
N_μ - EPOS-LHC	$2.0e5 - 8.7e5$	$1.5e6 - 6.9e6$	$8.8e6 - 5.6e7$	$1.6e8 - 4.5e8$
N_μ - M., K., U.	$5.5e5 - 8.7e5$	$5.1e6 - 7.6e6$	$4.8e7 - 6.8e7$	$4.4e8 - 5.8e8$

Table 4.2: The ranges of parameters presented in table 4.1 are inserted into the expressions for ΔX and N_μ . The resulting ranges of calculated ΔX and N_μ are compared to the corresponding ranges obtained from simulations done with CONEX using EPOS-LHC with vertical protons of different energies as primaries. N_μ is calculated combining the approaches from Matthews, Kampert and Unger. ΔX is calculated in two ways, using Matthews’ approach (M.) and using Kampert and Ungers’ approach (K., U.).

$\left(1 - (1 - \tilde{f}_{ch,en}^{FI})\kappa^{FI}\right) \left(1 - (1 - \tilde{f}_{ch,en})\kappa\right)^{n_g-1}$ is the fraction of the primary energy that stays in the hadronic channel after n_g interactions including the leading particle. $(1 - \tilde{f}_{ch,en})\kappa$ is namely the energy that is “lost” to the electromagnetic component. We describe this way the shower that is composed of all the subshowers initiated by the hadronically interacting particles from the first interaction. This energy is divided among all the charged pions present at the n_g -th interaction which decay to

$$N_\mu = (1 + N_{ch}^{FI}) (1 + N_{ch})^{n_g-1} \quad (4.4)$$

muons that reach ground. The first interaction has been separated from the consecutive ones, because the multiplicity N_{ch}^{FI} of the first interaction is not only remarkably higher than N_{ch} but also covers a wide range of possible values which should have a noticeable impact on the predictions. Separating $\left(1 - (1 - \tilde{f}_{ch,en}^{FI})\kappa^{FI}\right)$ from the rest of the fraction is worthwhile, as well, because even though there can be observed only a small difference, it turns out to be very important when using neural networks in the next chapter. In principle, the factors including the inelasticity and the hadronic fraction could be summarized in one parameter each (one corresponding to the first interaction and another one corresponding to the rest of the interactions) but separating them into two interpretable parameters gives the opportunity to later test the model in more detail.

Inserting our expression for the critical energy of charged pions (4.3) into equation (4.2) we obtain the following:

$$\begin{aligned} \frac{\rho(h)h_s}{n_g \cos(\theta)} &= \frac{\rho(h)c\tau_{\pi^\pm}\xi_c^\pi}{m_{\pi^\pm}c^2} \\ \Rightarrow \frac{h_s}{n_g \cos(\theta)} &= \frac{c\tau_{\pi^\pm}}{m_{\pi^\pm}c^2} \frac{\left(1 - (1 - \tilde{f}_{ch,en}^{FI})\kappa^{FI}\right) \left(1 - (1 - \tilde{f}_{ch,en})\kappa\right)^{n_g-1} E_0}{(1 + N_{ch}^{FI})(1 + N_{ch})^{n_g-1}} \\ \Rightarrow \frac{h_s}{\cos(\theta)} \frac{m_{\pi^\pm}c^2}{c\tau_{\pi^\pm}E_0} \frac{1 + N_{ch}^{FI}}{1 + N_{ch}} \frac{1 - (1 - \tilde{f}_{ch,en})\kappa}{1 - (1 - \tilde{f}_{ch,en}^{FI})\kappa^{FI}} &= n_g \left(\frac{1 - (1 - \tilde{f}_{ch,en})\kappa}{1 + N_{ch}} \right)^{n_g} \\ &= n_g \exp\left(n_g \ln\left(\frac{1 - (1 - \tilde{f}_{ch,en})\kappa}{1 + N_{ch}}\right)\right) \\ \Rightarrow n_g &= W_{-1}\left(\ln\left(\frac{1 - (1 - \tilde{f}_{ch,en})\kappa}{1 + N_{ch}}\right) \cdot \frac{h_s}{\cos(\theta)} \frac{m_{\pi^\pm}c^2}{c\tau_{\pi^\pm}E_0} \frac{1 + N_{ch}^{FI}}{1 + N_{ch}} \frac{1 - (1 - \tilde{f}_{ch,en})\kappa}{1 - (1 - \tilde{f}_{ch,en}^{FI})\kappa^{FI}}\right) \\ &\quad \Big/ \ln\left(\frac{1 - (1 - \tilde{f}_{ch,en})\kappa}{1 + N_{ch}}\right). \end{aligned} \quad (4.5)$$

The first and second rows in figure 4.6 show the result from calculating ΔX using equation (4.1) and N_μ using equations (4.4) and (4.5) for a primary energy of $E_0 =$

10^{20} eV. The values corresponding to this primary energy are summarized in table 4.3 and used to replace the parameters. The most noticeable variability happens when varying the parameters that describe the first interaction. Therefore, for simplification we use fixed mean values for the three effective parameters that are representative for the whole shower (see table 4.3). As a further simplification we show the results for the extreme values of κ , while allowing for a variety of values of N_{ch}^{FI} and $\tilde{f}_{ch,en}^{FI}$. In order to understand the effect of the parameters of the first interaction N_{ch}^{FI} , $\tilde{f}_{ch,en}^{FI}$ and κ on ΔX and N_μ we show in the first row of figure 4.6 the result of varying N_{ch}^{FI} and $\tilde{f}_{ch,en}^{FI}$ for the extreme upper value $\kappa = 0.88$, while the result for the extreme lower value $\kappa = 0.62$ is shown in the second row. For this we need the distribution of N_{ch}^{FI} , which is shown in figure 4.5 (left).

$\langle N_{ch} \rangle_{geom}$ and $\langle \kappa \rangle_w$ have already been inserted in the models from Matthews, Kampert and Unger and we use here their mean values. The only new effective parameter is $\tilde{f}_{ch,en}$ for which we use the geometric mean $\langle \tilde{f}_{ch,en} \rangle_{geom}$ whose distribution is shown in figure 4.5 (right). The product of the $\tilde{f}_{ch,en}$ values from the consecutive interactions in a shower is the fraction of energy that stays in the hadronic core, leaving the leading particle of each interaction out. Therefore, the product has the same meaning as individual values of $\tilde{f}_{ch,en}$, but for the whole shower. That is why the geometric mean is used again.

We see that higher values of the inelasticity of the first interaction stretch the distribution downwards to lower values of N_μ . This comes from the fact that the higher the inelasticity is, the more energy is available for the shower development. Consequently, there are more combinations of how to distribute the energy. There are combinations where more energy is “lost” to the electromagnetic component which implies a lower number of muons. One might think that this effect is artificial and that if we took into

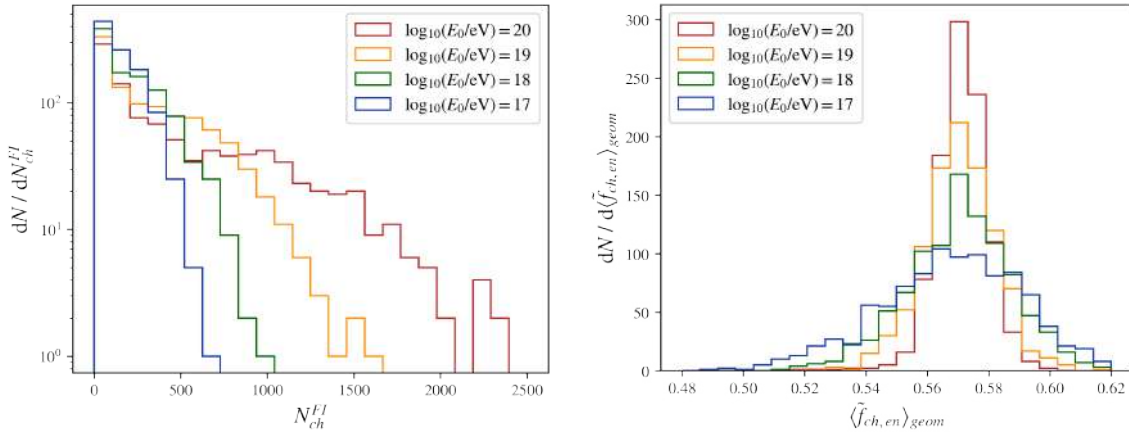


Figure 4.5: Distributions of the charged multiplicity N_{ch}^{FI} of the first interaction (left) and of the geometric mean $\langle \tilde{f}_{ch,en} \rangle_{geom}$ of the hadronic energy (right) for the sets of simulations from figure 4.3.

account what happens to the leading particles from all interactions another effect might be observed. We will see in a moment that the inelasticity is indeed well implemented in this model.

Since, when calculating ΔX , we don't take into account electromagnetic showers from later stages of the shower, this estimate tends to be low, as we already mentioned in the previous section. This is corrected with a multiplicative factor of 1.11. The number of muons is slightly overestimated which we correct dividing by a factor of 1.15. Doing this, a nice agreement between the simulations and our model is obtained. Furthermore, when looking at the distribution of N_{ch}^{FI} (left column) and $\langle \tilde{f}_{ch,en} \rangle_w$ (right column) in the $\Delta X - N_\mu$ plane, we can observe that the true distribution from simulations is fairly well reproduced (see figure 4.6, bottom left, and figure 4.7, top left, for comparison). The higher the multiplicity N_{ch}^{FI} (left column), the lower the energy of individual particles. This way, the shower develops in less generations and less energy is taken by the electromagnetic component, which lowers the value of ΔX . Thus, also more energy is kept in the hadronic channel, resulting in a higher number of muons. This is also and more efficiently achieved by a higher fraction $\langle \tilde{f}_{ch,en} \rangle_w$ (see right column).

Parameter	Replacement	Values	New name	Short description
N_{ch}^{FI}	N_{ch}^{FI}	1 – 2300	$M_0 = \ln(N_{ch}^{FI} + 1)$	Hadronic multiplicity of the first interaction
N_{ch}	$\langle N_{ch} \rangle_{geom}$	12.03	M_{eff}	Effective multiplicity of the rest of the shower
κ^{FI}	$\langle \kappa \rangle_w$	0.62, 0.88	K_0	Inelasticity mainly representing the first interaction
κ	mode(κ)	0.75	K_{eff}	Effective inelasticity of the rest of the shower
$\tilde{f}_{ch,en}^{FI}$	$\langle \tilde{f}_{ch,en} \rangle_w$	0.3 – 0.8	F_0	Hadronic energy fraction mainly representative of the first interaction
$\tilde{f}_{ch,en}$	$\langle \tilde{f}_{ch,en} \rangle_{geom}$	0.575	F_{eff}	Effective hadronic energy fraction of the rest of the shower

Table 4.3: In the following we test our model which can be summarized in expressions (4.1), (4.4) and (4.5) which depend on the six parameters N_{ch}^{FI} , N_{ch} , κ^{FI} , κ , $\tilde{f}_{ch,en}^{FI}$ and $\tilde{f}_{ch,en}$. In this table we summarize the interpretations we assign them (second column), the values we obtain from simulations of 10^{20} eV proton initiated showers (third column) and the new simpler name we assign them for the next chapter on neural networks (fourth column). For N_{ch}^{FI} and $\tilde{f}_{ch,en}^{FI}$ we are interested in ranges of values. For κ^{FI} we will only consider the upper and lower extreme values. For the other three effective parameters of the rest of the shower we consider mean values because their distributions are quite narrow. Furthermore, we redefine M_0 as $\ln(N_{ch}^{FI} + 1)$, which is the feature we will end up using when training neural networks. Finally, we give a short description of each parameter (fifth column).

The final step is to evaluate the model on a shower-to-shower basis. This is done repeating the procedure applied to the grid of realistic values as presented in figure 4.6 but this time on individual showers from simulations. For each shower the corresponding values of the three parameters from the first interaction are inserted in equations (4.1),

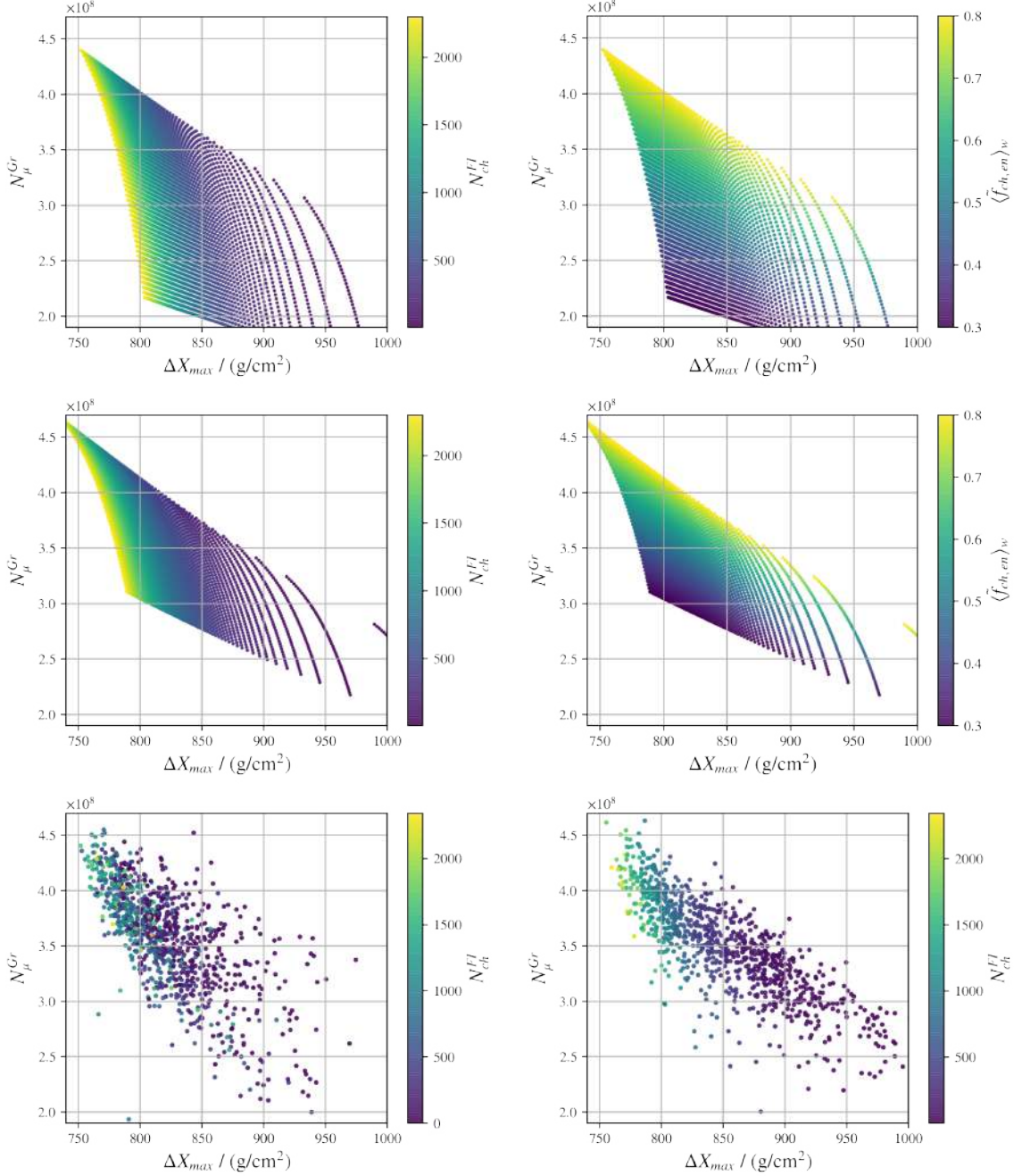


Figure 4.6: The first and second rows show the ΔX - N_μ distributions as calculated through our model given by expressions (4.1), (4.4) and (4.5) using the values summarized in table 4.3 for the parameters. The first row is calculated with a fixed value of $\langle \kappa \rangle_w = 0.88$, the second row with $\langle \kappa \rangle_w = 0.62$. In the last row we compare the behaviour of N_{ch}^{FI} in the ΔX - N_μ plane for the simulations obtained using EPOS-LHC for vertical protons of 10^{20} eV (left) with the distribution obtained through our model (right). Both distributions cover roughly the same area, have a similar anticorrelation and the parameter N_{ch}^{FI} presents the same behavior on both distributions.

(4.4) and (4.5). The same corrections are applied to the predictions of ΔX and N_μ as before. The only difference to the previous analysis is that we also insert the individual values of $\langle \tilde{f}_{ch,en} \rangle_{geom}$, while for the other two effective parameters mean values are used. Out of the three effective parameters is $\langle \tilde{f}_{ch,en} \rangle_{geom}$ the one with the biggest impact on the predicted values, so we want to see how the model works taking its individual values into consideration, as well.

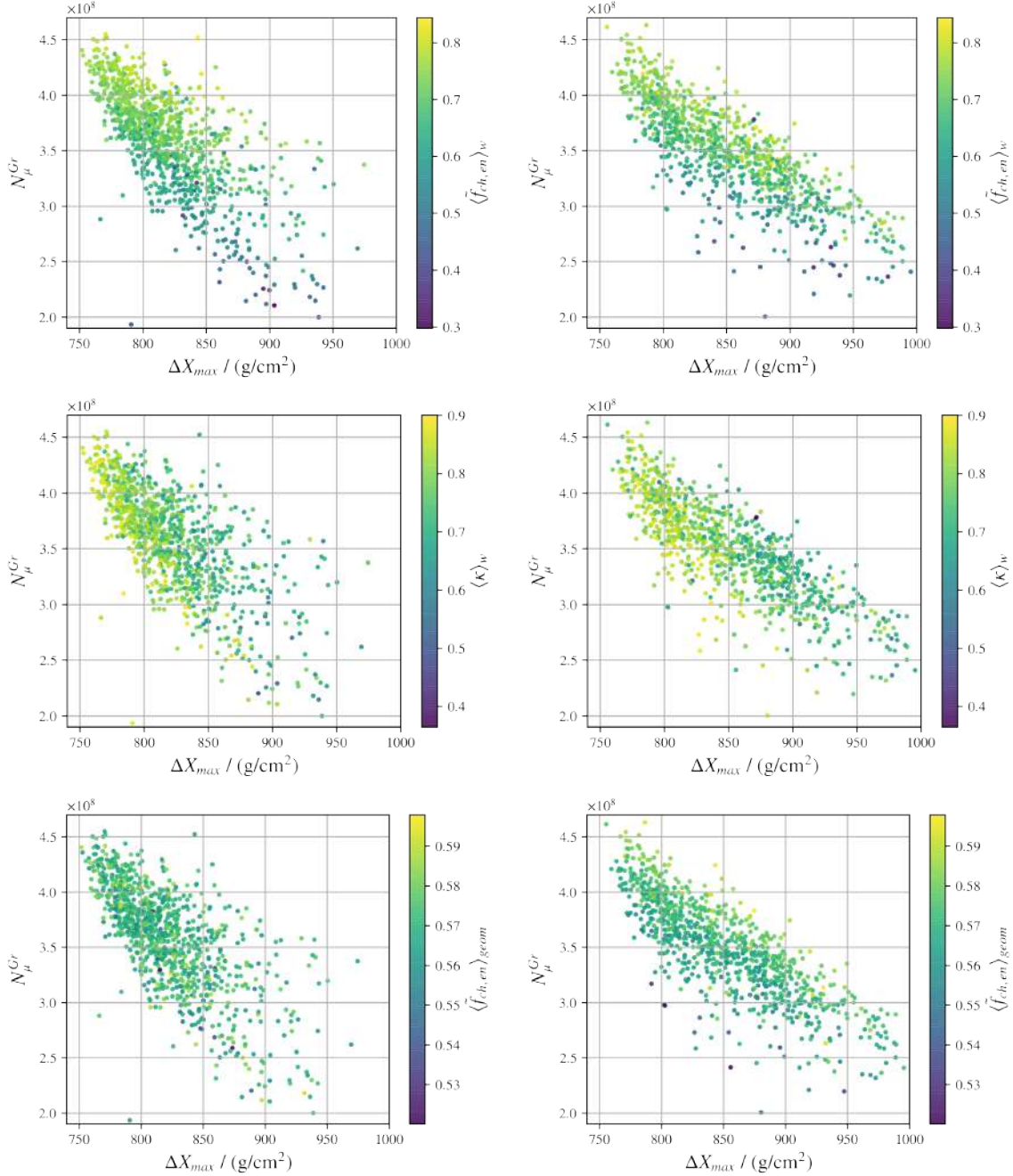


Figure 4.7: As in the last row of figure 4.6 we compare here the behaviour of the parameters $\langle \tilde{f}_{ch,en} \rangle_w$, $\langle \kappa \rangle_w$ and $\langle \tilde{f}_{ch,en} \rangle_{geom}$ on the distribution from simulations (left) with that on the distributions obtained through our model (right). The first two parameters show a remarkable coincidence.

The results are shown in figure 4.7 and the third row of figure 4.6. In each row we compare the behaviour of a different parameter in the $\Delta X-N_\mu$ plane for the simulations obtained using EPOS-LHC for vertical protons of 10^{20} eV (left) with the distribution obtained through our model (right). The $\Delta X-N_\mu$ distribution from our model covers the same region as simulations and the anticorrelation is represented fairly well. The fact that the distributions of the three parameters of the first interaction in the $\Delta X-N_\mu$ plane represent quite well those from simulations means that the behaviour of these parameters is qualitatively and even numerically well captured in the model. $\langle \tilde{f}_{ch,en} \rangle_{geom}$ follows the correct behaviour for lower values only.

This model is adequate to describe the $\Delta X-N_\mu$ anticorrelation as a function of a few physical parameters. The results in figures 4.6 and 4.7 show that the hadronic multiplicity and the inelasticity of the first interaction are correlated with ΔX and describe the anticorrelation moving horizontally. The hadronic energy of the first interaction, on the other hand, is more related to the number of muons and describes the anticorrelation moving almost vertically. None of these parameters alone describes the anticorrelation by itself. Instead, a combination of these parameters is needed to describe the inclined axis that the anticorrelation follows. In chapter 5 we will extend this model using neural networks and improve it even more.

Chapter 5

Construction of the model

“The laws of history are as absolute as the laws of physics, and if the probabilities of error are greater, it is only because history does not deal with as many humans as physics does atoms, so that individual variations count for more.”

— Isaac Asimov, *Foundation and Empire*

In this chapter we improve the model from section 4.2, which already gave good results, using neural networks. For the code we implement the functions provided by the SciPy ecosystem [121] (Numpy and Matplotlib), Scikit-learn [122] and Keras [123, 124] using Tensorflow [125], all interpreted by Python 3.6 [126]. The objective is now to reduce the number of features as much as possible while still being able to develop a neural network that captures all high-energy interaction scenarios simultaneously in the form of a unified or *universal model*. We show that this is possible when keeping M_0 , K_{eff} and F_0 as unknown features. After describing their distributions through suitable expressions we study the impact of changing the parameters that describe these distributions on the final observables $\langle X_{max} \rangle$, $\langle N_\mu \rangle$, $\sigma(X_{max})$ and $\sigma(N_\mu)$. We conclude performing a χ^2 fit using these observables to predict the unknown parameters as a means of testing our model.

5.1. Neural network architecture

A neural network is an algorithm inspired by biological neural networks that endeavors to recognize underlying patterns in a set of data. Neural networks shine when these are too complex to be understood using traditional approaches. Numerous techniques were developed in the last decades and several quite recent breakthroughs in computing power made these techniques computationally affordable. Many of these and other machine learning techniques are nowadays extensively used in the physical

sciences [127, 128]. In section 4.2 we developed a mathematical model for the X_{max} - N_μ anticorrelation based on the Heitler-splitting. It is a discrete model in the sense that processes involved are described discretely. It is to be expected that a neural network might identify other details not captured by the successive splitting, in which energy is equally divided among particles belonging to the same group. In order to build neural networks judiciously the previous work is crucial. We have already understood which parameters and which processes are important.

In *regression* tasks, like the one we are engaged in solving, one is interested in predicting *target* numerical values as a function of a set of input *features*. In our work the targets are X_{max} and N_μ . In our calculations, analyzing $\Delta X = X_{max} - X_0$ instead of X_{max} came quite naturally and there was no obstacle when comparing it to the corresponding values in simulations. When using CONEX the exact value of the depth of first interaction is accessible. In air shower experiments, however, we have no access to the value of X_0 on a shower-to-shower basis. Since we want to apply our final model to a dataset from the Pierre Auger Observatory, we need to cling to the value of X_{max} alone as one of the targets. N_μ , on the other hand, is the number of muons at ground (vertical atmospheric depth of 880 g/cm²), which corresponds to the mean position of the Pierre Auger Observatory. CONEX outputs several different longitudinal profiles. Among them is the profile of the total number of muons with energy above some predefined threshold traversing each depth. The energy threshold in our simulations is of 1 GeV. Since we will implement *supervised learning*, the training set that is fed to the algorithm will always include the desired and known solutions, called *labels*.

We start rightaway taking into account the features that are suggested by our model presented in section 4.2 and which are summarized in table 4.3: M_0 , K_0 and F_0 and the corresponding effective values M_{eff} , K_{eff} and F_{eff} representative for the whole shower. Furthermore, we add the depth of first interaction X_0 to the set of features, because in our mathematical model it was implicitly involved in the value of ΔX and we removed it from this expression of the target. Note that from here on we use $\ln(N_{tot}^{FI} + 1)$ instead of N_{tot}^{FI} as the feature we call M_0 because we need to avoid distributions with very long tails. When using neural networks, distributions concentrated in smaller regions are preferred. This is most of the times easily achieved by using the logarithm of the feature. When having only positive values, this compresses the values while keeping their order. The X_0 distribution has a long tail, as well, but in this case the original distribution tends to give better results. This can be explained by the fact that this distribution has a much shorter tail than the N_{tot}^{FI} distribution which makes the situation less problematic. In addition, the correlation is much higher between X_{max} and X_0 than between X_{max} and $\ln(X_0)$. And when looking into the relative importances of the features (see section 5.2), when switching from X_0 to $\ln(X_0)$ indeed importance is “lost” to the other features (we will discuss the concept of importance in section

5.2). We begin using these seven features but will reduce the number of features and discuss how and why in the next section.

The best way to analyze how well a model generalizes to new cases is splitting the data into two sets prior to training: the *training set* and the *test set*. The test set is put aside until the model has taken its final form and is ready to be tested on it. In order to be able to select the best and final model among several candidate models, one needs a way to evaluate how well these perform. A common solution to this necessity, that we use as well, is to subdivide the training set again. One fraction of it will be the proper reduced training set and the rest is the so-called *validation set*. Any candidate model can be tested on the validation set. Furthermore, since the complete process of training is repeated under many different conditions and in order to ensure that the test set is never used for training, we set a random number generator's seed for the separation of the different sets. We use the commonly employed values of 20% for the test set but only 10% out of the remaining set for the validation set since otherwise our training set gets too small.

It was shown by Banko et al. that very different machine learning techniques, including quite simple ones, performed almost identically well on a particular complex problem once they were given enough instances of data [129]. If the set is too small one might end up with unrepresentative data as a result of chance. In this situation a neural network might detect patterns in this noise which don't generalize. In order to examine the size of the dataset we need for accurate predictions, we started training neural networks on the training set extracted from a complete dataset of 1000 simulations, tested the resulting model on the validation set and analyzed if the model performed better for bigger datasets of up to 5000 simulations. Since the predictions of the neural networks did not change substantially, we continued working with complete datasets of 1000 instances from then on.

When separating the dataset into training, validation and test set, another care should be taken. It is always favourable to use a training set that is representative of all the cases one wants to generalize to, even those that are less probable. Our target distributions of X_{max} and N_μ have less values at the extreme upper and lower ends. It is possible that, by chance, when creating the training set not enough instances representative of some of these extremes are chosen. In order to guarantee that the less probable regions at the extremes of the X_{max} and N_μ distributions can be understood by the neural network, we use *stratified sampling* with respect to that feature which helps the most cover these critical regions. This consists in separating the training set in such a way that the distribution of the "most representative" feature is kept as close as possible to the original one.

We will explain in section 5.2 how to obtain relative *importances* for each feature. For now we use these values in order to decide with respect to which feature we will

use stratified sampling. In order to predict X_{max} , the most important feature is X_0 and then, to a lesser extent, M_0 . The disadvantage of X_0 , that follows an exponential distribution, is the accumulation of values near zero. Thus, there is no guarantee that enough low X_0 values, corresponding to the lower end of the X_{max} distribution, will be kept in the training set. This problem is not present in the M_0 distribution. We confirmed with a few example networks that with the latter option better predictions are obtained. Its distribution can be seen in figure 5.1 (left). The most important feature for predicting N_μ is F_0 , which is depicted in figure 5.1 (right). Both distributions ensure having instances of extreme and less frequent values of M_0 and F_0 from which to learn about these special air showers. Since for the two target observables it is justified to use a different set for this procedure of stratified sampling, we decide to develop a separate model for each of them instead of developing a single model with a 2-dimensional output.

Another aspect to take care of before training is that neural networks work best when all the features have the same scale. We *standardize* the usual way by centering and scaling to unit variance each feature independently. The advantage with respect to *min-max-scaling*, where values are shifted and rescaled so that they end up ranging from 0 to 1, is that standardization is less sensitive to outliers. With respect to the target values it suffices to divide them by appropriate constants so that they cover similar ranges.

Nodes in neural networks are grouped in layers, with the first layer being the input layer (see figure 5.2, blue circles) which introduces the values of the features. The last layer (red circle) returns the target value. All the layers in between (grey circles) are

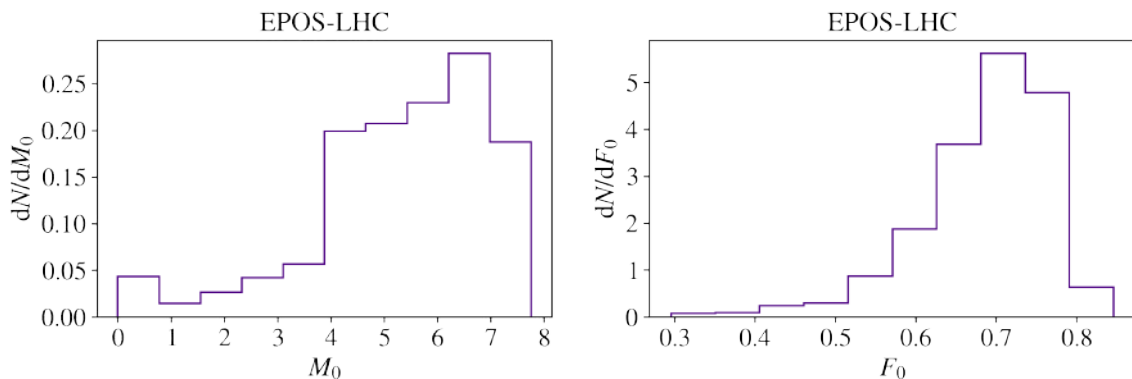


Figure 5.1: Correlations between features and targets suggest that M_0 (left) has the highest impact on X_{max} and F_0 (right) on N_μ . Since the distribution of N_{tot}^{FI} presents a very long tail, which typically cannot be properly dealt with in neural networks, we use from now on $\ln(N_{tot}^{FI} + 1)$ for the parameter we call M_0 . This distribution concentrates in a more compact region. Because of the relevance of these features, when we sample the training set out of the total dataset, we use stratified sampling with respect to these. Thus, we ensure that the original distributions are retained and extreme and less frequent values of M_0 and F_0 , which correspond to the extreme ends of the X_{max} and N_μ distributions, are present when learning.

the hidden layers. Each node from the hidden and output layers calculates a weighted sum (weights are symbolized by connecting arrows) of all the inputs it gets from the previous layer, adding a bias value (symbolized in figure 5.2 by the summatory). It then evaluates a predefined *activation function* f on this sum and outputs the result to the next layer. For example, for any node of the first hidden layer this can be summarized as:

$$f \left(\sum_i^n w_i x_i + b \right).$$

We will train neural networks with *fully connected* or *dense layers*, which means that all nodes of a layer are connected to all nodes from the adjacent layers.

Training a neural network means finding the values of the weights and biases (the model's parameters) for which the final nested function makes the most accurate predictions of the target values as a function of the features. This is done by means of the *backpropagation* training algorithm. It is essentially gradient descent with respect to the *loss function* which measures the error. The loss function can be any l_k norm of the difference between the vector of training instances and the vector of corresponding present predictions. The higher the norm index, the more it focuses on large values and neglects small ones, being so more sensitive to outliers. Since our $X_{max}-N_\mu$ distribution has quite some outliers we decide to work with the l_1 norm which is also called *mean absolute error*:

$$l_1(\bar{y}_{train}, \bar{y}_{predict}) = \frac{1}{m} \sum_{i=1}^m |y_{train}^{(i)} - y_{predict}^{(i)}|$$

where m is the number of instances in the training set. In order to minimize the error, weights and biases are updated following the gradient of the network's error with regard to every single model parameter.

Neural networks quickly get a large number of parameters (in 5.2 each arrow corresponds to a weight parameter, the biases are not shown there for simplicity but

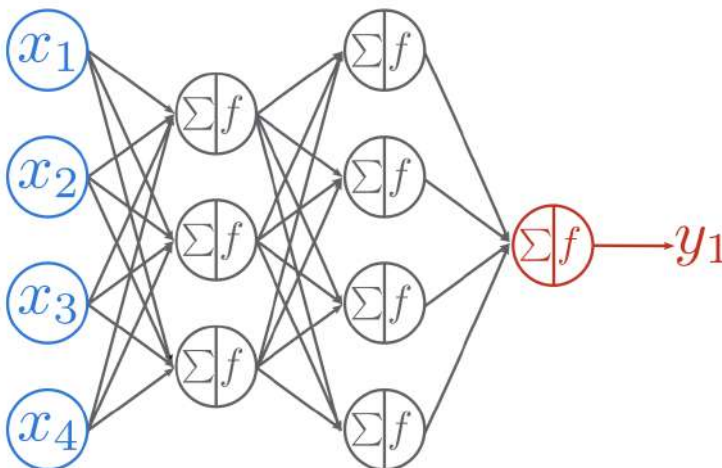


Figure 5.2: Outline of a neural network structure. The input layer (blue circles) introduces the input features. Nodes in hidden layers (grey circles) compute weighted sums of the outputs from previous layers and pass them on to the next layer through an activation function. After calculating optimal values for the weights and biases this nested function is the model which can associate to each combination of feature values a unique output value (red circle).

each one gives another set of parameters) which implies a lot of flexibility to fit the training data. On the counterpart, it also means that the model runs the risk of overfitting. Constraining the model in order to make it simpler and thus reduce the risk of overfitting is called *regularization*. Often, certain hyperparameters have the role of regularizing the model. A *hyperparameter* is a parameter of the learning algorithm and not the model. Hyperparameters are defined prior to training and remain constant during training. Examples of hyperparameters in this context are the depth of the neural network and the amount of nodes in each hidden layer. Our task is then to define all hyperparameters in such a way that there is enough room for flexibility when fitting but that there are enough constraints to avoid overfitting [130]. In the next few paragraphs the choices regarding regularization techniques, which give good results in the context of the present work, will be presented.

Regarding the backpropagation step several decisions need to be made. Instead of applying backpropagation to the complete training set we decide to compute gradients on small random sets of instances called *mini-batches*. In each *epoch* the training set is divided into mini-batches which are used in the algorithm one by one. The advantage of doing so is that the algorithm is much faster. Furthermore and even more importantly, it can also be favourable for convergence to a global minimum of the loss function [131]. Randomness usually helps escaping from local minima. Several different sizes of mini-batches were tested and a value of 16 gives the best results for both neural networks (for modeling X_{max} and N_μ). In each backpropagation step after calculating the direction of steepest descent, the weights and biases are updated through small corrections. These corrections are the opposite of the gradient times the *learning rate hyperparameter*. We chose a rather small value of 0.0005 for the learning rate in order to make sure that the solution will converge. As a way of avoiding that convergence is too slow we make use of *momentum optimization*. This method mimics friction by adding a momentum vector to the update. This vector is scaled by another hyperparameter that we set at a conservative value of 0.9. With respect to the number of epochs we give a rather high upper limit of 5000 which is never achieved because of another regularization technique we describe next.

All neural networks presented in this work are trained using *early stopping*. This means that after a predefined number of epochs with no improvement on the validation set, training is stopped. After a few examples of neural networks under different conditions, we set this number to 25. The weights and biases of the model that performed best on the validation set before stopping are stored. That is why we have set the number of epochs to a large value before. Training will be interrupted anyway as soon as there is no progress and setting the number of epochs to a high value ensures that training can continue until this optimal situation is reached.

A typical phenomenon when training a deep neural network with classical settings

is that the gradients of the loss function decrease as the algorithm progresses down to the lower layers (those close to the input). This means that when updating the weights' and biases' values, the lowest layers are left almost unchanged and thus convergence to a good solution is extremely slow. This could indeed be observed in our context. Glorot and Bengio discovered a way to overcome this issue [132]. It was understood that, in order to assign equal importance to all layers, the variance of the inputs of each layer needs to be equal to the variance of its outputs and that the gradients need to have equal variance before and after “flowing” through a layer during the backpropagation step. This can be achieved by certain combinations of activation functions and initialization strategies for the starting values of the weights. The combination that works best in this work is the *Relu function* (short for Rectified Linear Activation Function) together with the *He initialization* with a normal distribution. The Relu activation function is defined as:

$$f(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

He initialization with a normal distribution means that the starting weights are chosen from a normal distribution with mean 0 and variance $2/n_{in}$ where n_{in} is the number of inputs of the corresponding node.

Another regularization technique we employ is l_2 *regularization*. Here, a penalty of the form $\alpha \cdot (\sum w_i^2 + \sum b_j^2)$ is added to the loss function. It is intended to prevent the parameters of the model from growing indiscriminately, which is empirically been associated to model overfitting. The higher the value of α the stronger the constraint. We use $\alpha = 0.005$ in all our models, which in combination with the configuration chosen up to now has proven to give good results.

The configuration presented so far is chosen following suggested values from literature [133, 134, 135], keeping in mind what our goal is and the structure our data has. The only decisions we leave open is the determination of the number of layers and the number of nodes in each layer. For all datasets we are interested in modeling, we resort to randomized parameter optimization which will be explained in section 5.2. With this algorithm we get a rough idea of promising architectures. Then we finetune by training neural networks using many combinations close to these values.

The first dataset we are interested in modeling is the set of simulations done with vertical protons of 10^{20} eV as primaries. The anticorrelation between X_{max} and N_μ is strongest for this kind of primaries and we want to see if, using the features we decided upon, we can describe it. We have a dataset of 1000 instances for each of the following high-energy interaction models: EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d. As we have seen in section 3.3, each model handles physical processes of hadronic particle production at the highest energies differently. It is interesting to investigate to what extent it is possible to capture the differences between the high-energy interaction

models in form of feature variables. One might think that, since the models differ only at the highest energies, once those processes are summarized in the form of parameters, the description of the rest of the shower should be common to all three models. To answer this question we first train neural networks for each dataset separately and then on a joint ensemble of the three datasets.

When modeling X_{max} , we obtained the best results with a common architecture of 4 layers with 36 nodes each for all three high-energy interaction models. When modeling N_μ , architectures need to be different in order to get the best results individually: 5 layers of 60 nodes, 4 layers of 36 nodes and 5 layers of 40 nodes are used for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively. The learning curves for both models describing the EPOS-LHC data are shown in figure 5.3. These are the loss functions for the training and validation set. As is usual, the model performs better for the training set. The early stopping method prevents the gap between the two curves of becoming too large. In figures 5.4 and 5.5 the predictions are compared to the real values. The predictions of X_{max} have absolute errors around 20 g/cm². In the case of N_μ the relative errors vary more from model to model: 6.7%, 5.0%, and 5.9% for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively.

These small errors are evidence that X_{max} and N_μ can be very well described by the 7 features chosen. It also becomes evident that there is a bit more difficulty in describing the number of muons and that the high-energy interaction models are here captured slightly differently well by the chosen features. The comparison between the complete distributions is shown in figure 5.6. The anticorrelation is well reproduced for each scenario. All the axes are fixed at the same values to serve as a guidance. The differences in the high-energy interaction models lead to differently distributed X_{max} - N_μ pairs. While EPOS-LHC and SIBYLL-2.3d have similar distributions, QGSJETII-04 presents a quite different arrangement. The anti-correlation is less pronounced and X_{max} and N_μ reach lower values.

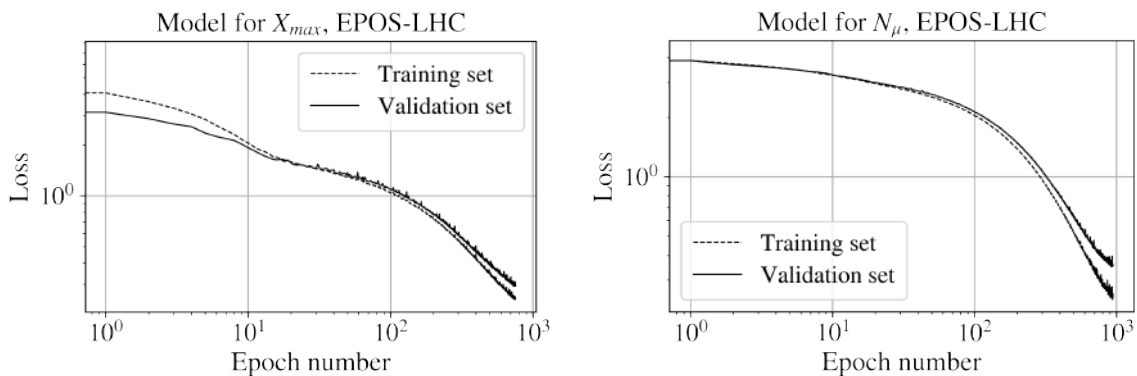


Figure 5.3: Learning curves of the neural networks trained to predict X_{max} (left) and N_μ (right), respectively. These networks are trained on the dataset of vertical proton primaries of 10²⁰ eV simulated with CONEX using EPOS-LHC.

Ultimately, we are interested in obtaining a unified model that describes all scenarios simultaneously and thus is model-independent. This way we can apply it to data knowing that it is valid whatever scenario represents reality best. Of course, this will only work if the real scenario is somewhat close to any of these three. Even though the high-energy interaction models used here have some caveats and inconsistencies, it is still to be expected that they represent the interactions quite well. With these assumptions made, we proceed now to developing a neural network that is trained on the joint ensemble consisting of simulations done with all three high-energy interaction models.

Since we want to apply our model to data and 10^{20} eV proton primaries are rather rare we also move on to another dataset in this step. We saw in section 1.2.4 that mass composition depends on primary energy and the high-energy interaction model

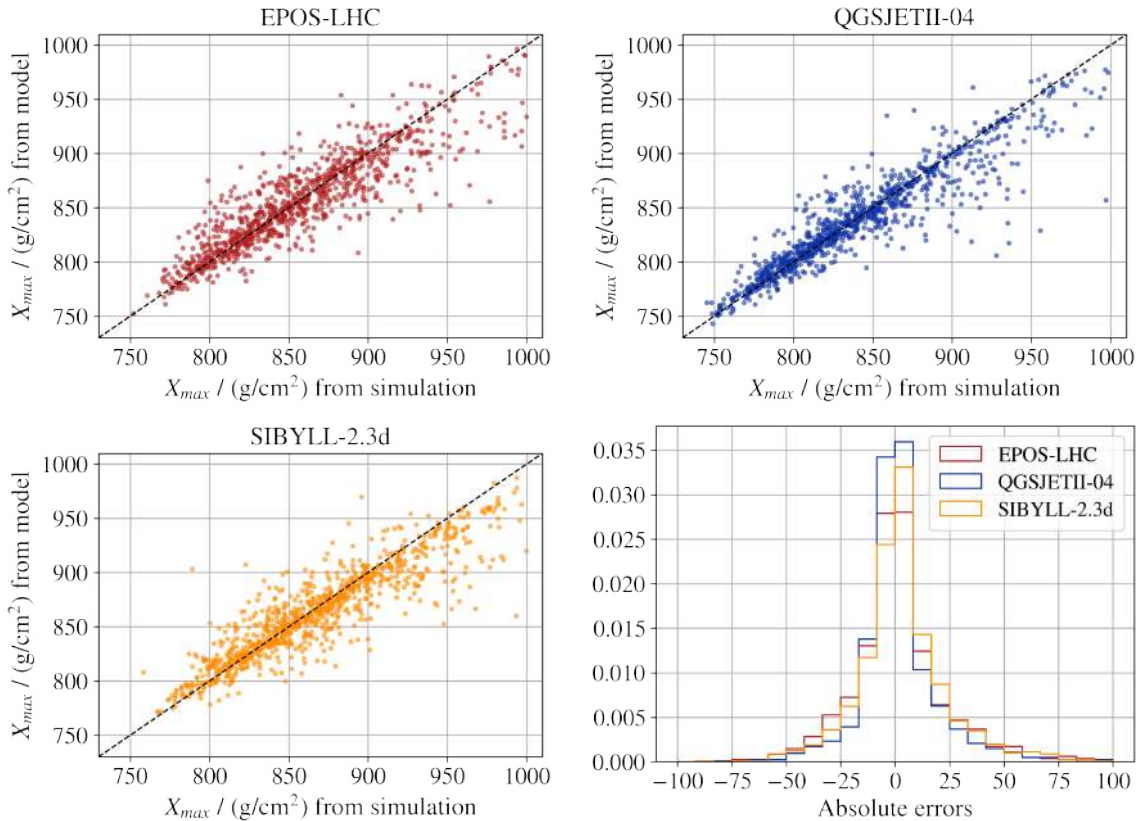


Figure 5.4: We investigate initially how well neural networks perform on showers generated by vertical protons of 10^{20} eV because the anticorrelation is strongest for this primary group and it is desirable to see it explained through these models. The first approach is to consider the datasets corresponding to the different high-energy interaction models (EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d) separately. One neural network is trained on each of the training sets corresponding to these high-energy interaction models in order to predict X_{max} . The comparison between the true values from simulation and the values obtained through each of the models is shown for EPOS-LHC (top left), QGSJETII-04 (top right) and SIBYLL-2.3d (bottom left). The black dashed lines representing $y = x$ are added for visual guidance. Finally, the relative errors between true values and predictions of each of the models are summarized in the histogram on the bottom right. The absolute errors are of about 22 g/cm^2 , 19 g/cm^2 , and 22 g/cm^2 , respectively.

considered. We want to avoid contamination of primaries other than protons as much as possible because our reasoning was done considering proton primaries and we don't expect to be able to develop a model that is also completely independent of the primary mass without introducing some new features. It is reasonable to think that we can deal with helium contamination though. The main problem with heavy primaries is that their first interaction looks very different from that of proton primaries. A considerable fraction of the primary energy is taken by individual nucleons that break off from the primary, while other nucleons remain together in fragments until the second interaction. In the case of helium primaries, this effect is the least pronounced. On the other hand, showers generated by iron primaries have so many muons that their $X_{max} - N_\mu$ distributions almost don't overlap with that for protons (see figure 3.3). The aim is then to decide on an energy range where the nitrogen contamination is as low as possible compared to the proton fraction. At the same time it should be feasible to obtain enough events in this energy range for which we know X_{max} and N_μ .

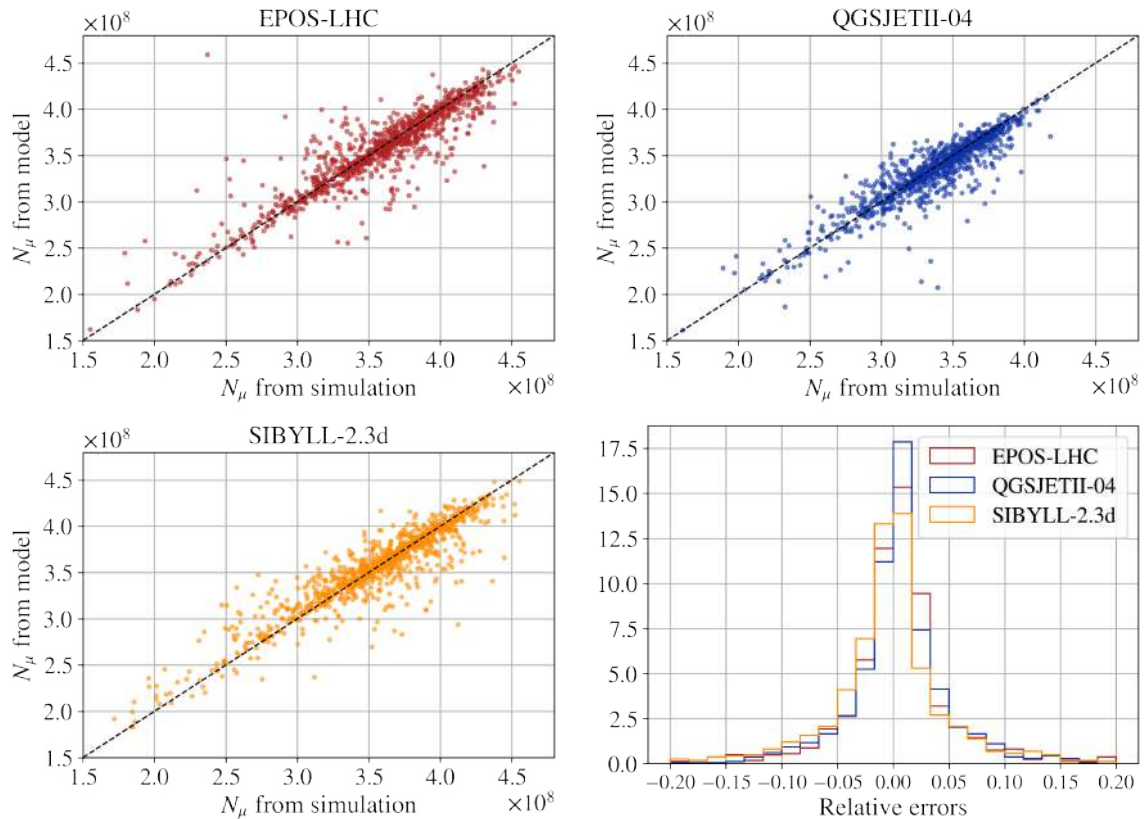


Figure 5.5: As presented in figure 5.4, a neural network is trained on each training set corresponding to the different high-energy interaction models separately, but this time in order to predict the number of muons N_μ . The comparison between the true values from simulation and the values obtained through each of the models is shown for EPOS-LHC (top left), QGSJETII-04 (top right) and SIBYLL-2.3d (bottom left). The black dashed lines representing $y = x$ are added for visual guidance. Finally, the relative errors between true values and predictions of each of the models are summarized in the histogram on the bottom right. The relative errors are of about 6.7%, 5.0%, and 5.9%, respectively.

With that in mind we decide on the energy range $\ln(E_0) = 18.098$ to 18.198 . We will come back to the subject of composition at these energies but for now we concentrate on datasets with proton showers only. We simulate 1000 showers for each high-energy interaction model and chose the primary energy from the distribution E^{-a} with spectral index $a = 3.27$ and in the chosen energy range. This time we use inclined showers with a zenith angle of 38° . The only change we need to make here is the addition of the primary energy as a new feature. The chosen energy interval seems small but we observed that the model improved considerably when distinguishing between primary energies during the training step. Furthermore, this range in energy may seem small, but a difference of 0.1 in $\log_{10}(E)$ implies a difference of about 25% in the number of muons. Thus, energy needs to be taken into account in the model.

We use the same setting as before with the exception that we search again for the best number of layers and nodes per layer. We obtained the best unified network for X_{max} prediction using 4 layers of 36 nodes each. The result is shown in figure 5.7. The absolute errors are slightly higher than for the separately trained networks but have similar distributions. For the network predicting N_μ 4 layers of 45 nodes gave the best result (see figure 5.8). Also here are the relative errors in general slightly higher

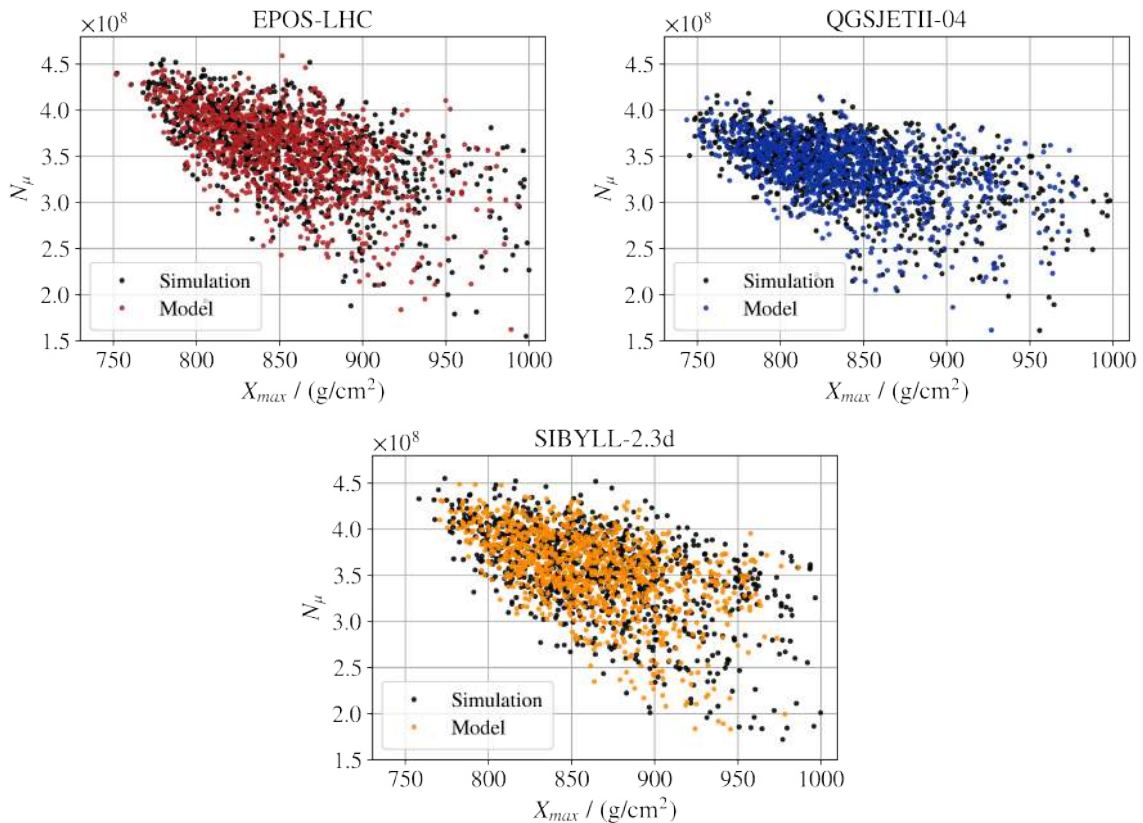


Figure 5.6: The results shown in figures 5.4 and 5.5 are combined in order to compare the X_{max} - N_μ distributions from simulations (black dots) to the distributions obtained through our models (coloured dots) for each high-energy interaction model. The anticorrelation is well reproduced in each case. All the axes are fixed at the same values to serve as a guidance.

than before but also distributed very similarly. Only a small bias for EPOS-LHC and QGSJETII-04 in opposite directions can be observed. The increase in the errors with respect to the previous models means that some compromise needs to be made in order to get a unified model. However, in this context these errors are small enough. Most importantly, the similarity in the error distributions means that the differences between the high-energy interaction models are captured by the features we used. If this was not the case, the model would need to compensate this missing ingredient. Since we are training with an ensemble that integrates all three high-energy interaction models this would result in an intermediate model. For the prediction of N_μ this is mildly observed in the form of small and opposite biases. The fact that these are small means that the model we obtained is model-independent.

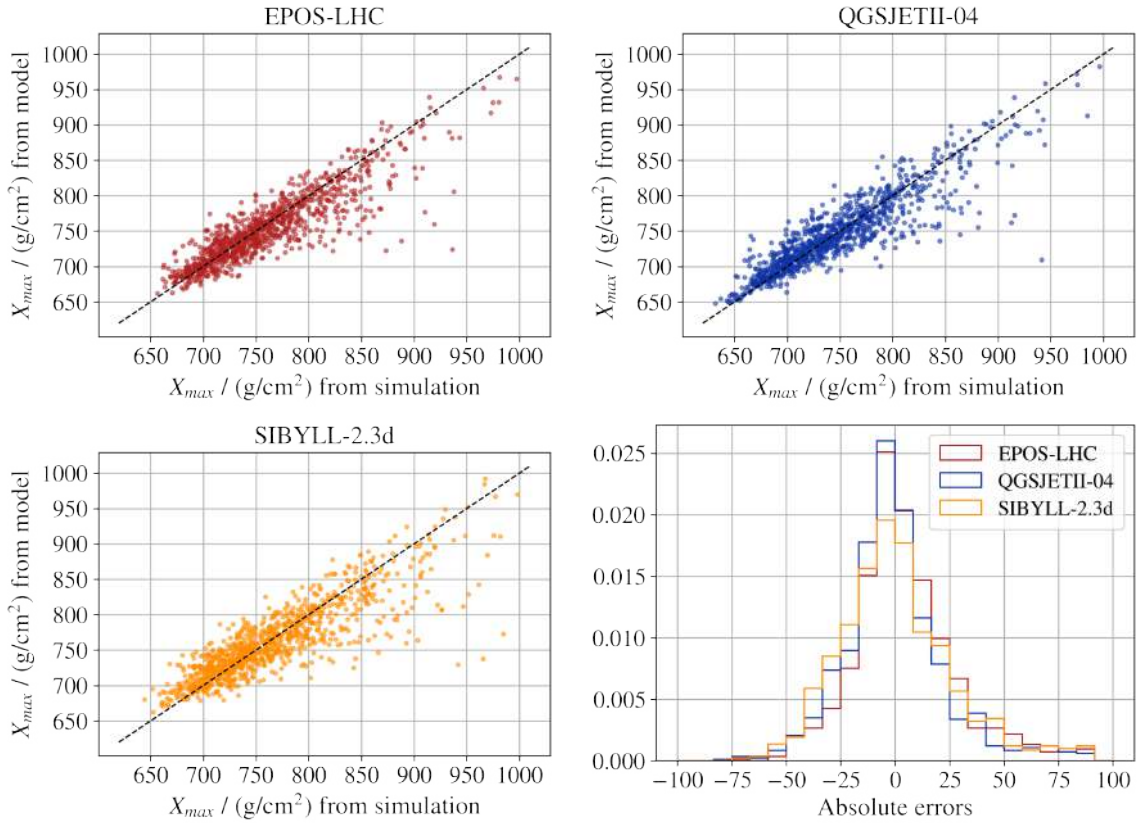


Figure 5.7: We now use the complete dataset of simulated showers generated by protons of primary energy between $\ln(E_0) = 18.098$ eV and 18.198 eV comprising all three high-energy interaction models. The goal is to obtain a model that predicts the X_{max} - N_μ anticorrelation independently of the high-energy interaction model used when simulating. This is done shuffling the ensemble of all three datasets and using the result as the new dataset. Thus the neural network is able to learn the structures that are common to the three scenarios. Since we now have a small interval of possible primary energies it turns out to be fruitful to incorporate E_0 as a new feature. The result for the network trained to obtain X_{max} is tested then on each subdataset separately and shown here. The absolute errors are slightly higher than for the neural networks trained on each data separately: 27 g/cm², 26 g/cm² and 32 g/cm² for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively. Errors are simultaneously centered at zero for all three scenarios and no relative bias is observed. From this we conclude that the goal to capture all three scenarios in one single and unified model is achieved.

Finally, we compare the complete X_{max} - N_μ distributions from simulations to the distributions obtained through our unified model (see figure 5.9). The overlapping is very good. All structures present in these distributions are well covered. As for 10^{20} eV protons the QGSJETII-04 distribution stands out with its lower X_{max} and N_μ values. For these primary energies the EPOS-LHC and SIBYLL-2.3d distributions differ more noticeably. The SIBYLL-2.3d distribution is a bit pointier at the highest N_μ values and is generally slightly shifted upwards. In order to come closer to a model suitable for application to data, however, we need to reduce the number of features.

5.2. Feature selection

We used in the previous section six features and then included the primary energy as input to our model and obtained a very good predictability. However, we know that real data is affected by uncertainties in the measurements and the reconstruction. Furthermore, the cosmic ray flux comprises all types of primaries ranging from proton

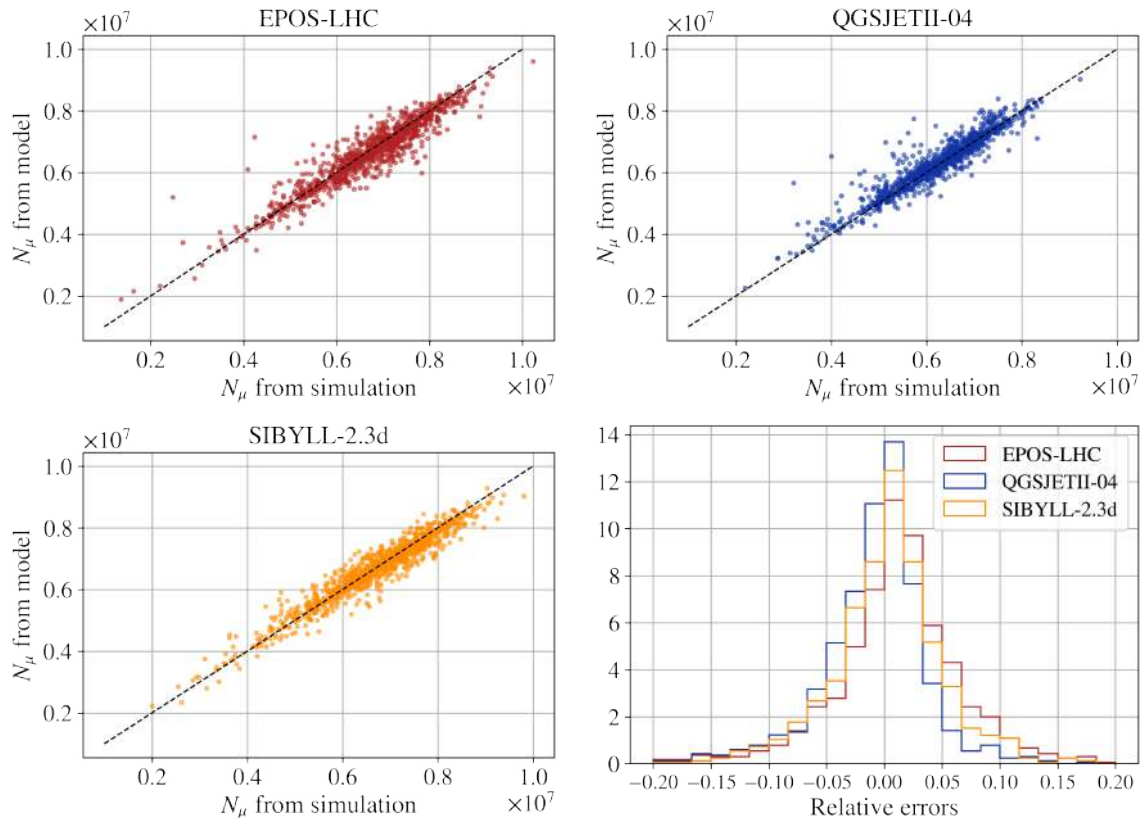


Figure 5.8: The same procedure as presented in figure 5.7 is carried out with N_μ as the target. The relative errors are again slightly higher (except for SIBYLL-2.3d) than for separately trained models: 7.4 %, 6.3 % and 5.1 % for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively. However, these errors are lower than typical errors involved when measuring or predicting the number of muons in other ways. The goal to capture all three scenarios in one single and unified model is achieved. A slight bias in the distributions of the relative errors for EPOS-LHC and QGSJETII-04 in opposite directions can be observed.

to iron. Since, as of today, we are not able to infer the primary mass on a shower-to-shower basis, we will need to consider this mixed composition scenario in our model. But for now it becomes clear that it is unrealistic to expect our model to be able to associate, under these conditions, to each combination of X_{max} and N_μ a specific combination of physical parameters, even if we were able to reduce the set of features to two elements (in order to avoid any ambiguity). The already mentioned uncertainties together with the fluctuations in the shower not taken into account by the reduced set of features, would make individual predictions very unreliable. It is more realistic to consider the complete distributions as our object of study and we will show that this indeed gives the desired result. This is by no means disappointing. If we were able to deduce the values of the physical parameters under study on a shower-to-shower basis, they by themselves would carry no information about the physical processes involved. We would still analyze their distributions to extract physical parameters anyway.

Concerning the further training and model construction, nothing has changed. We only need to keep in mind that with the final model we will analyze the relationship between the distributions of the features and those of the targets. In order to charac-

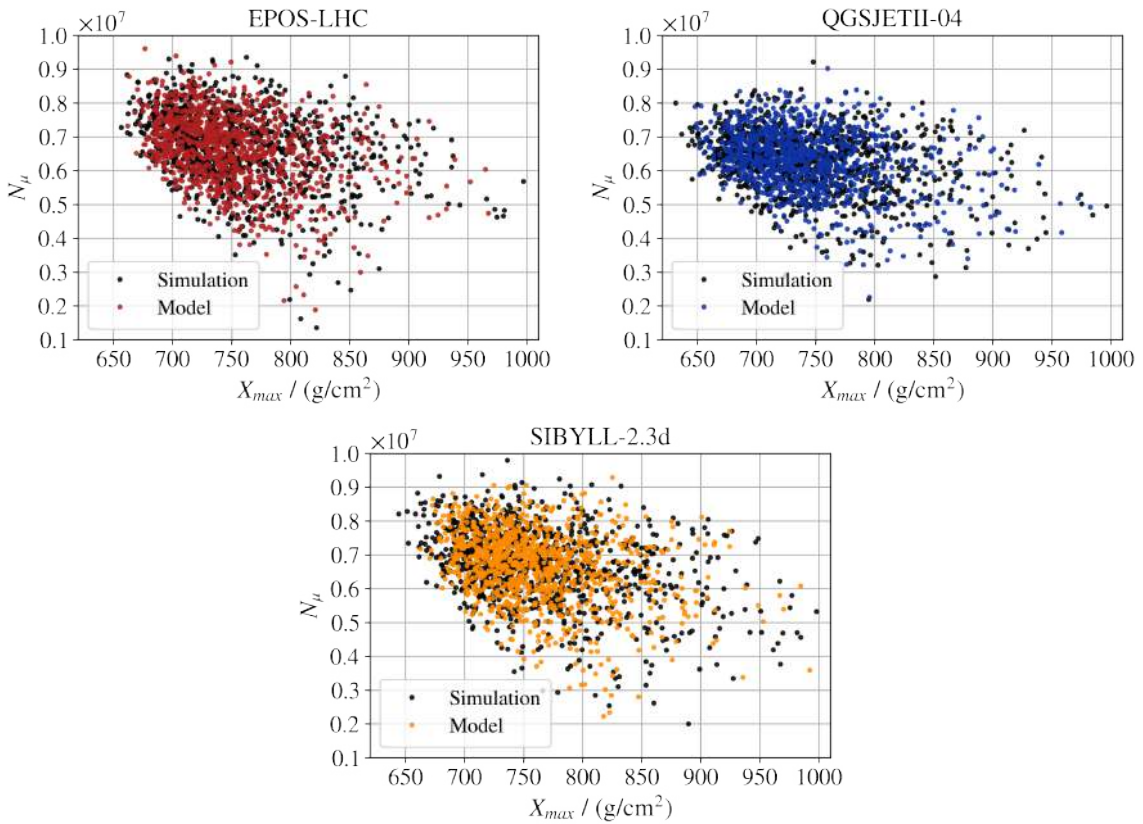


Figure 5.9: The results shown in figures 5.7 and 5.8 are combined in order to compare the X_{max} - N_μ distributions from simulations (black dots) to the distributions obtained through our unified model (coloured dots) for each high-energy interaction model. Also for this energy group and using a single model to capture all three high-energy scenarios, the anticorrelation is well reproduced in each case. All the axes are fixed at the same values to serve as a guidance.

terize these distributions we will need many parameters. Note that even though we are not able to measure individual values of X_0 at the Pierre Auger Observatory, its distribution is well known and therefore we can keep it in our model “for free”. Nonetheless, we still have too many features and need to reduce their number. One way to decide which features can be discarded is using *random forests* which can indicate the relative importance of each attribute for making accurate predictions with this method. Even though we use neural networks to develop our model, this is an appropriate tool to analyze the importance of each feature in a systematic way.

Decision trees are a supervised learning method used for classification and regression [124, 136]. The goal is to create a model that predicts the value of a target variable by learning simple if-then-else decision rules inferred from the data features. A trained tree can be seen as a piecewise constant approximation. In each region the predicted value is the average target value of the instances in that region. The algorithm splits during the training process each region in such a way that makes most training instances as close as possible to the predicted mean value. Decision trees are prone to overfitting on data because they are very sensitive to small variation in the training data.

A way to improve the algorithm substantially is considering an *ensemble* of decision trees. The algorithm in which each decision tree is trained on a different subset of the training set and then all predictions are combined into a mean value as the final prediction is called a *random forest*. The multiple subsets are formed by making bootstrap replicates of the training set, i.e. they are samples extracted from the training set with replacement. Despite its simplicity this is a very powerful machine learning algorithm [137, 138]. In essence, due to the *law of large numbers* this type of combined predictors achieves a better generalization than the best predictor in the ensemble [137].

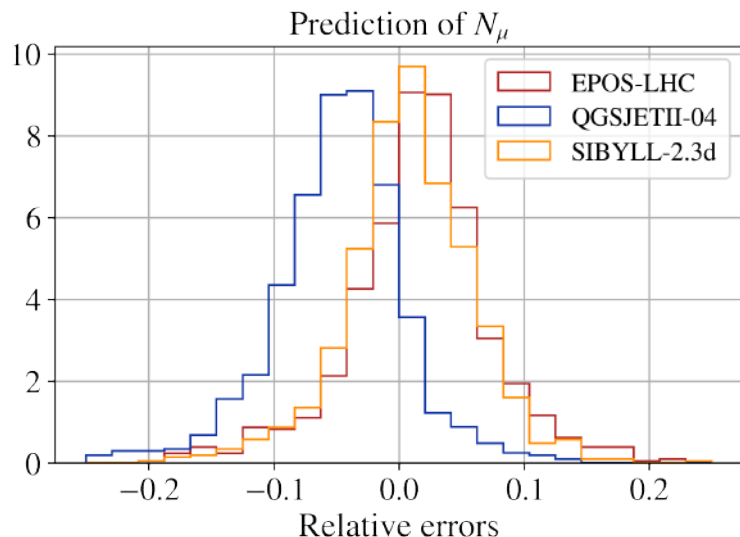
In order to use the attribute from the random forest regressor which gives the relative importance of each feature we need to fit this regressor to our data. In order to do so, many hyperparameters need to be tuned. We are interested in tuning the number of decision trees that will constitute the forest, the maximum depth of the trees (number of consecutive if-then-else decision rules), the number of features to consider when looking for the best split (these features are randomly chosen among the possible ones) and the minimum number of samples required to split an internal node (if a low number is allowed very detailed predictions that are only valid for a small region might be obtained, which leads to overfitting). In order to find the optimal hyperparameters we resort to *randomized parameter optimization*. From a grid of possible parameter settings that we propose, a predefined amount of combinations is randomly sampled. It has been shown empirically and theoretically that randomly chosen trials are more efficient for hyper-parameter optimization than trials on a complete grid [139]. Then *K-fold-cross-validation* is used to evaluate these combinations of hyperparameter values. Here the complete training set is separated in each evaluation into K subsets. Each subset is

used as a validation set after having trained the random forest on the rest of the subsets all together. The performance of each hyperparameter combination is the mean of the performances of the models on each of the K validation sets. Individual performances are measured via the mean squared errors in the corresponding validation set. For the hyperparameter combination that performs best we obtain relative importances of each feature. These are measured with the *Gini importance*, which counts the times a feature is used to split a node, weighted by the number of instances that are associated with the split. The higher this value the higher its importance.

We search 250 times among $\{50, 100, 500, 750, 1000, 1200\}$, $\{4, 5, 6, 8, 10, 12, 15, 20, 25\}$, $\{1, 2, 3, 4, 5, 6, 7\}$ and $\{5, 10, 15, 20, 25, 30\}$ for the number of decision trees, the maximum depth of the trees, the number of features to consider when looking for the best split and the minimum number of samples required to split an internal node, respectively. After performing 5-fold-cross-validation we obtain that the most important parameter for the prediction of X_{max} is X_0 with a Gini importance of 0.43 followed by M_0 , F_0 and K_0 with importances 0.25, 0.14 and 0.07, respectively. For the prediction of N_μ the most important features are F_0 , M_0 and K_0 with importances 0.68, 0.11 and 0.06, respectively. We already had an idea of which were the most important features for each model from analyzing the correlations and those are confirmed here. We now know in addition, which features are the next most important. It is an interesting result by itself to see that the three features inherent to the first interaction are the most important ones (taking X_0 out of the discussion since it will be treated as known). This is not only valid for the number of muons, which has already been discussed in the literature [66], but also for the depth of maximum development.

Since the next step will be to generate distributions of the features varying the parameters that describe them, the fact that K_0 is correlated with M_0 is a disadvantage.

Figure 5.10: Relative errors for N_μ prediction from a neural network trained using the same basic configuration as described in section 5.1 combined with 4 layers of 50 nodes each. The used features are E_0 , X_0 , M_0 and F_0 . These are only close to sufficient to capture all the differences between the high-energy interaction models. There is some mechanism missing in this reduced number of features which takes energy from the hadronic channel in QGSJETII-04 or which keeps energy in it for the other two models (or which does both).



Apart from needing parameters to describe the distributions themselves, we would need more parameters to describe the dependence, as well. The first attempt is to leave this feature out and keep the other most important features. For the reduced feature set comprising E_0, X_0, M_0 and F_0 a large variety of neural network architectures for N_μ prediction is evaluated. Just as a reminder, M_0 is the hadronic multiplicity of the first interaction and F_0 the hadronic energy fraction of the first interaction. To make sure that the best neural network possible is obtained, combinations varying all the hyperparameters are put to the test performing again a randomized search: combinations of 2 to 10 layers, 8 to 50 nodes per layer, learning rate between 0.001 and 0.0001, hyperparameter of l_2 -regularization between 0.001 and 0.01 and patience hyperparameter between 10 and 25 epochs were tested. After this broad search another more localized randomized search was performed.

When analyzing the relative errors between the real values and the predictions of N_μ , the result is always similar to the distributions shown in figure 5.10, which correspond to one particular case. The muon number is overestimated in showers generated

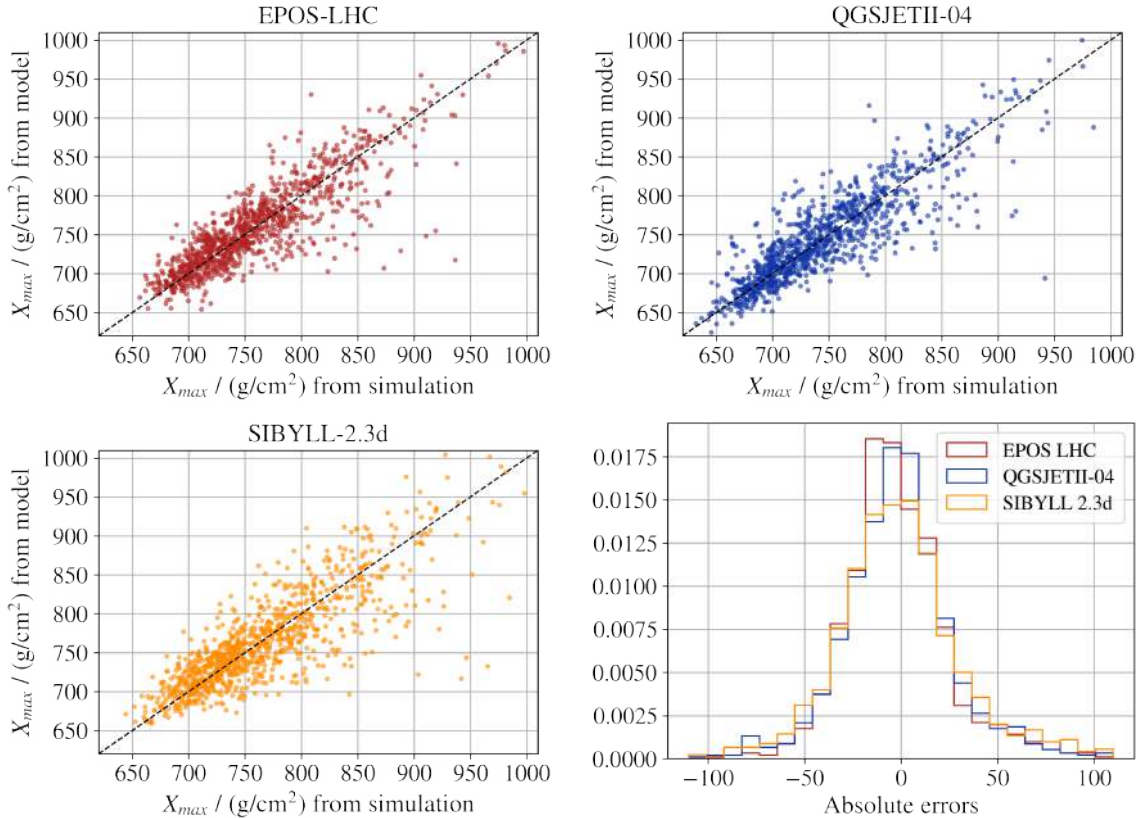


Figure 5.11: The same dataset of simulated showers as in figure 5.7 is used here. This time a reduced set of features comprising E_0, X_0, M_0, F_0 and K_{eff} is implemented. The result for the network trained to obtain X_{max} is tested then on each subdataset and shown here. The absolute errors are slightly higher than for the unified neural network trained using the complete feature set: 30 g/cm², 31 g/cm² and 37 g/cm² for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively. However, the errors are still small enough and the model makes predictions in a model independent way, as is needed.

with QGSJETII-04 and underestimated in showers generated with EPOS-LHC and SIBYLL-2.3d in a very similar way. This means that there is some mechanism missing in this reduced number of features which takes energy from the hadronic channel in QGSJETII-04 or which keeps energy in it for the other two models (or which does both). The best the neural network can do is find an “intermediate” model which has simultaneously for all high-energy models the smallest loss value. We will show in the next few paragraphs how to solve this issue.

The next attempt is to add one of the lesser important features to the feature set. When including M_{eff} or F_{eff} the missing mechanism is still not covered and results very similar to those observed in figure 5.10 are obtained, when attempting to model N_μ . The issue is solved when including K_{eff} , which represents the effective inelasticity of the whole shower, as a feature. Indeed M_0 , F_0 and K_{eff} are not correlated.

The original basic architecture described in section 5.1 together with 4 layers of 32 nodes each gives the best result for predicting X_{max} . The results are shown in figure 5.11. The absolute errors are around 30 g/cm², 31 g/cm² and 37 g/cm² for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively. They are a bit higher than the

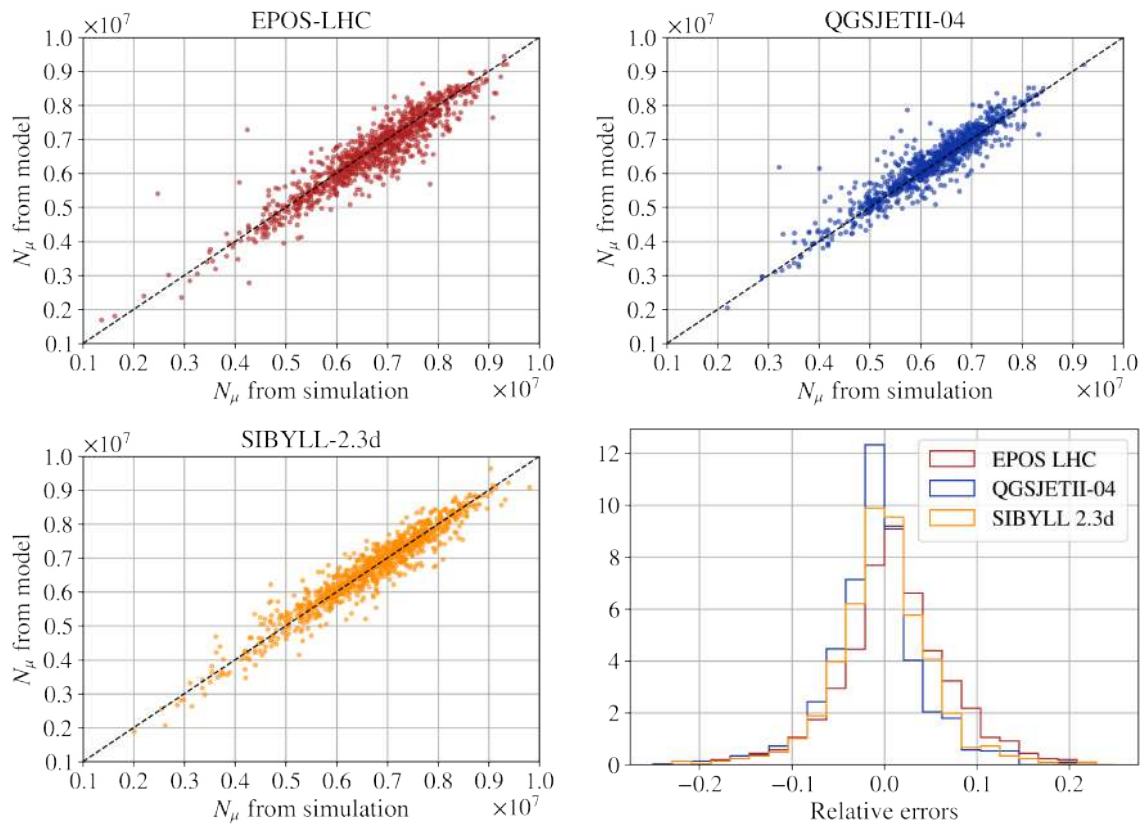


Figure 5.12: The same dataset of simulated showers and reduced set of features as in figure 5.11 is used to model the number of muons. The relative errors are slightly higher than for the unified neural network trained using the complete feature set: 7.7%, 6.5% and 5.4% for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively. However, the errors are still low and the model makes predictions in a model independent way, as is needed.

errors obtained when using the complete feature set, which is expected because leaving features out can lead to some ambiguity. But these errors are still small enough and the neural network gives predictions in a model independent way, as is needed.

For the neural network predicting N_μ the basic architecture described in section 5.1 also gives very good results. The best one is obtained using a neural network with 4 layers with 50 nodes each. The results can be seen in figure 5.12. The relative errors are

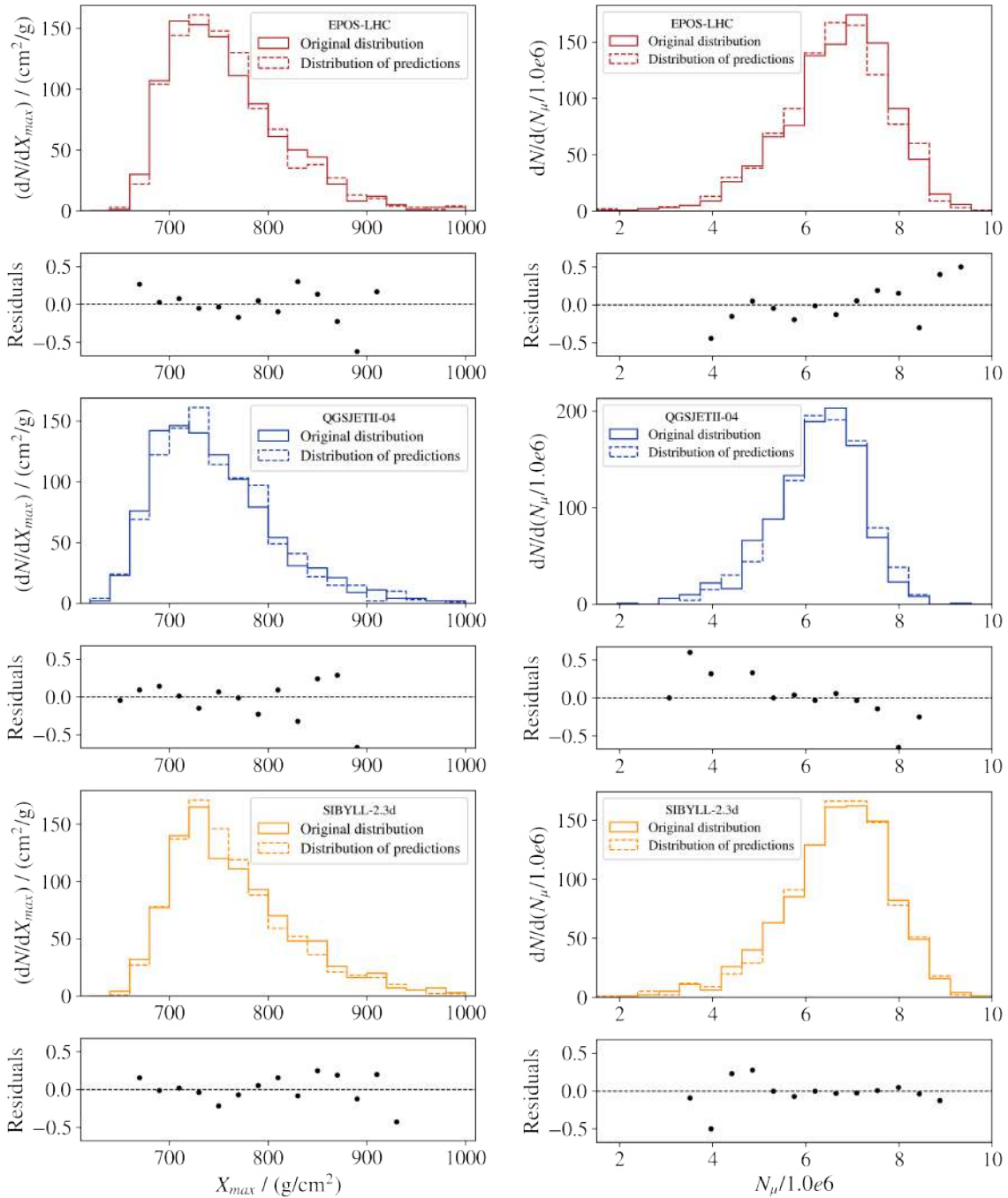


Figure 5.13: Comparison between the original distributions of X_{max} (left) and N_μ (right) and the distributions of the corresponding predictions obtained via the neural networks discussed in figures 5.11 and 5.12. Also the residuals for all combinations are shown.

slightly higher than for the unified neural network trained using the complete feature set: 7.7%, 6.5% and 5.4% for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively. It is interesting to note here that in the three settings for which we developed models for X_{max} and N_μ (based on neural networks) more nodes were required when modeling the number of muons and even so their predictions were always less accurate than when predicting X_{max} . This can be a symptom of the processes that build up the number of muons being more intricate than the combination of processes that lead to the value of X_{max} . After all, the main contribution to the electromagnetic channel comes from the neutral pions while the muons not only descend from charged pions and kaons, which already behave differently because of their different mean lifetimes, but also from other particles. Figure 5.12 also shows that the distributions of the relative errors are quite centered which means that we have indeed obtained a model for N_μ that is universal.

The ultimate goal is to compare the mean values and standard deviations of X_{max} and N_μ between distributions of interest and distributions obtained after inserting different possible distributions of the features from the reduced set into the neural networks presented in figures 5.11 and 5.12. Having that in mind, we compare now in figure 5.13 the original X_{max} and N_μ distributions corresponding to each high-energy interaction model (full lines) with the X_{max} and N_μ distributions obtained from inserting the corresponding original distributions of the features into these neural networks (dashed lines). The distributions are quite well represented. The mean value of X_{max} is through our predictions off by around 1 g/cm² in all cases. The dispersion of X_{max} is generally underestimated by around 1 g/cm². The mean value of the number of muons is very accurate for SIBYLL-2.3d. For EPOS-LHC it is underestimated by about 1.3% and for QGSJETII-04 overestimated by a around 1.2%. The dispersion is quite accurate for all cases and only high by about 0.4% for QGSJETII-04.

The neural networks we trained in this section use a low enough number of features and give remarkably good results when comparing the predicted final observables with the true ones. As a consequence this model can be used for further analysis.

5.3. Description of the features

In the previous section we trained neural networks to predict X_{max} and N_μ for a shower with specific values of E_0 , X_0 , F_0 , M_0 and K_{eff} . M_0 is the hadronic multiplicity of the first interaction, F_0 the energy fraction taken by hadronically interacting particles and K_{eff} the effective multiplicity of the rest of the shower. We now switch to thinking of this designation as one between distributions (see the depiction on figure 5.14). If we insert in our neural networks the distributions of the features we obtain as an output distributions of X_{max} and N_μ . The distribution of E_0 can be obtained directly from

the dataset of interest itself and is also already described (see section 1.2.1). As was reviewed in section 1.3, the distribution of X_0 is well known and thus the same for all three high-energy model scenarios. For the small interval of energies we are using it can be described by the exponential distribution

$$f(x) = \lambda e^{-\lambda x}$$

(for $x \geq 0$ and zero otherwise) with a common value of $\lambda = 47 \text{ g/cm}^2$ for all three high-energy interaction models (see figure 5.15, left).

The distributions of F_0 , M_0 and K_{eff} differ between high-energy interaction models and are unknown. We will see in this section that each of these features can be described with a known probability distribution function (pdf) which depends on two parameters: a *location* parameter denoting the position of the maximum of the distribution and a *scale* parameter describing the width of the distribution. We will refer to these with superscripts *loc* and *scale*, respectively. At this stage, the fact that the three features F_0 , M_0 and K_{eff} are not correlated presents a clear advantage. We can use the pdfs, that we define in the next paragraph, to generate these distributions, without the need to force a correlation in some way. Furthermore, at a later stage, when we generate distributions with slightly different basic parameters, it would be impossible to know which relation between such new parameters is the correct one.

F_0 and M_0 can be described using the Gumbel minimum distribution with pdf

$$f(x) = \frac{1}{b} e^{\frac{x-a}{b}} e^{-e^{\frac{x-a}{b}}}$$

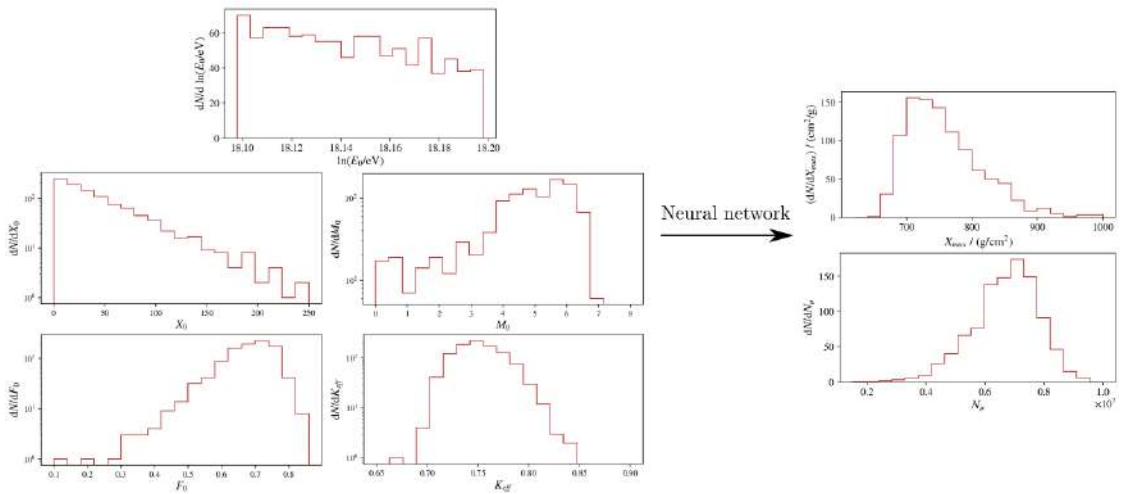


Figure 5.14: Depiction of our method. We developed in sections 5.1 and 5.2 a neural network that yields, on a shower-to-shower basis, X_{max} as a function of E_0 , X_0 , F_0 , M_0 and K_{eff} and another network that returns N_μ . Even though the model was trained on individual instances we now switch to thinking of this designation as one between distributions. For given distributions of the features used to train the neural networks we obtain distributions of the outputs X_{max} and N_μ . Therefore we need to describe the distributions of the features next.

where a is the location parameter and b the scale parameter. For the multiplicity M_0 it turns out that the scale parameter is very close to 1.05 for all three high-energy interaction models. Because of this accordance and because the impact this parameter has on the observables we will use is small (see explanation to figure 5.17), we will fix it at this mean value: $M_0^{scale} = 1.05$. Fitting the Gumbel distribution (see figure 5.15, right) with the scale parameter fixed gives $M_0^{loc}(E) = 5.32$, $M_0^{loc}(Q) = 5.78$ and $M_0^{loc}(S) = 5.40$ (E , Q and S in parenthesis stand for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively). In the case of the hadronic fraction F_0 there are no coincidences, so we need to leave both parameters free. The fitting parameters we obtain are $F_0^{loc}(E) = 0.725$, $F_0^{loc}(Q) = 0.708$ and $F_0^{loc}(S) = 0.714$ and $F_0^{scale}(E) = 0.066$, $F_0^{scale}(Q) = 0.058$ and $F_0^{scale}(S) = 0.068$ (see figure 5.16, left). K_{eff} follows approximately a normal distribution (see figure 5.16, right) whose width is very similar for all high-energy interaction models and is thus fixed at the mean value of $K_{eff}^{scale} = 0.024$. In figure 5.17 it becomes evident that changing this parameter has no significant effect on the observables. The mean values we obtained are $K_{eff}^{loc}(E) = 0.752$, $K_{eff}^{loc}(Q) = 0.813$ and $K_{eff}^{loc}(S) = 0.732$.

We have just seen that the three features with unknown distributions can be approximately described using 6 parameters, two of which we have fixed at their mean values. It is interesting to investigate the impact of changing these basic parameters on the observables we will use. Just like in [111] we vary these parameters by multiplying them by a factor. We do this for each parameter while leaving the other 5 parameters fixed at their values of the corresponding high-energy interaction model. We then generate distributions based on these altered parameters and feed them into our neural networks (like depicted in 5.14) to obtain the corresponding distributions of X_{max} and N_μ from which we calculate their mean values and standard deviations. The changes in the basic parameters are applied in such a way that the resulting ‘‘artificial’’ distributions of the features have some overlap with the region the original distributions covered and where the network was trained.

We also show as full black horizontal lines the true values of the corresponding observable and as dashed lines the 1σ error in the measurement of this value. These estimates are calculated using the *Bootstrap Method* [140]. This method is a statistical technique for estimating quantities about a population by averaging estimates from multiple small data samples. Data samples are extracted with replacement from the original dataset, in our case from the distributions of X_{max} and N_μ , respectively. From these distributions we extract the means and standard deviations of X_{max} and N_μ . Repeating this process we obtain distributions of these observables from which the mean values and standard deviations are represented by full and black lines, respectively.

The results can be seen in figures 5.17, 5.18 and 5.19 for the different high-energy interaction models. Each circle is the mean value of the corresponding observable after

repeating 15 times the procedure of generating the distributions of the features and extracting the observables. Each repetition consists of a set of 15000 showers. Thus we make sure that the prediction is very accurate. There are slight shifts between the values obtained using the proper original distributions and the fitted distributions (corresponding to $f = 1$). This comes from the fact that the distributions don't fit perfectly.

It is directly visible that changing K_{eff}^{scale} (purple circles) has almost no effect on any of the observables, which justifies fixing it at the constant mean value of the corresponding values from the different hadronic interaction models. Even though M_0^{scale} (yellow circles) shows influence on $\sigma(X_{max})$ it is the one parameter that has the least impact of all, after K_{eff}^{scale} . Furthermore, the true values for the different hadronic interaction models are very close. If we vary this parameter keeping it close to the realistic values, the effect on $\sigma(X_{max})$ is very small.

The location parameter of the charged multiplicity M_0 (red circles) has the strongest

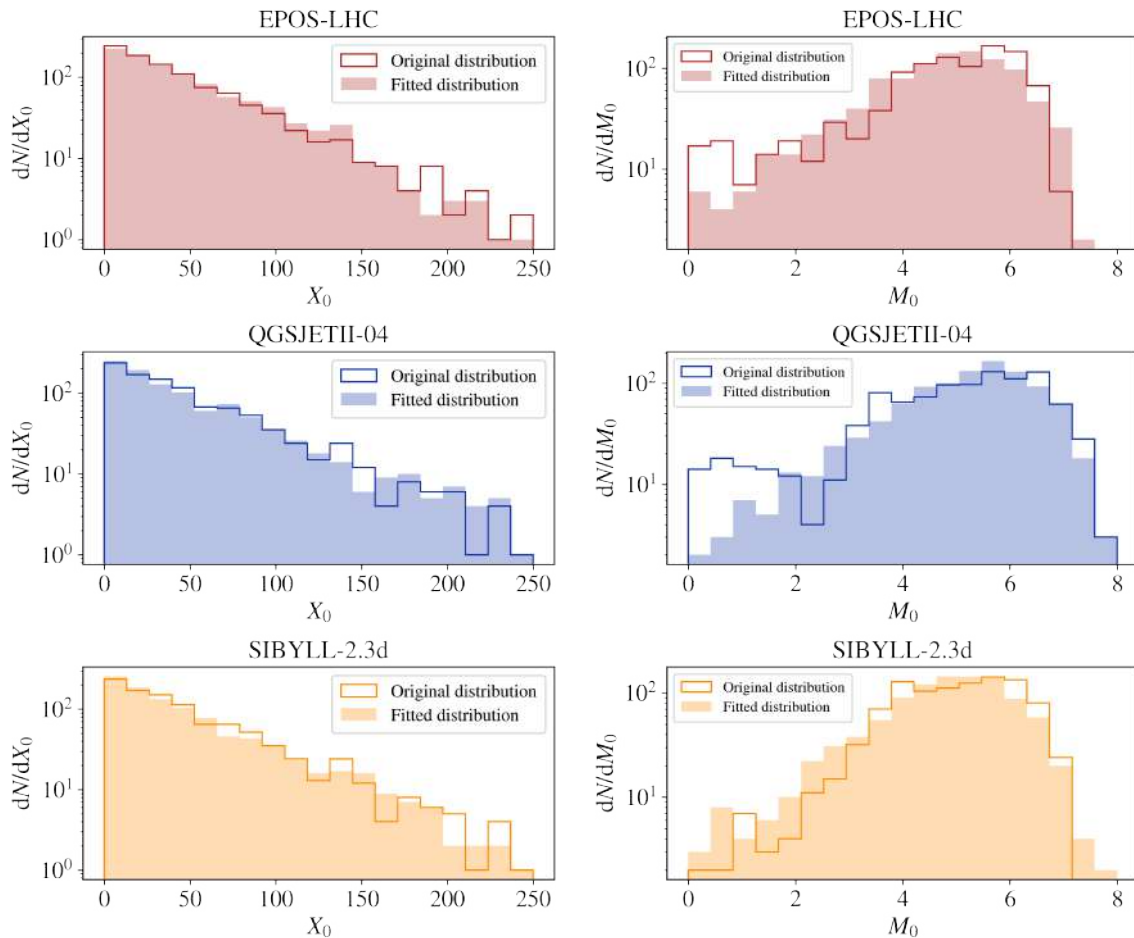


Figure 5.15: Distributions of X_0 (left) and M_0 (right). The depth of first interaction follows an exponential distribution with $\lambda = 47 \text{ g/cm}^2$. The hadronic multiplicity can be described with the pdf of the Gumbel minimum distribution. The scale parameter is very close to 1.05 for all three high-energy interaction models and is therefore fixed at that value.

impact in $\langle X_{max} \rangle$. For a higher value of M_0^{loc} the distribution of M_0 is shifted to higher values. For higher M_0 values the energy is distributed among more particles in the first interaction. This generates more subshowers of less energy, which develop quicker in the atmosphere and finally give a smaller overall value of X_{max} . At the same time more energy is kept in the hadronic channel resulting in a higher number of muons visible as an increasing value of $\langle N_\mu \rangle$ as a function of the factor f . The effect on $\sigma(X_{max})$ is more intricate as can be deduced from the fact that it varies between high-energy interaction models. It might depend on other factors such as the value of F_0 . For QGSJETII-04 the distribution of F_0 is narrower. Having less variability in the energy that is available for the hadronically interacting particles of the first interaction might lead to less ways to distribute energy among particles anyway, leading to a more or less constant value of $\sigma(X_{max})$ in this case. Whereas for EPOS-LHC and SIBYLL-2.3d, which have wider distributions of F_0 , an increasing value of $\sigma(X_{max})$ with f can be observed. An increasing value of the scale parameter of the charged multiplicity M_0

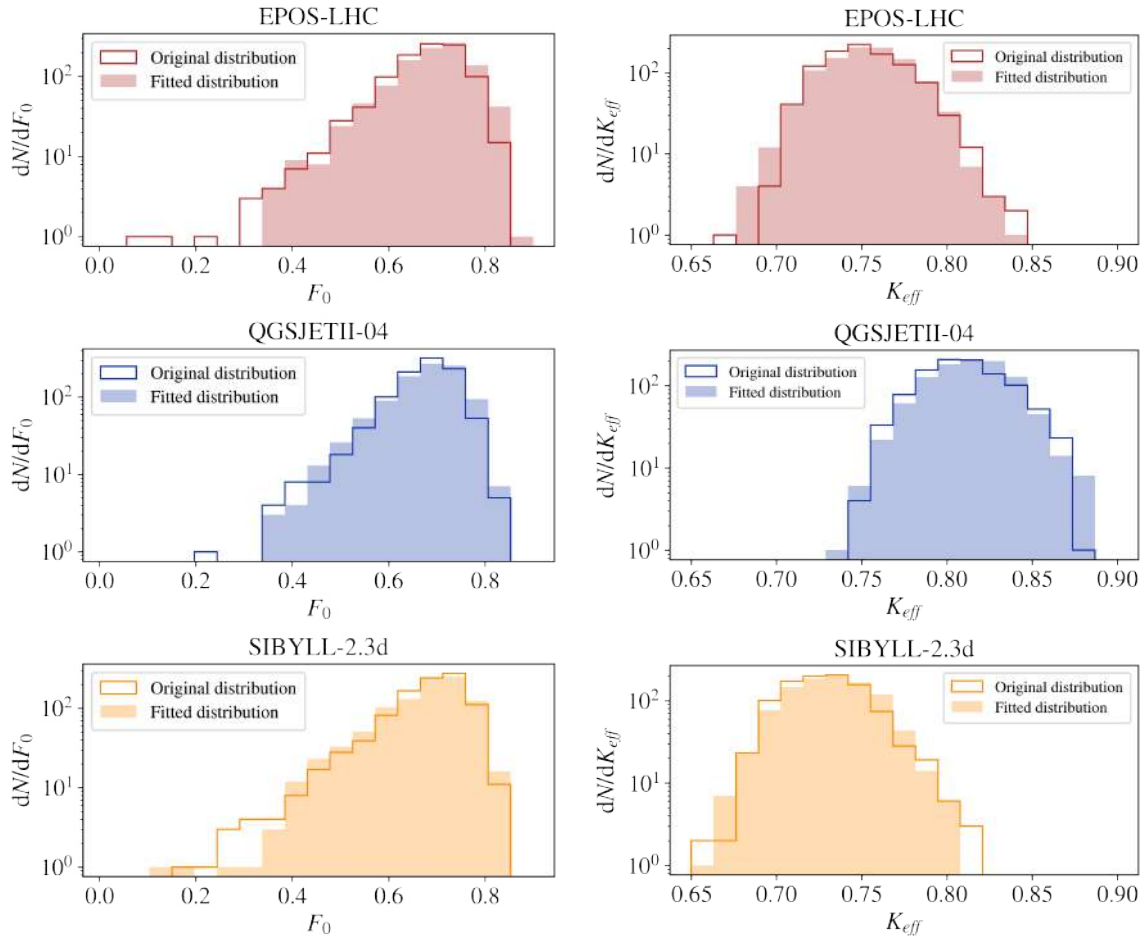


Figure 5.16: Distributions of F_0 (left) and K_{eff} (right). The hadronic fraction can be described with the probability distribution function of the Gumbel minimum distribution. The location and scale parameters vary for the hadronic interaction models and are thus left free. The distribution of K_{eff} can be approximated by a normal distribution with fixed standard deviation.

makes its distribution wider. Having more options for the hadronic multiplicity leads directly to more variability in the value of X_{max} and thus to a higher value of $\sigma(X_{max})$.

The hadronic fraction F_0 has a similar effect on $\langle X_{max} \rangle$ and $\langle N_\mu \rangle$ as the multiplicity M_0 , the difference being that the impact on the latter is stronger. The higher this fraction (obtained through a higher value of F_0^{loc} , green circles), the more energy is kept in the hadronic channel leading to a higher number of muons. At the same time the electromagnetic channel has less energy and gives a lower value of X_{max} . An increasing value of F_0^{loc} also leads to more variability as to how energy is distributed in the hadronic channel and thus to a higher value of $\sigma(N_\mu)$. $\sigma(X_{max})$ on the other hand has a parabolic shape with a minimum. The decreasing behaviour might be associated with the fact that with increasing value of F_0 less energy is diverted to the electromagnetic channel and thus there are less ways to distribute it among neutral pions and η particles in the first few interactions. From a certain point however, having a higher fraction in the hadronic channel of the first interactions also leads to more variability in the

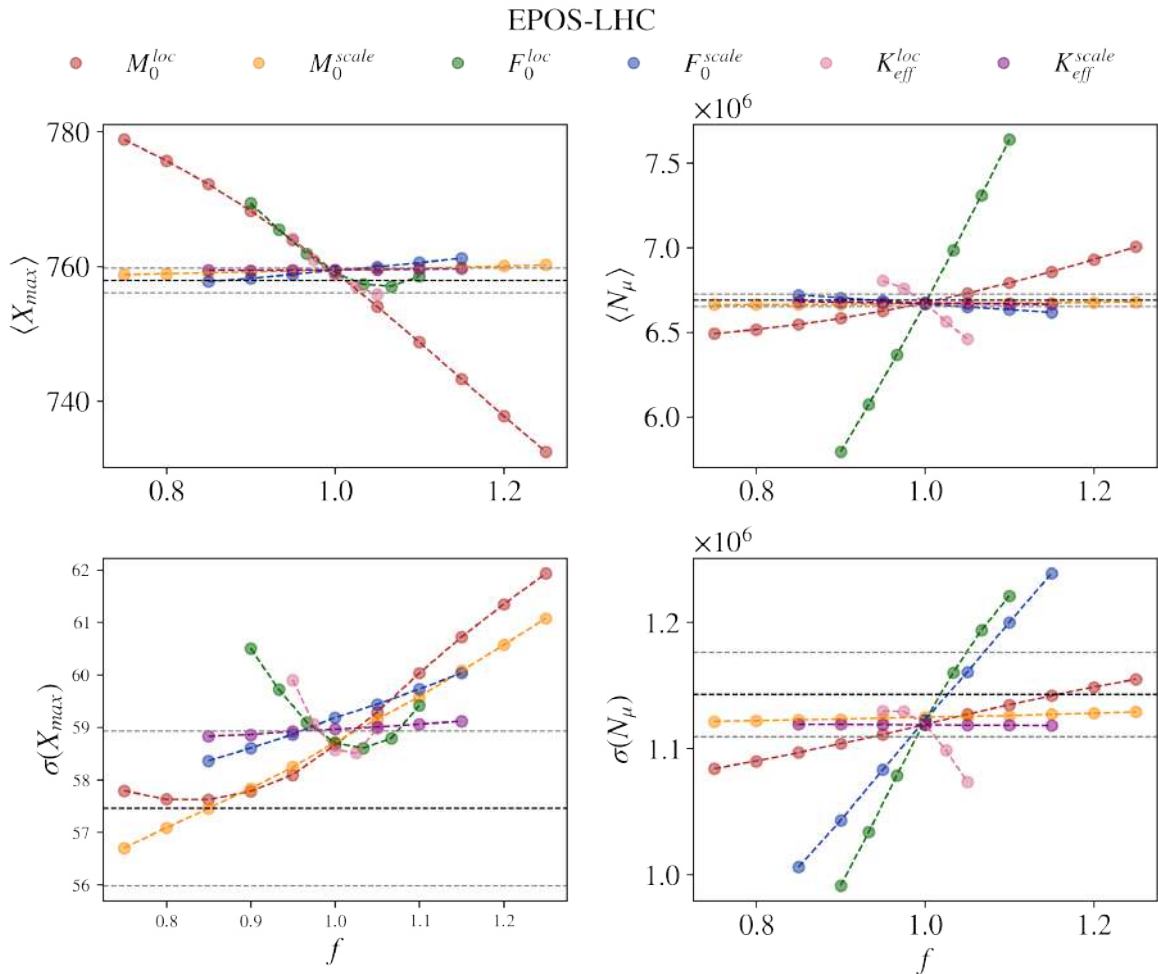


Figure 5.17: Impact of changing basic parameters used to describe the distributions of the features on the means and standard deviations of X_{max} and N_μ for EPOS-LHC.

electromagnetic component that descends from this group of particles and $\sigma(X_{max})$ increases again. An increasing value of the scale parameter of the fraction F_0 (blue circles) which widens its distribution leads to a higher variability in both observables.

The inelasticity K_{eff} has a more intricate behaviour. Its location parameter (pink circles) has in general a similar effect on $\langle X_{max} \rangle$ and $\sigma(X_{max})$ as the location parameter of F_0 , which is also in accordance with figure 3.16, even though our feature K_{eff} is a constant effective value while in [111] the elasticity is adapted in each interaction individually. It becomes evident that the leading particles of each interaction drive the bulk of electromagnetic particles deeper into the atmosphere. The effect on the number of muons is less pronounced. Evidently this number depends more on the fraction of energy which indeed ends up in the hadronic channel.

It is particularly interesting to notice that by increasing the location parameter of the F_0 distribution one quickly obtains a higher number of muons. This is important because we are interested in extracting the best fitting parameters for a dataset from

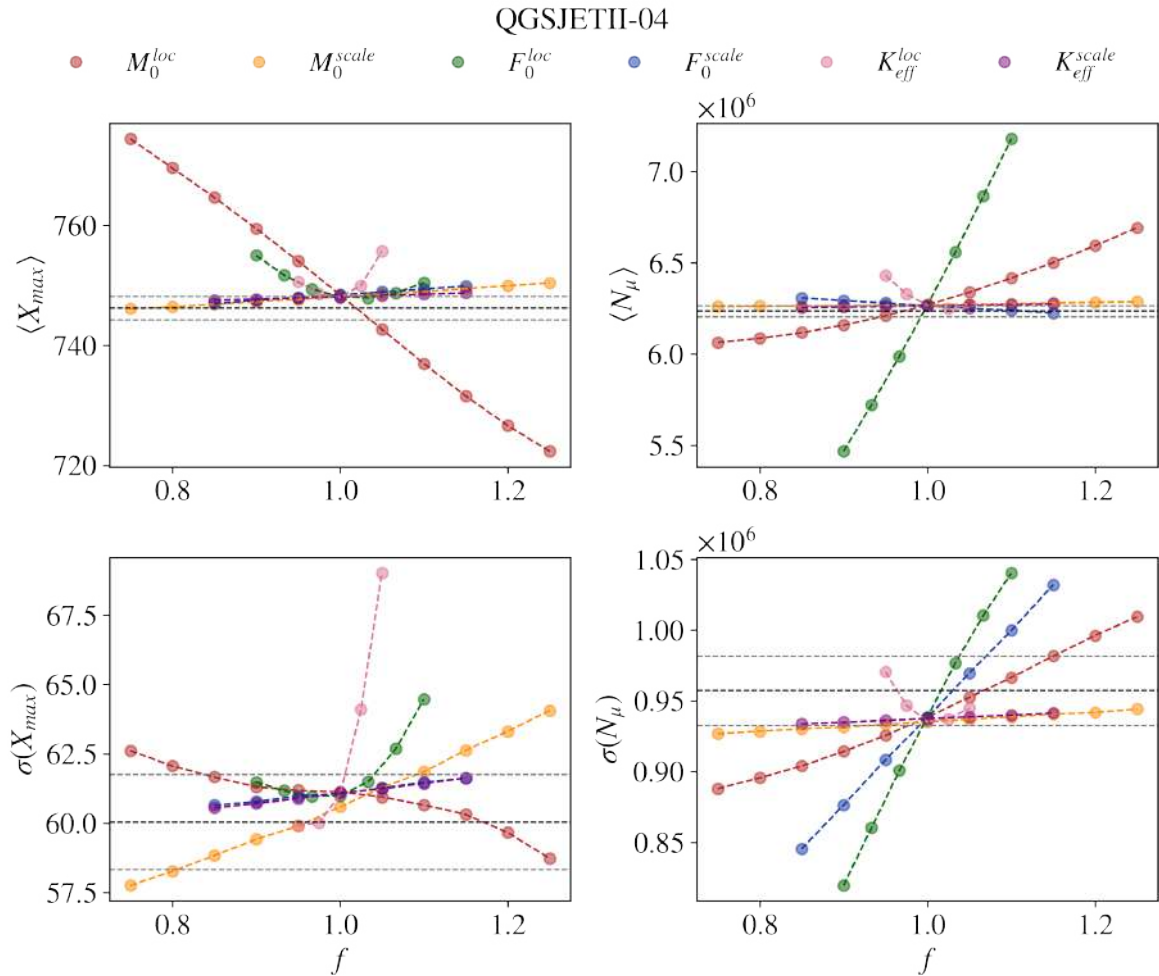


Figure 5.18: Impact of changing basic parameters on the means and standard deviations of X_{max} and N_μ for QGSJETII-04.

the Pierre Auger Observatory and, as was discussed in section 3.5, it is well known that the observed number of muons is significantly higher than the number obtained in simulations.

5.4. Performance of the model

We now proceed to perform a χ^2 -analysis comparing the observables $\langle X_{max} \rangle$, $\langle N_\mu \rangle$, $\sigma(X_{max})$ and $\sigma(N_\mu)$ in a 4-dimensional grid of possible values for the parameters M_0^{loc} , K_{eff}^{loc} , F_0^{loc} and F_0^{scale} . We do this for each high-energy interaction model in order to test the performance of our model in predicting the unknown parameters and in particular to test how well the model can distinguish between the high-energy interaction models. More in detail, we define

$$\chi^2(\bar{\theta}) = (\hat{x} - \bar{\mu})^T V^{-1} (\hat{x} - \bar{\mu}) \quad (5.1)$$

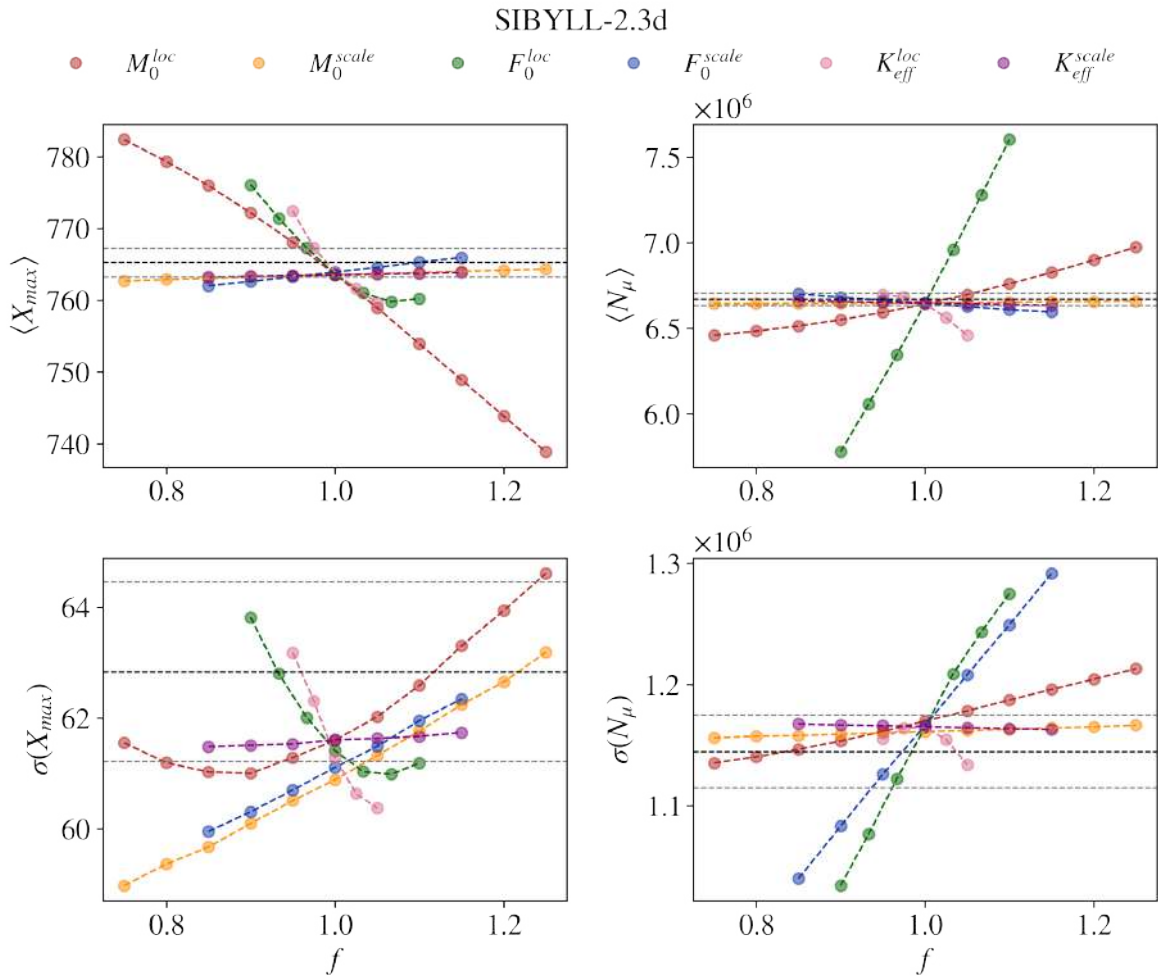


Figure 5.19: Impact of changing basic parameters on the means and standard deviations of X_{max} and N_μ for SIBYLL-2.3d.

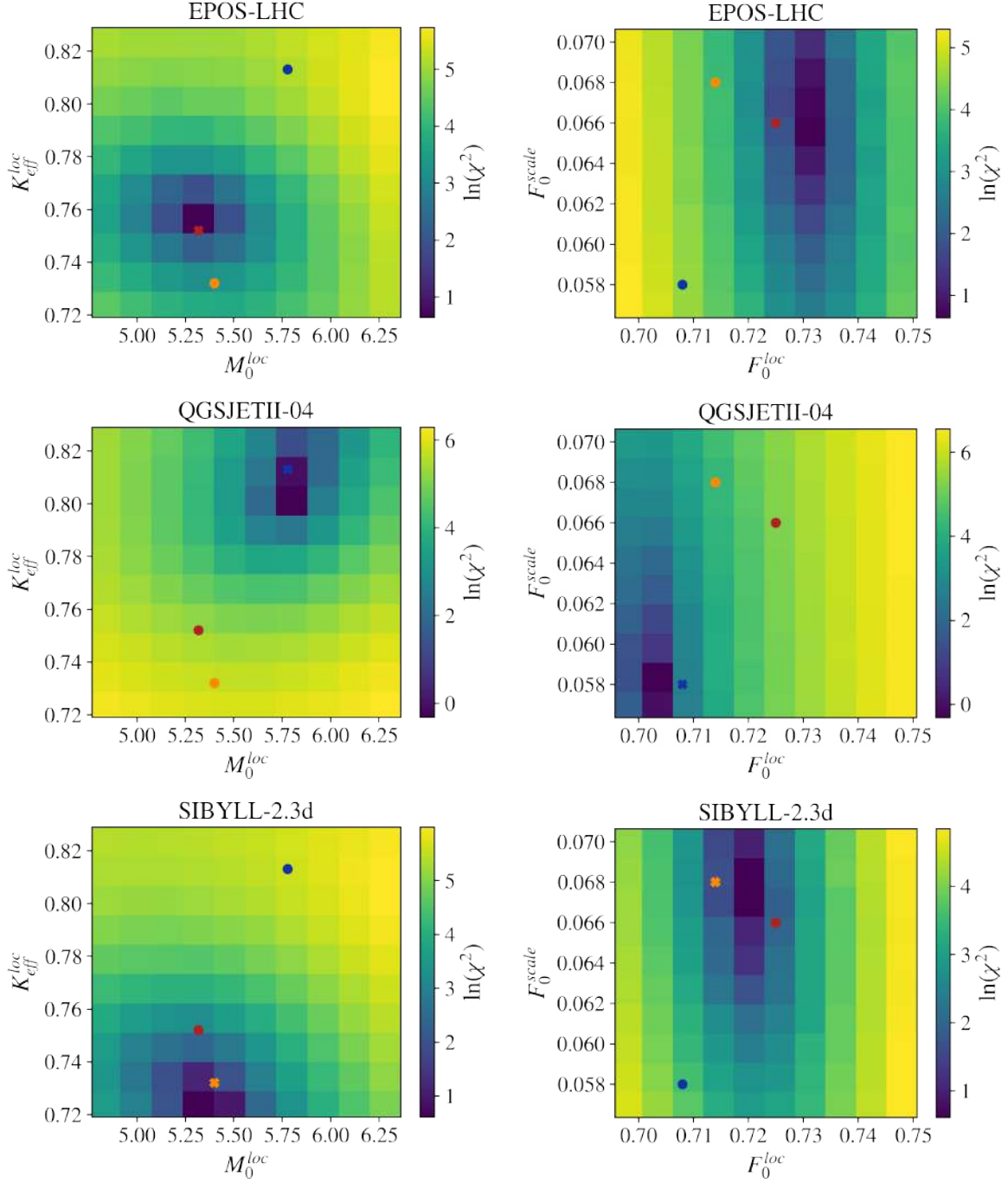


Figure 5.20: Logarithm of the 4-dimensional χ^2 -function calculated to get the unknown parameters M_0^{loc} , K_{eff}^{loc} , F_0^{loc} and F_0^{scale} . The real values for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d are shown as red, blue and yellow markers, respectively. In each figure the corresponding value that is being fitted is marked with a star, while the values for the other hadronic interaction models are put for comparison and represented with filled circles. In this χ^2 -analysis 4 observables are compared: $\langle X_{max} \rangle$, $\langle N_\mu \rangle$, $\sigma(X_{max})$ and $\sigma(N_\mu)$. Since we are interested in minimizing the χ^2 -function we show here 2-dimensional slices of the 4-dimensional space where the minimum occurs. We fix the values of F_0^{loc} and F_0^{scale} where the minimum is obtained and show the logarithm of the χ^2 as a function of M_0^{loc} and K_{eff}^{loc} (left). Equivalently, the χ^2 as a function of F_0^{loc} and F_0^{scale} is shown on the right. The minimum is achieved very close to the position of the real values. The logarithm of the χ^2 -function is used here only to make the position of the minimum more visible.

where $\bar{\theta}$ is the vector of the parameter values that are tested for M_0^{loc} , K_{eff}^{loc} , F_0^{loc} and F_0^{scale} . \hat{x} is the array containing the true values of the observables $\langle X_{max} \rangle$, $\langle N_\mu \rangle$, $\sigma(X_{max})$ and $\sigma(N_\mu)$. These are the values calculated using the bootstrap method. The vector $\bar{\mu}$ contains the values of the observables corresponding to the particular combination of parameters $\bar{\theta}$. This vector is obtained after evaluating our model on this combination of parameters $\bar{\theta}$. We include in the covariance matrix V the contributions due to statistical and systematic errors:

$$V = V_{stat} + V_{syst},$$

as suggested in [142, 143]. Since their measurements are correlated the covariance matrix $(V_{stat})_{ij} = cov[\hat{x}_i, \hat{x}_j]$ needs to be used here [144]. For the systematic errors we use $V_{syst} = \bar{\mathbf{s}}\bar{\mathbf{s}}^T$ where \mathbf{s} contains the difference between the real values of the observables and the predicted values using the parameters from the real distributions of M_0 , K_{eff} and F_0 . These matrices are also obtained using the bootstrap method. The minimum of the χ^2 -function in equation (5.1) defines the least-squares estimators $\hat{\theta}$ we are looking for.

The results for the 4-dimensional χ^2 -function are shown in figure 5.20 for all high-energy interaction models. Since it is not possible to deal with 4 dimensions visually, we only show there the slices corresponding to the regions where the minimum χ^2 -value is achieved. We fix the values of F_0^{loc} and F_0^{scale} at the values where the minimum is obtained and show the logarithm of the χ^2 as a function of M_0^{loc} and K_{eff}^{loc} on the left. Similarly, we show the χ^2 as a function of F_0^{loc} and F_0^{scale} on the right. The minimum is achieved very close to the position of the real values. The logarithm of the χ^2 -function is used here only to make the position of the minimum more visible.

Fitting the χ^2 -function as presented in figure 5.20 by a quadratic function in each

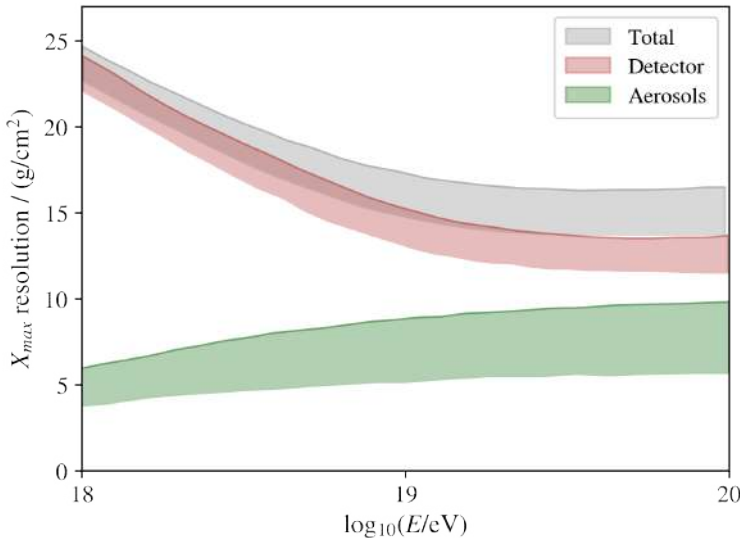


Figure 5.21: We need for our analysis the X_{max} value that is observed by the fluorescence telescopes of the Pierre Auger Observatory. Here the X_{max} resolution as a function of energy is shown (reproduced from [141]). Bands denote the estimated systematic uncertainties. Contributions due to the performance of the detector system (including the atmosphere itself) and due to the fluctuations of the night sky background as well as the time-variability of the aerosol content are shown together with the total contribution.

two-dimensional slice and searching for the minimum gives the predictions (stars in figure 5.22) for the four parameters M_0^{loc} , K_{eff}^{loc} , F_0^{loc} and F_0^{scale} for EPOS-LHC (red), QGSJETII-04 (blue) and SIBYLL-2.3d (yellow). The 1σ (full lines) and 2σ (dashed lines) contours in figure 5.22 correspond to $\chi_{min}^2 + 4.72$ and $\chi_{min}^2 + 9.7$, respectively. These are the values that need to be used in the case of 4 dimensions [145]. Comparing the real values (filled circles) with the predictions, one can see that M_0^{loc} , K_{eff}^{loc} and F_0^{scale} are remarkably well reproduced and contained within a 1σ error. F_0^{loc} for QGSJETII-04 is the only parameter that falls on the 2σ contour line. The prediction of this parameter for EPOS-LHC and SIBYLL-2.3d is better and close to the 1σ contour line. A feature of particular importance is that our model is capable of differentiating the 3 high-energy scenarios.

Before moving on to the dataset from the Pierre Auger Observatory in the final chapter, we can test the performance of the model including the effect of the resolution of the measurements. As can be seen in figure 5.21 (reproduced from [141]) for the showers in the small energy range we used for our simulations an absolute value of around 24 g/cm^2 is present. For the number of muons it is usual to assume a value of 20% which is considered quite conservative. We include this effect by smearing the X_{max} values according to a normal distribution centered at the true value and with a standard deviation of 24 g/cm^2 . The N_μ values are spread also according to a normal distribution centered at the true value but with a standard deviation equal to 0.2 times the real muon number. We do this with the original distributions simulated with EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d and with the ones that are created for the χ^2 -analysis. For each combination of possible values of the parameters M_0^{loc} , K_{eff}^{loc} , F_0^{loc} and F_0^{scale} under study, the values of X_{max} and N_μ that are obtained from our neural network are also subject to these modifications. When repeating the χ^2 -analysis the same way as was done for figure 5.22, but with these modified distributions, very similar and remarkably good results are obtained. These are summarized in figure 5.23. The predictions are slightly biased compared to the performance without including the effect of the resolution of the measurements, especially regarding the prediction of F_0^{loc} for EPOS-LHC. Nevertheless, the three high-energy interaction models are well distinguished and the rest of the predictions are within a 1σ or 2σ range.

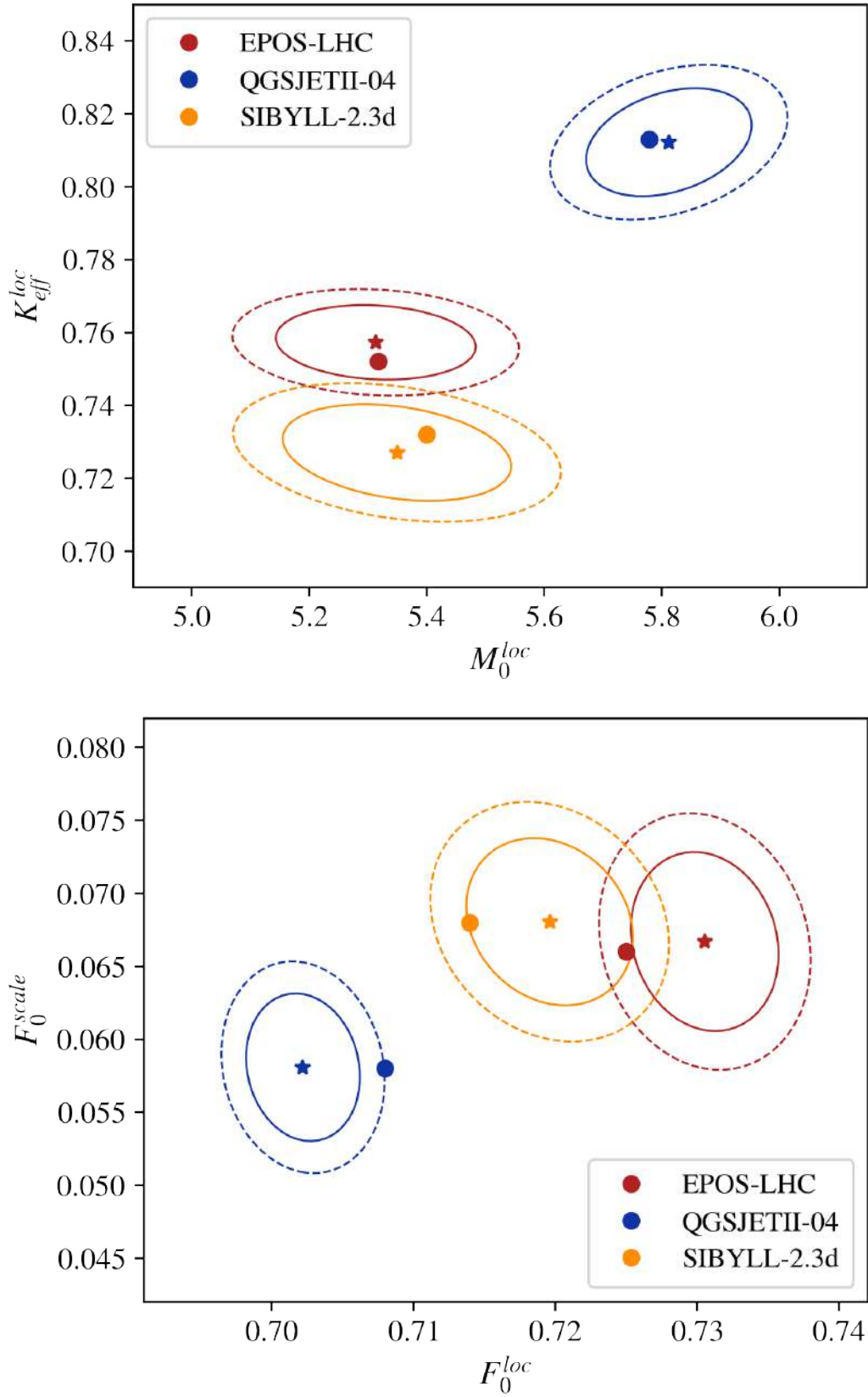


Figure 5.22: Minimum χ^2 fit to obtain the predictions (stars) of the four parameters M_0^{loc} , K_{eff}^{loc} , F_0^{loc} and F_0^{scale} for EPOS-LHC (red), QGSJETII-04 (blue) and SIBYLL-2.3d (yellow). The χ^2 -function as shown in figure 5.20 is used for the fit. The 1 σ (full lines) and 2 σ (dashed lines) contours correspond to $\chi_{min}^2 + 4.72$ and $\chi_{min}^2 + 9.7$, respectively. These are the values for the case of 4 dimensions. The filled circles are the real values.

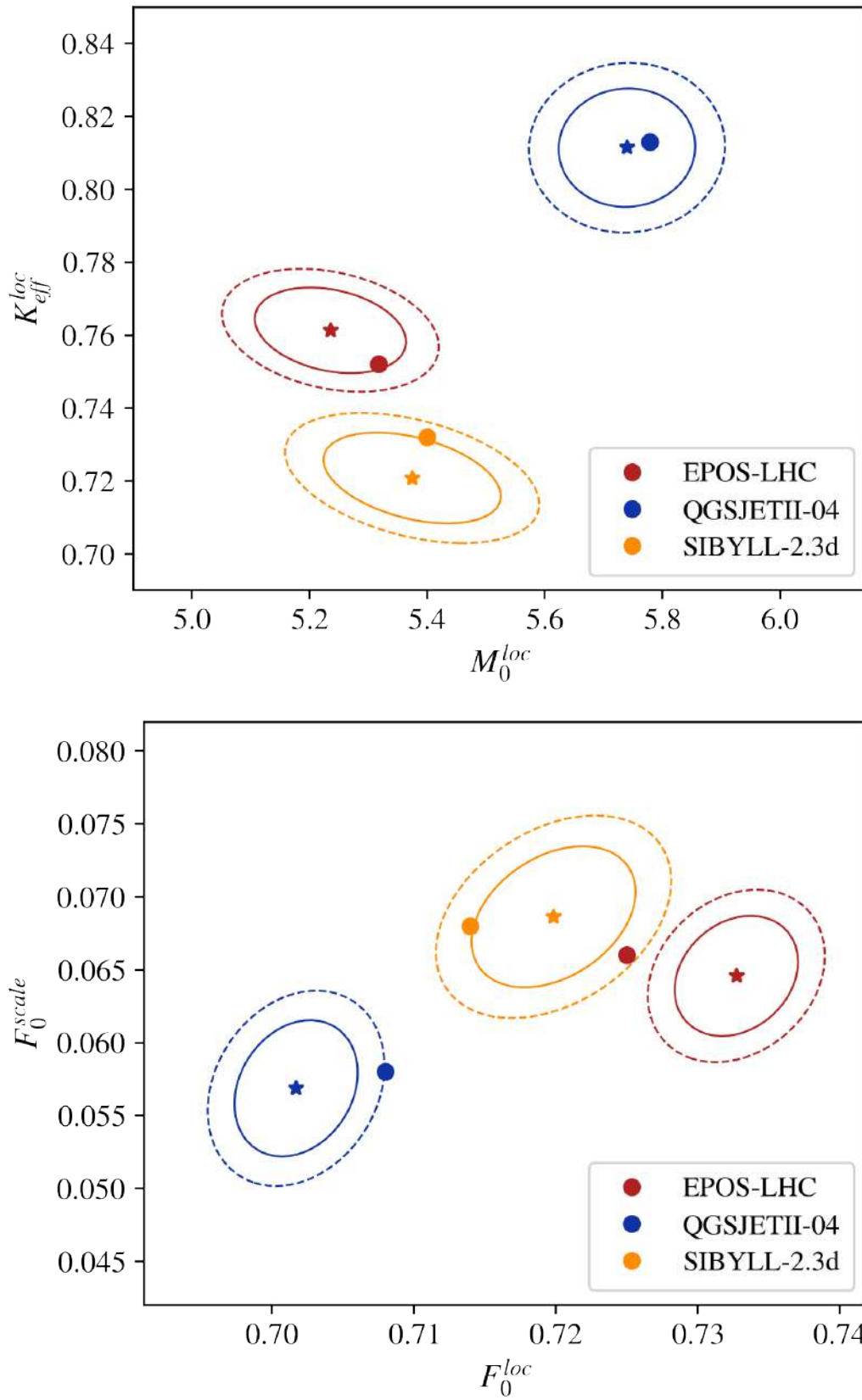


Figure 5.23: Minimum χ^2 fit to obtain the predictions (stars) of the four parameters M_0^{loc} , K_{eff}^{loc} , F_0^{loc} and F_0^{scale} for EPOS-LHC (red), QGSJETII-04 (blue) and SIBYLL-2.3d (yellow) as calculated for 5.22. The difference is that for this analysis we performed modifications on the $X_{max}-N_\mu$ distributions in order to include the effect of the resolution of the measurements.

Chapter 6

Application to Auger data

“El que para no avanza.”

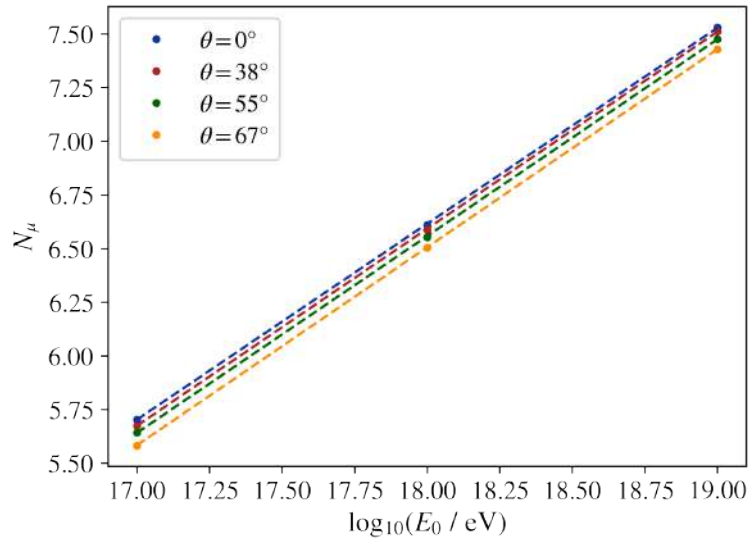
— Gustavo Eber Gonzalez

In this final chapter we adapt our formalism in order to apply it to a dataset from Auger. For this purpose, the mixed composition in the cosmic ray flux is taken into account [36]. However, having primaries other than protons introduces a difficulty, because our neural networks were trained on pure proton fluxes. A solution to this issue is proposed. Finally, we apply our formalism to a dataset from Auger which meets all the requirements we need. As a final result, we are able to infer the distributions of M_0 , F_0 and K_{eff} for this dataset. These reveal hints as to how the Muon Puzzle could be solved.

6.1. Dataset

The final goal of this work is to apply our formalism to a dataset from the Pierre Auger Observatory. For this purpose we need a dataset which contains unbiased values of the depth of maximum development X_{max} and the total number of muons at ground N_μ for vertical showers ($\theta < 60^\circ$). In [146] a model-independent calibration method based on the concept of *Universality* is applied to a dataset of hybrid events. On the one hand, having hybrid events ensures unbiased measurements of X_{max} . On the other hand, this calibration renders the number of muons we need, as we will explain later in this section. It is stated in [146] that an unbiased value can be obtained if one corrects for the fact that muon numbers for 10^{19} eV showers are systematically overestimated by around 14%. Furthermore, showers with zenith angle up to 60° are considered in this dataset. As a consequence, the conditions needed are met and we can apply our formalism to this dataset. The hybrid events used in it are from the period between 2005 and 2012. 7145 events are considered in total in the dataset. The estimated primary

Figure 6.1: Mean number of muons at ground N_μ as a function of primary energy for different zenith angles obtained from simulations performed using QGSJETII-03. Knowing the primary energy and angle of incidence of an event we can extract the corresponding number of muons from the relations presented here. The corresponding value from the function shown here times the normalization factor $N_\mu^n(E_0)$ given in the dataset gives the experimental number of muons we need.



energy E_0 , angle of incidence θ and value of X_{max} together with the estimator used to infer N_μ are collected in a file. This calibration method was not applied to events detected since then by the Pierre Auger Observatory. Nevertheless, the validity of the method remains solid.

The dataset contains X_{max} estimations from the SD and the FD, which are essentially equal because the SD estimations are constrained by those from the FD. The muon number is determined through the universality method in a model-independent way. The concept of air shower universality states that to a high degree of precision, extensive air showers can be characterized by a set of only three parameters: the primary energy E_0 , the depth of shower maximum X_{max} and the overall normalization of the muon component N_μ^n . Once these values are known the electromagnetic ground signal and the evolution of the muon signal are understood and can be used to parameterize the total ground signal:

$$S(E_0, \theta) = S_{em}(E_0, \theta, \langle X_{max} \rangle) + N_\mu^n(E_0) \cdot S_\mu(\theta, \langle X_{max} \rangle).$$

S_μ is a fixed mean muon reference signal for proton showers of 10^{19} eV simulated using QGSJETII-03. Consequently is $N_\mu^n(E_0)$ the relative muon normalization needed to obtain the signal for primary energy E_0 . Here E_0 and $N_\mu^n(E_0)$ are the only unknowns which are calculated in [146].

The factor $N_\mu^n(E_0)$ can be used to obtain an estimation of the number of muons at ground for a shower of primary energy E_0 . For that we need to know the mean number of muons at ground given a primary energy and zenith angle for showers generated by QGSJETII-03 (see figure 6.1). This mean number times $N_\mu^n(E_0)$, which is the value stored in the calibration dataset, gives the muon number at ground we are looking for. From the dataset we keep those events with primary energy between

$\log_{10}(E_0/\text{eV}) = 18.098$ and 18.198 , the range of energies covered by the simulations used to train the final model. With respect to the zenith angle, our simulations are done for $\theta = 38^\circ$. In order to not change the conditions significantly, we keep those showers with zenith angle between $\theta = 34^\circ$ and 40° . Furthermore, discarding events for which the calibration did not converge properly, we obtain a dataset with 263 events. The corresponding distributions of X_{max} and N_μ are displayed in figure 6.2. The distributions obtained for pure proton simulations using EPOS-LHC with the same energies and $\theta = 38^\circ$ are added for comparison.

6.2. Mixed composition

Now that we are ready to apply this formalism to a dataset from Auger, apart from including the resolution of the measurements as we have already done in section 5.4, we need to take into account the mixed composition present in the cosmic ray flux (see section 1.2.4). We chose the small energy interval $\log_{10}(E_0/\text{eV}) = 18.098 - 18.198$ because contamination by particles other than proton is minimized, as can be seen in figure 1.6 [36, 147]. We chose a small interval here because the fractions of the different elements quickly change when varying the energy and for our study we need a more or less fixed combination of fractions. In order to involve mixed composition into our analysis we repeat the calculations done in 5.4. This time, we create X_{max} - N_μ distributions from simulations respecting the fractions of primaries valid for our energy range. In addition, we apply the smearing as done in section 5.4. The resulting distributions can be appreciated in figure 6.3. The distribution for QGSJETII-04 has only a small fraction of helium with respect to proton. For EPOS-LHC the ‘‘contamination’’ through nitrogen is more pronounced, while the distribution for SIBYLL-2.3d is the most affected by helium as well as nitrogen.

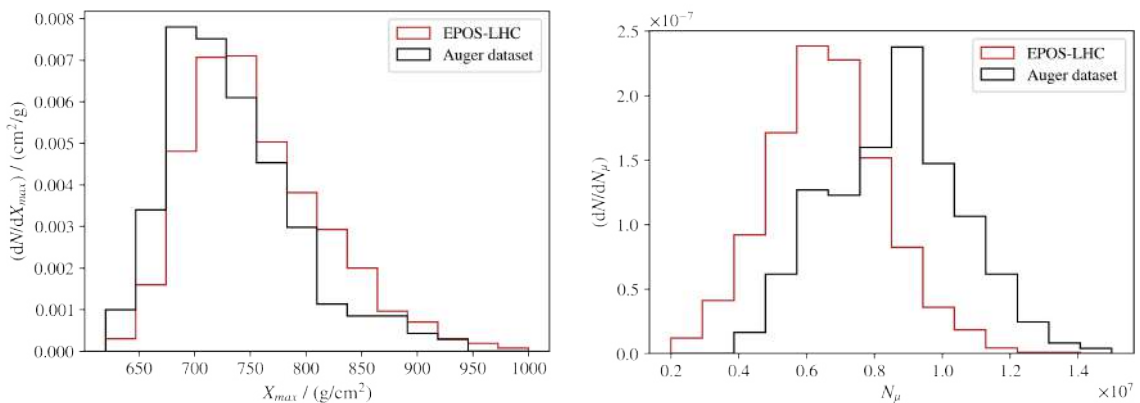


Figure 6.2: X_{max} and N_μ distributions from the Auger dataset. The corresponding distributions for EPOS-LHC simulations of proton primaries with $\theta = 38^\circ$ and $\log_{10}(E_0/\text{eV})$ between 18.098 and 18.198 are added for comparison.

Since the neural networks developed in chapter 5 were trained with simulations of showers generated by primary protons and the reasoning that led to the election of features used there was also focused on primary protons (see section 4.2), we expect the presence of other primaries to worsen the predictions obtained at the end of section 5.4. Our neural network has not “learned” from instances other than protons. Furthermore, we don’t expect to be able to build a unified model independent of primary mass with this set of features, besides from already being independent from the hadronic interaction model used. Showers generated by heavy primaries behave very differently from proton initiated ones in the first interaction. The inelasticity for example, as used in the mathematical expressions in section 4.2, does not make sense for showers generated by heavy primaries. In those showers a considerable fraction of the primary energy is taken by individual nucleons that break off from the primary, while other nucleons remain together in fragments until the second interaction occurs. In the case of helium primaries, this effect is the least pronounced.

Consequently, we need to apply a cut on the distributions in order to get rid of

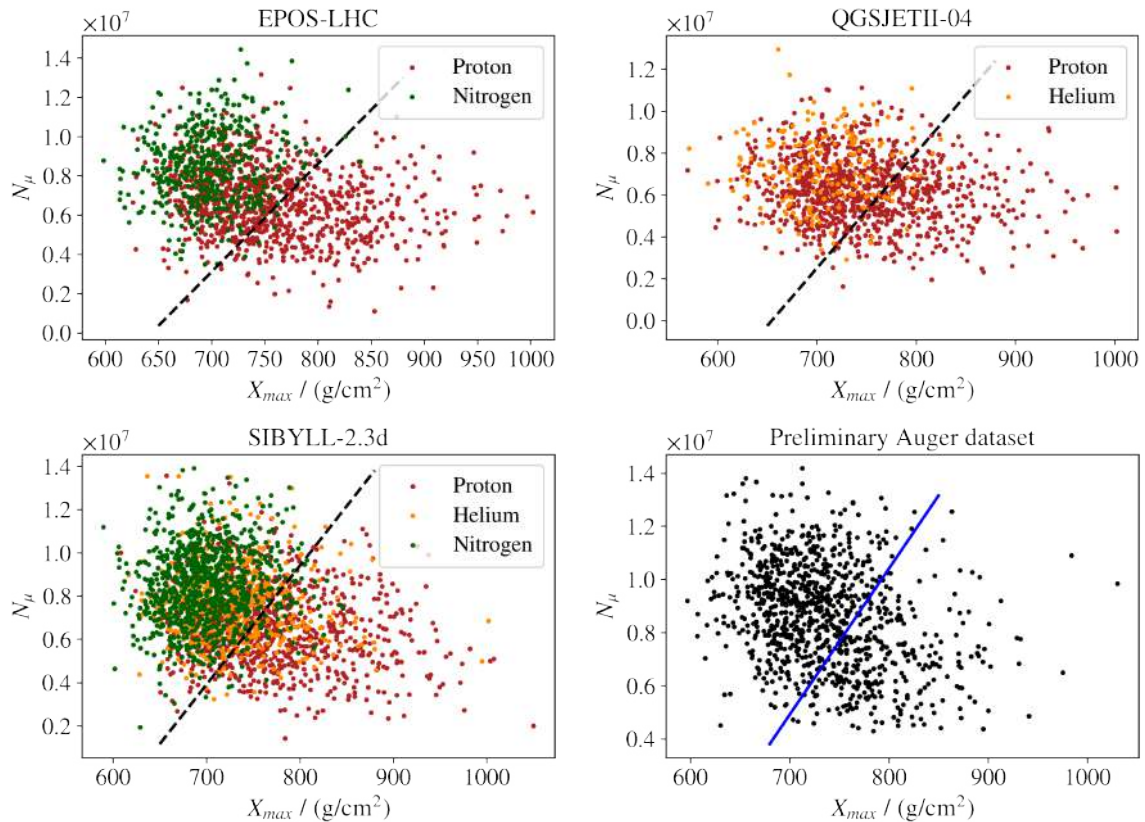


Figure 6.3: X_{max} - N_μ distribution for EPOS-LHC, QGSJETII-04, SIBYLL-2.3d and the Auger dataset after applying artificial smear in order to mimic the measurement resolution as described in section 5.4. Also the effect of mixed composition is taken into account. For each high-energy interaction model the fractions of proton (red dots), helium (yellow), nitrogen (green) and iron from [36] are used. The tilted line sets the limit for the cut we use in the rest of the study. The objective is to keep a sample of events with a high fraction of protons while keeping as many events as possible.

the most “contaminated” areas in the X_{max} - N_μ plane. We chose to do this by keeping a certain percentage of the “tail” of the X_{max} - N_μ distribution. We use an inclined line to separate a predefined percentage of the events we discard, from those we keep (tilted lines in figure 6.3). Considering only the events below the cut enhances the contribution of protons in the sample. The lower the line representing the cut, the higher the fraction of protons in the sample. However, this also reduces the number of events available for the analysis. Since no hadronic interaction model is preferred,

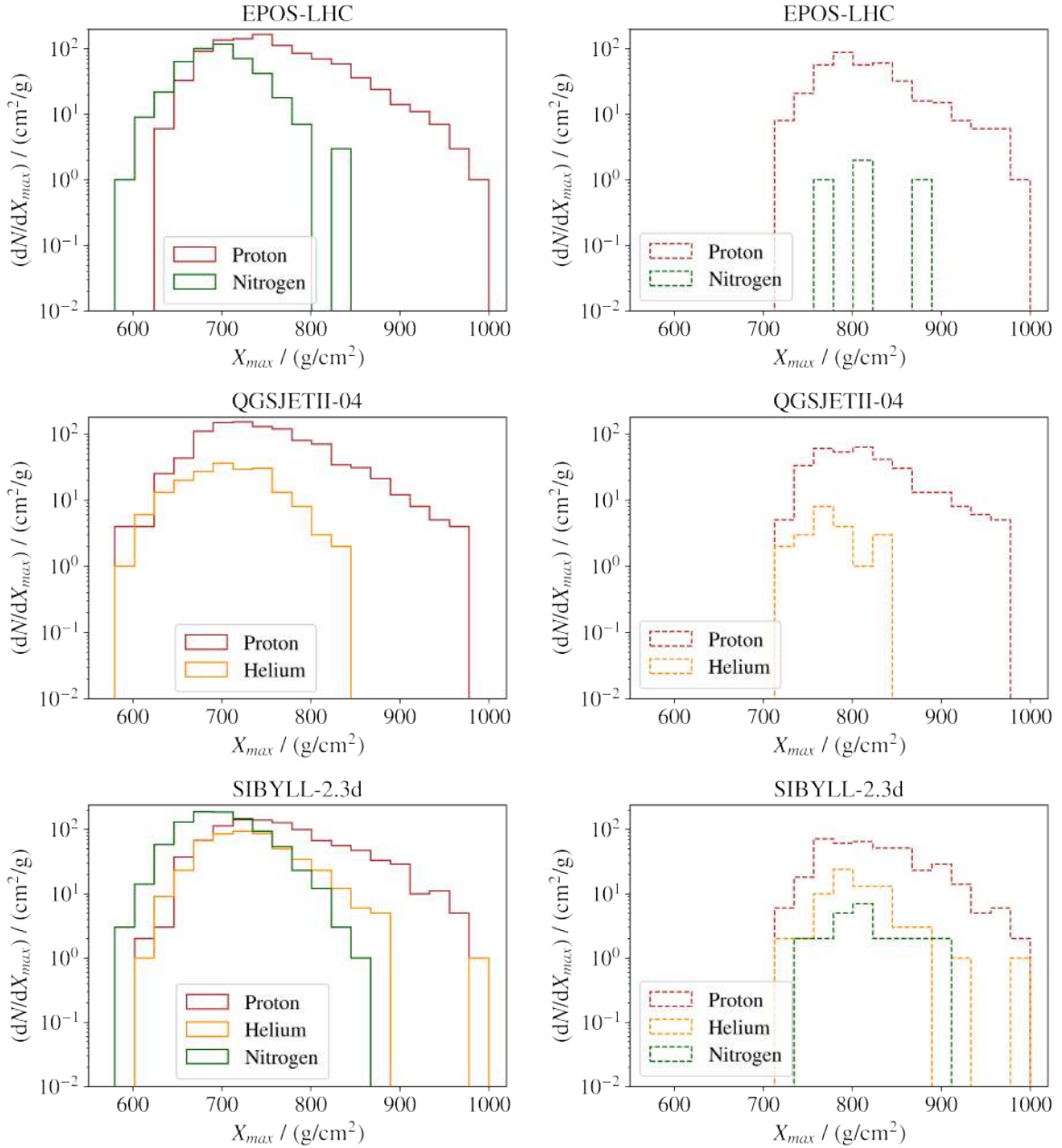


Figure 6.4: X_{max} distribution of different primaries present in the cosmic ray flux using the fractions calculated for different high-energy interaction models [36] before (left) and after (right) performing the cut shown in figure 6.3. The axes are fixed at equal values in this comparison in order to see which portion of events is selected out when performing the cut.

a cut with respect to a common percentage needs to be applied to all distributions in an equal way. We perform this cut on the distributions of interest, from which we wish to infer the values of the parameters (the distributions from all three high-energy interaction models and the Auger dataset), and on the distributions generated for each combination of parameters θ with which they will be compared in the χ^2 -analysis.

For the definition of the cut we tested the values of 40 %, 30 % and 20 %. These values ensured having a high enough fraction of protons. Using a value of 30 % we obtained the best results. For 40 % and 20 % the predictions of the parameters for the

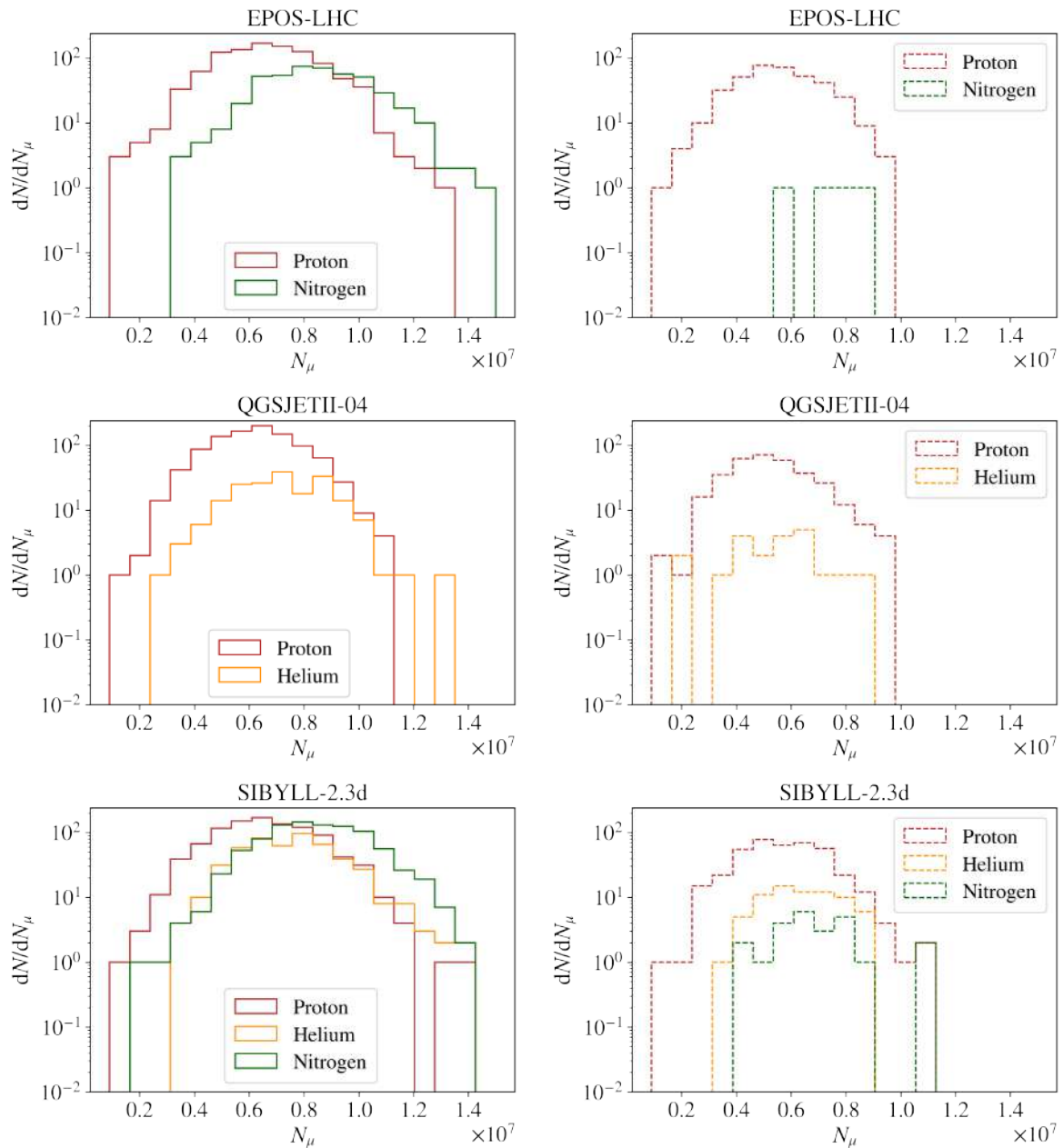


Figure 6.5: N_μ distribution of different primaries present in the cosmic ray flux using the fractions calculated for different high-energy interaction models [36] before (left) and after (right) performing the cut shown in figure 6.3. The axes are fixed at equal values in this comparison in order to see which portion of events is selected out when performing the cut.

simulated distributions had a stronger bias than for the cut corresponding to 30%. In figures 6.4 and 6.5 we show on the left the distributions of X_{max} and N_μ , respectively. Each color corresponds to a different primary present for the high-energy interaction model considered (only protons, helium, nitrogen and iron are considered). The fraction of each particle type is from [36]. On the right we present the corresponding distributions after performing the cut. We leave all axes fixed so that the distributions can be compared. After performing the cut a considerable amount of “undesired” primaries is left out for EPOS-LHC and QGSJETII-04. For SIBYLL-2.3d the presence of primaries other than proton is the highest.

As already mentioned, we don’t expect to be able to develop a model that is independent of the primary mass, using just the three features kept until now. Nonetheless, we retrain our neural networks with the same architecture as presented in section 5.2 on a joint ensemble of the distributions described in this chapter: distributions obtained with the three high-energy interaction models including the effect of measurement resolution and the effect of mixed composition and subjected to the cut corresponding to 30%. We thus allow for the neural network to understand in addition, at least partially,

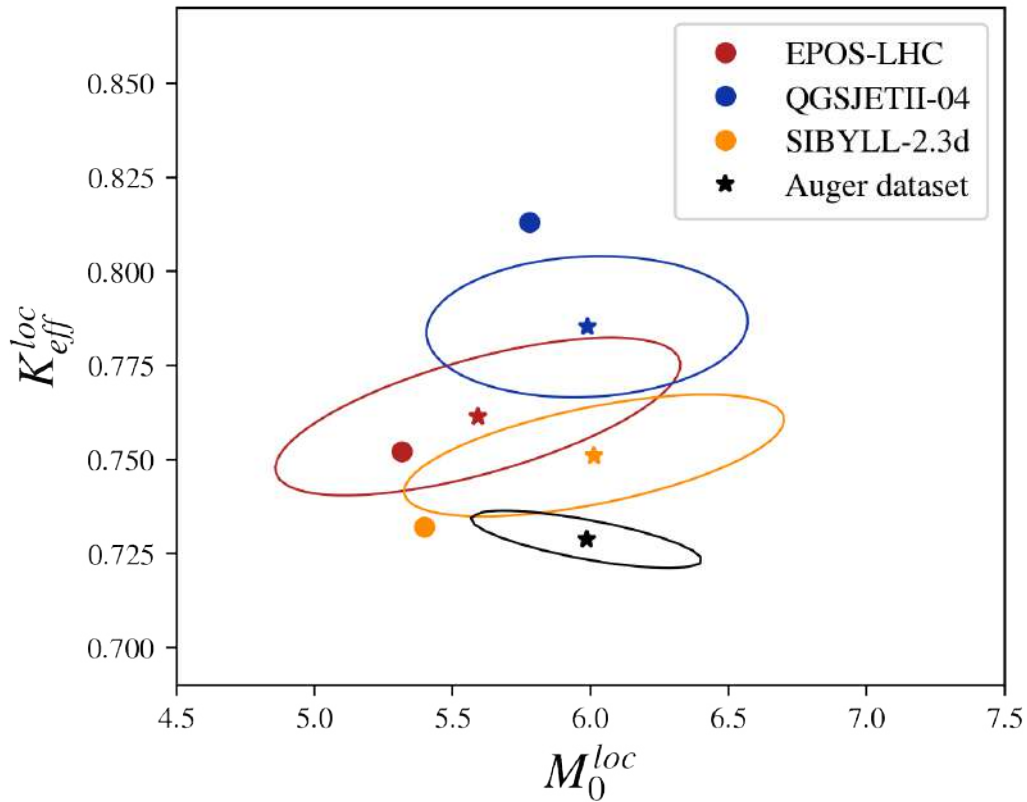


Figure 6.6: Minimum χ^2 fit to obtain the calculated values (stars) of the parameters M_0^{loc} and K_{eff}^{loc} for EPOS-LHC (red), QGSJETII-04 (blue) and SIBYLL-2.3d (yellow) as calculated for 5.22. In this analysis we include the effect of having mixed composition in the cosmic ray flux. In this setting it is possible to include the calculated values corresponding to the Auger dataset (black star). The ellipses correspond to a 1σ region.

this new setting where showers generated by other primaries are present. Performing a χ^2 -analysis as in section 5.4 using these updated neural networks and applying it to the tails of the distributions as shown in figure 6.3, we obtain the results presented in figures 6.6 and 6.7.

Despite the difficulty introduced by the incorporation of the effect of mixed composition, the predicted values are very close to the real ones. For EPOS-LHC all predictions are within the 1σ region. For the predictions of M_0^{loc} and F_0^{scale} for QSGJETII-04 and SIBYLL-2.3d this is also the case. The rest of the predictions are within the 2σ region (not shown in figures 6.6 and 6.7). The 1σ region for the prediction corresponding to the Auger dataset is smaller than for the different hadronic interaction scenarios because for the dataset we do not know the true values of the parameters. Consequently, we do not have the matrix V_{syst} at our disposal. Taking into account the size of the 1σ region for the simulations one can get a notion of which size this region would have if the systematic errors were included.

The inferred values of the unknown parameters are $M_0^{loc} = 5.99$, $K_{eff}^{loc} = 0.73$, $F_0^{loc} = 0.77$ and $F_0^{scale} = 0.058$. The distributions of M_0 , F_0 , K_{eff} and N_{ch}^{FI} obtained

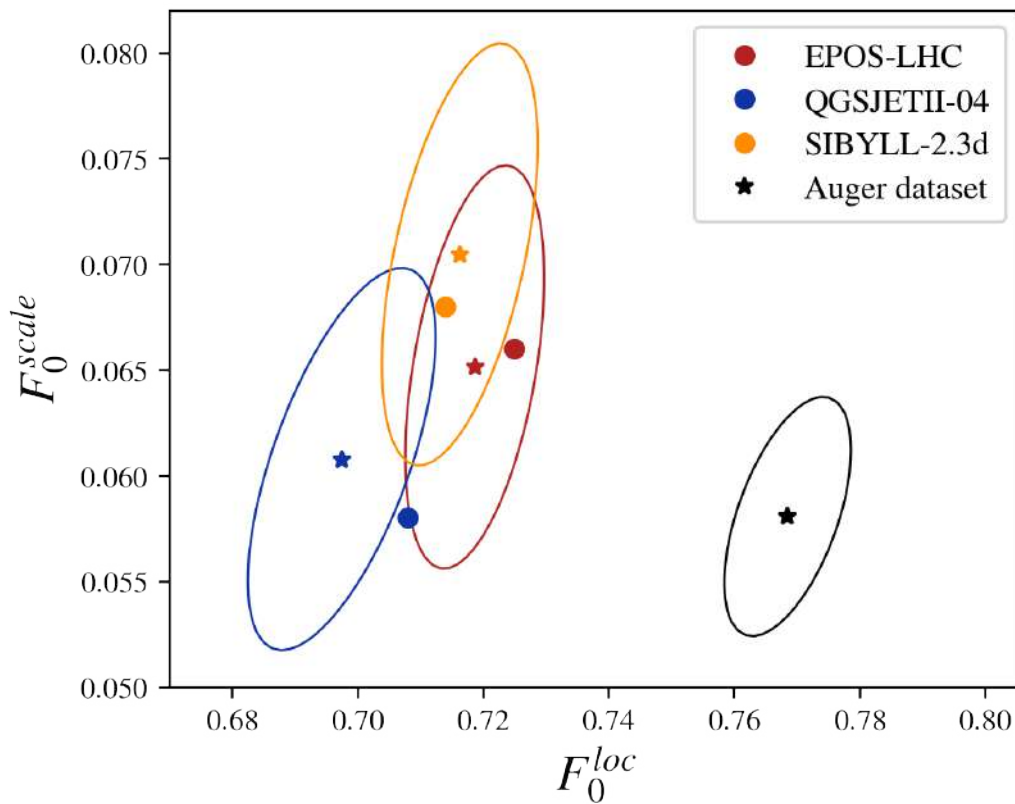


Figure 6.7: Minimum χ^2 fit to obtain the calculated values (stars) of the parameters F_0^{loc} and F_0^{scale} for EPOS-LHC (red), QGSJETII-04 (blue) and SIBYLL-2.3d (yellow) as calculated for 5.22. In this analysis we include the effect of having mixed composition in the cosmic ray flux. In this setting it is possible to include the calculated values corresponding to the Auger dataset (black star). The ellipses correspond to a 1σ region.

for these values are presented in figure 6.8 together with those corresponding to EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d. The distributions for N_{ch}^{FI} are included here, because the multiplicity is typically presented this way in the literature.

The hadronic multiplicity of the first interaction M_0 tends to be higher for the dataset than for any of the models used in simulations. This makes sense because having a higher hadronic multiplicity at the first interaction implies having more subshowers of less energy. Thus, the shower develops in less generations, keeping more energy in the hadronic channel which eventually means a higher number of muons. This also occurs when the hadronic fraction of the first interaction F_0 is higher, as happens for the Auger dataset. Having a higher value for F_0^{loc} than any model shifts its distribution to higher values. Having a small value of F_0^{scale} concentrates the peak around these higher values. As for the resulting K_{eff} distribution it is quite close to that corresponding to SIBYLL-2.3d and also to the EPOS-LHC distribution.

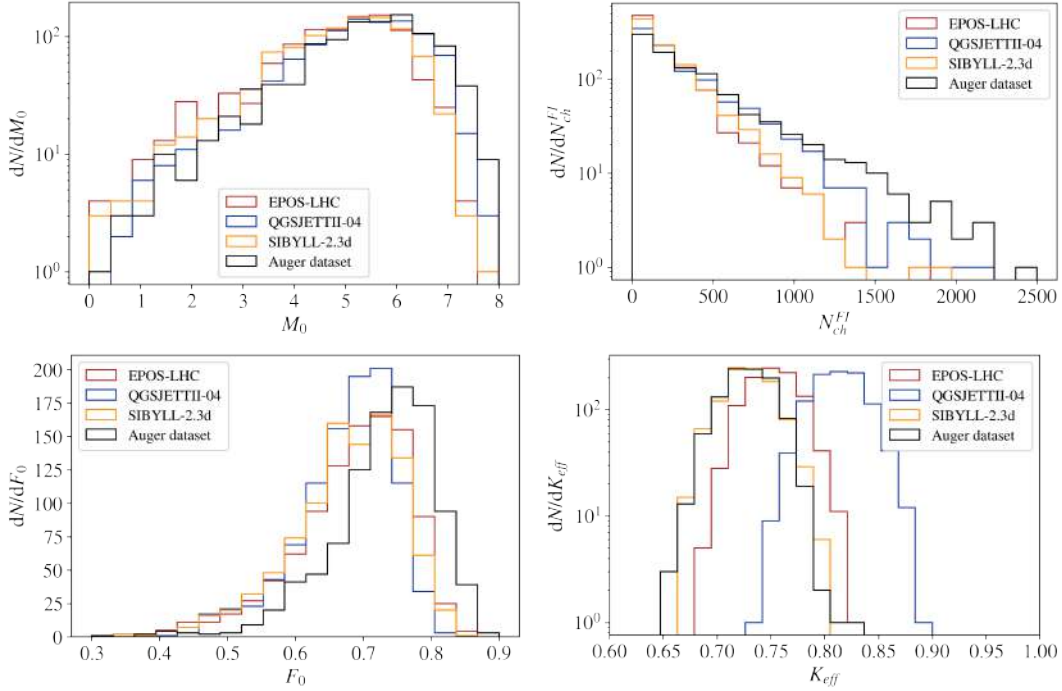


Figure 6.8: Predicted distributions of the unknown features M_0 , N_{ch}^{FI} , F_0 and K_{eff} for the Auger dataset obtained using the predicted basic parameters (see figures 6.6 and 6.7). The hadronic multiplicity M_0 tends to be higher than for any of the models used in simulations. This is in accordance with having more muons in the dataset compared to simulations. This also can be associated with a higher hadronic fraction F_0 , as happens for the Auger dataset. Having a higher value for F_0^{loc} than any model shifts its distribution to higher values. Having a small value of F_0^{scale} concentrates the peak around these higher values. As for the resulting K_{eff} distribution it is quite close to that corresponding to SIBYLL-2.3d and also to the EPOS-LHC distribution.

These final results reveal which conditions need to be met in order to solve the Muon Puzzle. The energy fraction of hadronically interacting particles of the first interaction needs to be higher than for current high-energy interaction models. Furthermore, its distribution needs to be narrower. An increase in the hadronic energy fraction could be

obtained with the *core-corona model* [148]. The basic idea is that a certain fraction of the volume of an interaction behaves like a quark-gluon plasma and decays according to statistical hadronization (core), while the other part produces particles via string fragmentation (corona). The existence of such a core would increase the value of F_0 because statistical hadronization produces more heavy particles and less pions compared to string fragmentation [149, 150].

Chapter 7

Conclusions

Cosmic rays of the highest energies are excellent probes to study physical properties at energies beyond those achievable in man-made accelerators. Many studies are being carried out in which characteristics of X_{max} or N_μ alone are used to extract information about extensive air showers. The purpose of the present work is the study of the joint distribution of X_{max} and N_μ . The fact that these two observables are correlated means that analyzing their distributions simultaneously should render new information not accessible when considering them separately.

During the first part of the work we studied the anticorrelation by developing a model that describes X_{max} and N_μ as functions of certain parameters. The main result is a model which is based on a combination of the semi-empirical model developed by Heitler and Matthews and the expression of the critical energy as calculated by Kampert and Unger. Considerable improvement is achieved when modeling the first interaction separately from the rest of the shower. This separation is not only reflected in the mathematical equations but also in the set of parameters used to predict X_{max} and N_μ . These are the multiplicity of hadronically interacting particles of the first interaction M_0 , a parameter K_0 mainly related to the inelasticity of the first interaction and a parameter F_0 mainly related to the fraction of energy that is taken by the hadronically interacting particles of the first interaction together with three effective versions M_{eff} , K_{eff} and F_{eff} of these parameters representative for the rest of the shower.

The ΔX - N_μ distribution is represented very well with this model, not only qualitatively but also quantitatively, which is remarkable for a relatively simple discrete model. In particular, the anticorrelation is very visible in the predicted distribution, as well. In order to make sure that this behaviour is no coincidence, the distributions of all used parameters were compared between the original anticorrelation and the predicted one. The three parameters related to the first interaction, which are the most important ones, show good coincidence in their behaviour. The three effective para-

eters related to the rest of the shower have the role of introducing a scale around which these values are to be expected. These values from simulations are indeed very concentrated in small regions and show little variability. The hadronic multiplicity M_0 and the inelasticity K_0 of the first interaction are correlated with ΔX and describe the anticorrelation moving horizontally. The hadronic energy fraction F_0 of the first interaction, on the other hand, is more related to the number of muons and describes the anticorrelation moving almost vertically. None of these parameters alone describes the anticorrelation. Instead, a combination of these parameters is needed to describe the inclined axis that the anticorrelation follows.

In the second part of the work we use the fact that we know which parameters are important to describe the X_{max} - N_μ anticorrelation in order to develop a model using neural networks. The versatility of neural networks is expected to allow for a model that captures more details than through the equations from the analytical model. The principal model is trained on a reduced set of features. From the parameter set introduced previously the multiplicity of hadronically interacting particles of the first interaction M_0 , the parameter F_0 mainly related to the fraction of energy that is taken by the hadronically interacting particles of the first interaction and the effective inelasticity K_{eff} of the rest of the shower are sufficient to use as features in neural networks that output X_{max} and N_μ , separately. K_{eff} turned out to be more important than K_0 . A joint mixed ensemble of simulations performed with CONEX using the high-energy interaction models EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d was used for training.

A particular characteristic that was attained for this main model is that of universality: predictions of X_{max} and N_μ are calculated by this network essentially equally well for all three high-energy interaction models considered, which is a very valuable quality. This means that the differences between the hadronic interaction models are captured in the parameters that were kept, while the behaviour that is common to all three models is included in the final neural network. Furthermore, the error in the predictions of X_{max} and N_μ are for all high-energy interaction scenarios below 36.6 g/cm² and 7.7%, respectively.

The performance of this model was tested emulating the process that is applied at a final stage to a dataset from Auger. The distributions of the features used for training were described through suitable distributions. Four parameters M_0^{loc} , K_{eff}^{loc} , F_0^{loc} and F_0^{scale} are sufficient to describe the unknown distributions. A χ^2 -analysis was carried out where $\langle X_{max} \rangle$, $\langle N_\mu \rangle$, $\sigma(X_{max})$ and $\sigma(N_\mu)$ from the true distributions from simulations were compared to those generated by our model for a grid of possible values of the parameters M_0^{loc} , K_{eff}^{loc} , F_0^{loc} and F_0^{scale} . Remarkably good results were obtained for the predicted values of the parameters. The predictions of M_0^{loc} , K_{eff}^{loc} and F_0^{scale} are well within a 1σ error from the corresponding true values. In the case of the position

of the peak F_0^{loc} the prediction is between 1σ and 2σ away from the true value. In particular, the model is clearly able to distinguish between the scenarios from different theories.

Finally, this process was applied to a hybrid dataset from Auger in which the number of muons at ground is provided using the concept of universality. Since in a realistic scenario the cosmic ray flux presents mixed composition, this needs to be taken into account in the process. This introduces a difficulty because our model is, in principle, not designed to deal with primaries other than protons. However, this issue was solved performing a cut on the distributions in order to obtain samples with high proton fraction. Furthermore, the model was retrained on the sets of simulations that respect the estimated fractions of proton, helium, nitrogen and iron. This way, it was possible to obtain predictions for the parameters M_0^{loc} , K_{eff}^{loc} , F_0^{loc} and F_0^{scale} for the Auger dataset. These are sufficient to deduce the distributions of M_0 , F_0 and K_{eff} .

It would be interesting to repeat this analysis considering other small ranges of primary energy which are associated with other combinations of primary fractions. One would need to have simulations in every energy interval of interest, train different neural networks with these sets of simulations because the feature distributions change with energy. Finally, the χ^2 -analysis would need to be performed considering the resolution of the measurements and the combination of fractions of different primaries for the corresponding primary energy. Furthermore, it will be an excellent opportunity to apply our formalism to data obtained with AugerPrime. This upgrade of the Pierre Auger Observatory will provide a complementary measurement of the showers allowing the reconstruction of muons and electromagnetic particles.

In this work, the distributions of the hadronic energy fraction of the first interaction, of the hadronic multiplicity of the first interaction and of the effective inelasticity were inferred for a preliminary dataset from the Pierre Auger Observatory. The first two distributions are highly interesting properties related to high-energy hadronic interactions which cannot be studied at accelerators. It is the first time that such distributions are observed, that show hadronic properties at energies 10 times higher than the highest achievable at the LHC (in the laboratory frame). The differences with the hadronic interaction models give hints as to how the Muon Puzzle could be solved. The obtained distribution of the hadronic energy fraction of the first interaction is significantly different from the ones obtained using current hadronic interaction models. F_0 is generally higher for the dataset compared to simulations. This result suggests that mechanisms that increase the hadronic fraction F_0 should be implemented in the high-energy interaction models used in simulations.

Bibliography

- [1] Montanus, J. An extended heitler–matthews model for the full hadronic cascade in cosmic air showers. *Astroparticle Physics*, **59**, 4–11, 2014. [vii](#), [28](#), [30](#), [37](#), [38](#), [42](#), [43](#)
- [2] Kampert, K.-H., Unger, M. Measurements of the cosmic ray composition with air shower experiments. *Astroparticle Physics*, **35** (10), 660–678, May 2012. URL <http://dx.doi.org/10.1016/j.astropartphys.2012.02.004>. [vii](#), [28](#), [38](#), [39](#), [42](#), [44](#), [45](#)
- [3] Hess, V. On the observations of the penetrating radiation during seven balloon flights, 2018. [1](#)
- [4] Anderson, C. D. The positive electron. *Phys. Rev.*, **43**, 491–494, Mar 1933. URL <https://link.aps.org/doi/10.1103/PhysRev.43.491>. [2](#)
- [5] Bazilevskaya, G. Skobeltsyn and the early years of cosmic particle physics in the soviet union. *Astroparticle Physics*, **53**, 01 2013. [2](#)
- [6] Bonolis, L. Walther bothe and bruno rossi: The birth and development of coincidence methods in cosmic-ray physics. *American Journal of Physics*, **79** (11), 1133–1150, Nov 2011. URL <http://dx.doi.org/10.1119/1.3619808>. [2](#)
- [7] Dirac, P. A. M. The quantum theory of the electron. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, **117** (778), 610–624, 1928. [2](#)
- [8] Yukawa, H. On the interaction of elementary particles. i. *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series*, **17**, 48–57, 1935. [2](#)
- [9] Anderson, C. D., Neddermeyer, S. H. Cloud chamber observations of cosmic rays at 4300 meters elevation and near sea-level. *Phys. Rev.*, **50**, 263–271, Aug 1936. URL <https://link.aps.org/doi/10.1103/PhysRev.50.263>. [2](#)
- [10] Ellis, J. The discovery of the gluon. *International Journal of Modern Physics A*, **29** (31), 1430072, Dec 2014. URL <http://dx.doi.org/10.1142/S0217751X14300725>. [3](#)

- [11] Weise, W. Yukawa's pion, low-energy qcd and nuclear chiral dynamics. *Progress of Theoretical Physics Supplement*, **170**, 161–184, 2007. [3](#)
- [12] Chao, C. Y. Mixed cosmic-ray showers at sea level. *Phys. Rev.*, **75**, 581–590, Feb 1949. URL <https://link.aps.org/doi/10.1103/PhysRev.75.581>. [3](#)
- [13] Fretter, W. B. Penetrating showers. *Phys. Rev.*, **76**, 511–517, Aug 1949. URL <https://link.aps.org/doi/10.1103/PhysRev.76.511>. [3](#)
- [14] Lattes, C. M. G., Muirhead, H., Occhialini, G. P. S., Powell, C. F. Processes involving charged mesons. *Nature*, **159**, 694–697, 1947. [3](#)
- [15] Rochester, G., Butler, C. Evidence for the existence of new unstable elementary particles. *Nature*, **160** (4077), 855–857, 1947. [3](#)
- [16] Seriff, A. J., Leighton, R. B., Hsiao, C., Cowan, E. W., Anderson, C. D. Cloud-chamber observations of the new unstable cosmic-ray particles. *Phys. Rev.*, **78**, 290–291, May 1950. URL <https://link.aps.org/doi/10.1103/PhysRev.78.290>. [3](#)
- [17] Rossi, B. Misura sulla distribuzione angolare di intensità della radiazione penetrante all'asmara. *Supplemento a la Ricerca Scientifica*, **1**, 579, 1934. [4](#)
- [18] Schmeiser, K., Bothe, W. Die harten ultrastrahlschauer. *Annalen der Physik*, **424** (1-2), 161–177, 1938. [4](#)
- [19] Kolhörster, W., Matthes, I., Weber, E. Gekoppelte höhenstrahlen. *Naturwissenschaften*, **26** (35), 576–576, 1938. [4](#)
- [20] Auger, P., Ehrenfest, P., Maze, R., Daudin, J., Fréon, R. A. Extensive cosmic-ray showers. *Rev. Mod. Phys.*, **11**, 288–291, Jul 1939. URL <https://link.aps.org/doi/10.1103/RevModPhys.11.288>. [4](#)
- [21] Bethe, H., Heitler, W. On the stopping of fast particles and on the creation of positive electrons. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, **146** (856), 83–112, 1934. URL <http://www.jstor.org/stable/2935479>. [5](#)
- [22] Carlson, J. F., Oppenheimer, J. R. On multiplicative showers. *Phys. Rev.*, **51**, 220–231, Feb 1937. URL <https://link.aps.org/doi/10.1103/PhysRev.51.220>. [5](#)
- [23] Kamata, K., Nishimura, J. The Lateral and the Angular Structure Functions of Electron Showers. *Progress of Theoretical Physics Supplement*, **6**, 93–155, 02 1958. URL <https://doi.org/10.1143/PTPS.6.93>. [5](#), [15](#)

- [24] Misaki, A. Mean Square Angular and Lateral Spreads of Electrons and Photons in a Cascade Shower. *Progress of Theoretical Physics Supplement*, **32**, 82–103, 04 1964. URL <https://doi.org/10.1143/PTPS.32.82>. 5
- [25] Kampert, K. H., Collaboration, K.-G. Cosmic rays in the knee-region - recent results from kascade -, 2004. 6
- [26] Apel, W. D., Arteaga-Velázquez, J. C., Bekk, K., Bertaina, M., Blümer, J., Bozdog, H., *et al.* Kneelike structure in the spectrum of the heavy component of cosmic rays observed with kascade-grande. *Physical Review Letters*, **107** (17), Oct 2011. URL <http://dx.doi.org/10.1103/PhysRevLett.107.171104>. 7
- [27] Hillas, A. M. The origin of ultra-high-energy cosmic rays. *Annual review of astronomy and astrophysics*, **22** (1), 1984. 7, 8, 15
- [28] Collaboration, P. A., *et al.* Bounds on the density of sources of ultra-high energy cosmic rays from the pierre auger observatory. *Journal of Cosmology and Astroparticle Physics*, **2013** (05), 2013. 7, 8
- [29] Alves Batista, R., Biteau, J., Bustamante, M., Dolag, K., Engel, R., Fang, K., *et al.* Open questions in cosmic-ray research at ultrahigh energies. *Frontiers in Astronomy and Space Sciences*, **6**, Jun 2019. URL <http://dx.doi.org/10.3389/fspas.2019.00023>. 7, 8
- [30] Apel, W., Arteaga-Velazquez, J., Bekk, K., Bertaina, M., Blumer, J., Bozdog, H., *et al.* Ankle-like feature in the energy spectrum of light elements of cosmic rays observed with kascade-grande. *Physical Review D*, **87** (8), 2013. 8
- [31] Abraham, J., Abreu, P., Aglietta, M., Aguirre, C., Allard, D., Allekotte, I., *et al.* Observation of the suppression of the flux of cosmic rays above 4×10^{19} ev. *Physical review letters*, **101** (6), 2008. 8
- [32] Fang, K., Kotera, K., Olinto, A. V. Ultrahigh energy cosmic ray nuclei from extragalactic pulsars and the effect of their galactic counterparts. *Journal of Cosmology and Astroparticle Physics*, **2013** (03), 2013. 8
- [33] Aab, A., Abreu, P., Aglietta, M., Al Samarai, I., Albuquerque, I., Allekotte, I., *et al.* Observation of a large-scale anisotropy in the arrival directions of cosmic rays above 8×10^{18} ev. *Science*, **357** (6357), 2017. 9
- [34] Aab, A., Abreu, P., Aglietta, M., Samarai, I. A., Albuquerque, I., Allekotte, I., *et al.* Combined fit of spectrum and composition data as measured by the pierre auger observatory. *Journal of Cosmology and Astroparticle Physics*, **2017** (04), Apr 2017. URL <http://dx.doi.org/10.1088/1475-7516/2017/04/038>. 8

- [35] Aab, A., Abreu, P., Aglietta, M., Ahn, E. J., Samarai, I. A., Albuquerque, I. F. M., *et al.* Searches for anisotropies in the arrival directions of the highest energy cosmic rays detected by the pierre auger observatory. *The Astrophysical Journal*, **804** (1), Apr 2015. URL <http://dx.doi.org/10.1088/0004-637X/804/1/15>. 9
- [36] Bellido, J., Collaboration, P. A., *et al.* Depth of maximum of air-shower profiles at the pierre auger observatory: Measurements above 10 17.2 ev and composition implications. En: 35th International Cosmic Ray Conference, tomo 301. SISSA Medialab, 2018. 10, 11, 103, 105, 106, 107, 108, 109
- [37] Abreu, P., Aglietta, M., Ahn, E., Albuquerque, I. F. d. M., Allard, D., Allekotte, I., *et al.* Measurement of the proton-air cross section at $s = 57$ tev with the pierre auger observatory. *Physical review letters*, **109** (6), 2012. 11, 12
- [38] Ellsworth, R., Gaisser, T., Stanev, T., Yodh, G. Ultrahigh-energy cross section from study of longitudinal development of air showers. *Physical Review D*, **26** (1), 1982. 12
- [39] Engel, R., Heck, D., Pierog, T. Extensive air showers and hadronic interactions at high energy. *Annual review of nuclear and particle science*, **61**, 2011. 13
- [40] Aab, A., Abreu, P., Aglietta, M. Pierre auger collab. *Phys. Rev. D*, **91**, 2015. 5, 13
- [41] Allekotte, I., Barbosa, A., Bauleo, P., Bonifazi, C., Civit, B., Escobar, C., *et al.* The surface detector system of the pierre auger observatory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **586** (3), Mar 2008. URL <http://dx.doi.org/10.1016/j.nima.2007.12.016>. 13, 17
- [42] Abraham, J., Abreu, P., Aglietta, M., Aguirre, C., Ahn, E., Allard, D., *et al.* The fluorescence detector of the pierre auger observatory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **620** (2-3), Aug 2010. URL <http://dx.doi.org/10.1016/j.nima.2010.04.023>. 14
- [43] Shellard, R. C. First results from the pierre auger observatory. *Brazilian journal of physics*, **36** (4A), 2006. 5, 14
- [44] Abraham, J., Abreu, P., Aglietta, M., Ahn, E., Allard, D., Allekotte, I., *et al.* Trigger and aperture of the surface detector array of the pierre auger observatory.

- Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **613** (1), 2010. 14
- [45] *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **798**, Oct 2015. URL <http://dx.doi.org/10.1016/j.nima.2015.06.058>. 5, 15, 16, 17
- [46] Greisen, K. The extensive air showers. *Progress in cosmic ray physics*, **3** (1), 1956. 15
- [47] Newton, D., Knapp, J., Watson, A. The optimum distance at which to determine the size of a giant air shower. *Astroparticle Physics*, **26** (6), Jan 2007. URL <http://dx.doi.org/10.1016/j.astropartphys.2006.08.003>. 15
- [48] Schulz, A. the measurement of the cosmic ray spectrum above 3×10^{17} ev with the pierre auger observatory. En: International Cosmic Ray Conference, tomo 33. 2013. 15
- [49] Pesce, R., *et al.* Energy calibration of data recorded with the surface detectors of the pierre auger observatory: an update. *Proc. 32nd ICRC, Beijing, China*, 2011. 16
- [50] Heat – a low energy enhancement of the pierre auger observatory. *Astrophysics and Space Sciences Transactions*, **7** (2), 183–186, May 2011. URL <http://dx.doi.org/10.5194/astra-7-183-2011>. 17
- [51] Unger, M., Dawson, B., Engel, R., Schüssler, F., Ulrich, R. Reconstruction of longitudinal profiles of ultra-high energy cosmic ray showers from fluorescence and cherenkov light measurements. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **588** (3), 2008. 18
- [52] Barbosa, H. M., Catalani, F., Chinellato, J. A., Dobrigkeit, C. Determination of the calorimetric energy in extensive air showers. *Astroparticle Physics*, **22** (2), 2004. 18
- [53] Aab, A., Abreu, P., Aglietta, M., Ahn, E., Samarai, I. A., Albuquerque, I., *et al.* The pierre auger observatory upgrade-preliminary design report. *arXiv preprint arXiv:1604.03637*, 2016. 18, 19
- [54] Stasielak, J. Augerprime—the upgrade of the pierre auger observatory. *arXiv preprint arXiv:2110.09487*, 2021. 18, 20

- [55] Letessier-Selvon, A., Billoir, P., Blanco, M., Maris, I. C., Settimo, M. Layered water cherenkov detector for the study of ultra high energy cosmic rays. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **767**, 2014. [19](#)
- [56] Gora, D., Collaboration, P. A., *et al.* The pierre auger observatory: review of latest results and perspectives. *Universe*, **4** (11), 2018. [19](#)
- [57] Suarez, F. The amiga muon detectors of the pierre auger observatory: overview and status. En: Proceedings, 33rd International Cosmic Ray Conference (ICRC2013): Rio de Janeiro, Brazil, tomo 712. Citeseer, 2013. [19](#)
- [58] Müller, S. N. Measurement of the cosmic ray composition with air showers detected by the amiga extension at the pierre auger observatory, 2019. [19](#)
- [59] Bergmann, T., Engel, R., Heck, D., Kalmykov, N., Ostapchenko, S., Pierog, T., *et al.* One-dimensional hybrid approach to extensive air shower simulation. *Astroparticle Physics*, **26** (6), 2007. [21](#), [25](#)
- [60] Pierog, T., Alekseeva, M., Bergmann, T., Chernatkin, V., Engel, R., Heck, D., *et al.* First results of fast one-dimensional hybrid simulation of eas using conex. *arXiv preprint astro-ph/0411260*, 2004. [21](#), [25](#)
- [61] Bossard, G., Drescher, H., Kalmykov, N., Ostapchenko, S., Pavlov, A., Pierog, T., *et al.* Cosmic ray air shower characteristics in the framework of the parton-based gribov-regge model nexus. *Physical Review D*, **63** (5), 2001. [21](#), [25](#), [26](#)
- [62] Heitler, W. The quantum theory of radiation, Second edition. International Series of Monographs on Physics. Oxford University Press, 1936. [21](#), [28](#)
- [63] Matthews, J. Energy flow in extensive air showers. págs. 261–264, 01 2001. [21](#), [28](#), [30](#), [34](#), [41](#)
- [64] Matthews, J. A Heitler model of extensive air showers. *Astropart. Phys.*, **22**, 387–397, 2005. [21](#), [28](#), [30](#), [41](#)
- [65] Cazon, L. Working group report on the combined analysis of muon density measurements from eight air shower experiments, 2020. [21](#), [47](#), [48](#), [49](#)
- [66] Cazon, L. Probing high-energy hadronic interactions with extensive air showers, 2019. [21](#), [32](#), [33](#), [53](#), [58](#), [86](#)
- [67] Cazon, L., Conceição, R., Martins, M. A., Riehn, F. Probing the π^0 spectrum at high-x in proton-air interactions at ultra-high energies, 2019. URL <http://dx.doi.org/10.1051/epjconf/201921002006>. [21](#), [53](#), [58](#)

- [68] Boezio, M., Mocchiutti, E. Chemical composition of galactic cosmic rays with space experiments. *Astroparticle Physics*, **39-40**, 95–108, Dec 2012. URL <http://dx.doi.org/10.1016/j.astropartphys.2012.05.007>. 21
- [69] Kampert, K.-H., Unger, M. Measurements of the cosmic ray composition with air shower experiments. *Astroparticle Physics*, **35**, 660–678, 2012. 21
- [70] Zyla, P., *et al.* Review of Particle Physics. *PTEP*, **2020** (8), 083C01, 2020. 22, 32, 33, 34
- [71] Grieder, P. Extensive air showers. High energy phenomena and astrophysical aspects. A tutorial, reference manual and data book. Vol. I and II, tomo 1. 2010. 23
- [72] Kobal, M., Collaboration1Argentina, P. A., CNEA, C.-I., Republic, C., ASU, F., *et al.* A thinning method using weight limitation for air-shower simulations. *Astroparticle Physics*, **15** (3), 2001. 26
- [73] Risse, M., Heck, D., Knapp, J. Eas simulations at auger energies with corsika. En: International Cosmic Ray Conference, tomo 2. 2001. 26
- [74] Pierog, T., Engel, R., Heck, D., Ostapchenko, S., Werner, K. Latest results from the air shower simulation programs corsika and conex. *arXiv preprint arXiv:0802.1262*, 2008. 26
- [75] Pierog, T., Engel, R., Heck, D., Poghosyan, G. Future of monte carlo simulations of atmospheric showers. En: EPJ Web of Conferences, tomo 89. EDP Sciences, 2015. 26
- [76] Pierog, T., Engel, R., Heck, D., Ulrich, R. 3d hybrid air shower simulation in corsika. En: Proc. of the International Cosmic Ray Conference. 2011. 26
- [77] Pierog, T., Engel, R., Heck, D. 3d air shower simulations using conex in corsika. En: Contributions to the 31st International Cosmic Ray Conference, Poland, 2009. 2009. 26
- [78] Alekseeva, M., Bergmann, T., Chernatkin, V., Engel, R., Heck, D., Kalmykov, N., *et al.* Extensive air shower simulation program conex: matching monte carlo and numerical methods. En: 29th International Cosmic Ray Conference (ICRC29), Volume 7, tomo 7. 2005. 26
- [79] Kalmykov, N., Ostapchenko, S., Werner, K. Results of eas characteristics calculations in the framework of the universal hadronic interaction model nexus. En: International Cosmic Ray Conference, tomo 2. 2001. 26

- [80] Aab, A., Abreu, P., Aglietta, M., Albuquerque, I., Albury, J. M., Allekotte, I., *et al.* Measurement of the average shape of longitudinal profiles of cosmic-ray air showers at the pierre auger observatory. *Journal of Cosmology and Astroparticle Physics*, **2019** (03), 2019. [26](#)
- [81] Antcheva, I., Ballintijn, M., Bellenot, B., Biskup, M., Brun, R., Buncic, N., *et al.* Root—a c++ framework for petabyte data storage, statistical analysis and visualization. *Computer Physics Communications*, **182** (6), 2011. [26](#)
- [82] Lang, T., van Hees, H., Steinheimer, J., Bleicher, M. Heavy quark transport in heavy ion collisions at rhic and lhc within the urqmd transport model. *arXiv preprint arXiv:1211.6912*, 2012. [27](#)
- [83] Werner, K., Liu, F.-M., Pierog, T. Parton ladder splitting and the rapidity dependence of transverse momentum spectra in deuteron-gold collisions at the bnl relativistic heavy ion collider. *Physical Review C*, **74** (4), 2006. [27](#)
- [84] Pierog, T., Werner, K. Epos model and ultra high energy cosmic rays. *Nuclear Physics B-Proceedings Supplements*, **196**, 2009. [27](#)
- [85] Pierog, T., Karpenko, I., Katzy, J. M., Yatsenko, E., Werner, K. Epos lhc: Test of collective hadronization with data measured at the cern large hadron collider. *Physical Review C*, **92** (3), 2015. [27](#)
- [86] Ostapchenko, S. Total and diffractive cross sections in enhanced pomeron scheme. *Physical Review D*, **81** (11), 2010. [27](#)
- [87] Ostapchenko, S. On the re-summation of enhanced pomeron diagrams. *Physics Letters B*, **636** (1), 2006. [27](#)
- [88] Ostapchenko, S. Nonlinear screening effects in high energy hadronic interactions. *Physical Review D*, **74** (1), 2006. [27](#)
- [89] Ahn, E.-J., Engel, R., Gaisser, T. K., Lipari, P., Stanev, T. Cosmic ray interaction event generator sibyll 2.1. *Physical Review D*, **80** (9), 2009. [27](#)
- [90] Riehn, F., Dembinski, H. P., Engel, R., Fedynitch, A., Gaisser, T. K., Stanev, T. The hadronic interaction model sibyll 2.3 c and feynman scaling. *arXiv preprint arXiv:1709.07227*, 2017. [27](#)
- [91] Riehn, F., Engel, R., Fedynitch, A., Gaisser, T. K., Stanev, T. Hadronic interaction model sibyll 2.3 d and extensive air showers. *Physical Review D*, **102** (6), 2020. [27](#)

- [92] Thakuria, C. C., Boruah, K. Comparison of epos and qgsjet-ii in eas simulation using corsika. *arXiv preprint arXiv:1202.3661*, 2012. 27
- [93] Linsley, J. Structure of large air showers at depth 834 g/sq cm. iii-applications. En: International Cosmic Ray Conference, tomo 12, págs. 89–96. 1977. 32
- [94] Pierog, T. Hadronic Interactions and Air Showers: Where Do We Stand? *EPJ Web Conf.*, **208**, 2019. 34, 36, 46
- [95] Airapetian, A. Multiplicity of charged and neutral pions in deep-inelastic scattering of 27.5 gev positrons on hydrogen. *The European Physical Journal C-Particles and Fields*, **21**, 599–606, 2001. 35
- [96] Adam, J., Adamova, D., Aggarwal, M. M., Rinella, G. A., Agnello, M., Agrawal, N., *et al.* *The European Physical Journal C*, **75** (5), 1–23, 2015. 36
- [97] Gaisser, T. K., Engel, R., Resconi, E. Cosmic Rays and Particle Physics. 2^a ed^{ón}. Cambridge University Press, 2016. 37, 48
- [98] Veberič, D. Lambert w function for applications in physics. *Computer Physics Communications*, **183** (12), 2622–2628, Dec 2012. URL <http://dx.doi.org/10.1016/j.cpc.2012.07.008>. 39
- [99] Abu-Zayyad, T., Belov, K., Bird, D., Boyer, J., Cao, Z., Catanese, M., *et al.* Evidence for changing of cosmic ray composition between 1017 and 1018 ev from multicomponent measurements. *Physical Review Letters*, **84**, 4276–4279, 05 2000. 47
- [100] Abreu, P., Andringa, S., Assis, P., Brogueira, P., Cazon, L., Conceição, R., *et al.* Muons in air showers at the pierre auger observatory: Mean number in highly inclined events, 2015. 47, 53
- [101] Dembinski, H. Computing mean logarithmic mass from muon counts in air shower experiments. *Astroparticle Physics*, **102**, 89–94, Nov 2018. URL <http://dx.doi.org/10.1016/j.astropartphys.2018.05.008>. 48
- [102] Dembinski, H., Engel, R., Fedynitch, A., Gaisser, T., Riehn, F., Stanev, T. Data-driven model of the cosmic-ray flux and mass composition from 10 gev to gev. *arXiv preprint arXiv:1711.11432*, 2017. 48
- [103] Meurer, C., Blümer, J., Engel, R., Haungs, A., Roth, M. Muon production in extensive air showers and its relation to hadronic interactions. *Czechoslovak Journal of Physics*, **56** (S1), A211–A219, Sep 2006. URL <http://dx.doi.org/10.1007/s10582-006-0156-9>. 50

- [104] Drescher, H.-J., Bleicher, M., Soff, S., Stöcker, H. Model dependence of lateral distribution functions of high energy cosmic ray air showers. *Astroparticle Physics*, **21** (1), 87–94, 2004. URL <https://www.sciencedirect.com/science/article/pii/S0927650503002640>. 50
- [105] Drescher, H., Hladik, M., Ostapchenko, S., Pierog, T., Werner, K. Parton-based gribov–regge theory. *Physics Reports*, **350** (2-4), 93–289, Sep 2001. URL [http://dx.doi.org/10.1016/S0370-1573\(00\)00122-8](http://dx.doi.org/10.1016/S0370-1573(00)00122-8). 50
- [106] Fletcher, R. S., Gaisser, T. K., Lipari, P., Stanev, T. sibyll: An event generator for simulation of high energy cosmic ray cascades. *Phys. Rev. D*, **50**, 5710–5731, Nov 1994. URL <https://link.aps.org/doi/10.1103/PhysRevD.50.5710>. 50
- [107] Engel, J., Gaisser, T. K., Lipari, P., Stanev, T. Nucleus-nucleus collisions and interpretation of cosmic-ray cascades. *Phys. Rev. D*, **46**, 5013–5025, Dec 1992. URL <https://link.aps.org/doi/10.1103/PhysRevD.46.5013>. 50
- [108] Glauber, R. J., Matthiae, G. High-energy scattering of protons by nuclei. *Nuclear Physics*, **21**, 135–157, 1970. 50
- [109] Khachatryan, V., Sirunyan, A. M., Tumasyan, A., Adam, W., Asilar, E., Bergauer, T., *et al.* Pseudorapidity distribution of charged hadrons in proton–proton collisions at $s = 13$ tev. *Physics Letters B*, **751**, 143–163, 2015. 50
- [110] Albrecht, J., Cazon, L., Dembinski, H., Fedynitch, A., Kampert, K.-H., Pierog, T., *et al.* The muon puzzle in cosmic-ray induced air showers and its connection to the large hadron collider, 2021. 51, 52
- [111] Ulrich, R., Engel, R., Unger, M. Hadronic multiparticle production at ultrahigh energies and extensive air showers. *Physical Review D*, **83** (5), Mar 2011. URL <http://dx.doi.org/10.1103/PhysRevD.83.054026>. 50, 92, 96
- [112] Baur, S., Dembinski, H., Perlin, M., Pierog, T., Ulrich, R., Werner, K. Core-corona effect in hadron collisions and muon production in air showers, 2020. 52, 53
- [113] Aab, A., Abreu, P., Aglietta, M., Ahn, E., Al Samarai, I., Albuquerque, I., *et al.* Muons in air showers at the pierre auger observatory: Mean number in highly inclined events. *Physical Review D*, **91** (3), 2015. 52
- [114] Adam, J., Adamova, D., Aggarwal, M. M., Rinella, G. A., Agnello, M., Agrawal, N., *et al.* Enhanced production of multi-strange hadrons in high-multiplicity proton–proton collisions, 2017. 53

- [115] Ostapchenko, S. Qgsjet: physics, recent improvements, and results for air showers. En: EPJ Web of Conferences, tomo 52. EDP Sciences, 2013. 53
- [116] Pierog, T., Werner, K. Muon production in extended air shower simulations. *Phys. Rev. Lett.*, **101**, 171101, Oct 2008. URL <https://link.aps.org/doi/10.1103/PhysRevLett.101.171101>. 53
- [117] Cazon, L., Conceição, R., Riehn, F. Probing the energy spectrum of hadrons in proton air interactions at ultrahigh energies through the fluctuations of the muon content of extensive air showers, Sep 2018. URL <http://dx.doi.org/10.1016/j.physletb.2018.07.026>. 53, 58
- [118] Cazon, L., Conceição, R., Martins, M., Riehn, F. Probing the high energy spectrum of neutral pions in ultra-high energy proton-air interactions, 2019. 53, 58
- [119] Cazon, L., Conceição, R., Martins, M. A., Riehn, F. Constraining the energy spectrum of neutral pions in ultra-high-energy proton-air interactions, Jan 2021. URL <http://dx.doi.org/10.1103/PhysRevD.103.022001>. 53, 58
- [120] Alner, G., Alpgård, K., Ansorge, R., Åsman, B., Böckmann, K., Booth, C., *et al.* Scaling violation favouring high multiplicity events at 540 gev cms energy. *Physics Letters B*, **138** (4), 1984. 54
- [121] Jones, E., Oliphant, T., Peterson, P., *et al.* Scipy: Open source scientific tools for python, 2001. 71
- [122] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, **12**, 2011. 71
- [123] Chollet, F., *et al.* keras, 2015. 71
- [124] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., *et al.* Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, 2013. 71, 85
- [125] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., *et al.* Tensorflow: A system for large-scale machine learning. En: 12th symposium on operating systems design and implementation 16. 2016. 71
- [126] Van Rossum, G., *et al.* Python programming language. En: USENIX annual technical conference, tomo 41. 2007. 71

- [127] Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., *et al.* Machine learning and the physical sciences. *Reviews of Modern Physics*, **91** (4), Dec 2019. URL <http://dx.doi.org/10.1103/RevModPhys.91.045002>. 72
- [128] A living review of machine learning for particle physics,. 72
- [129] Banko, M., Brill, E. Scaling to very very large corpora for natural language disambiguation. En: Proceedings of the 39th annual meeting of the Association for Computational Linguistics, págs. 26–33. 2001. 73
- [130] Smith, L. N. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, 2018. 76
- [131] Ge, R., Huang, F., Jin, C., Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. En: Conference on learning theory. PMLR, 2015. 76
- [132] Glorot, X., Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. En: Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010. 77
- [133] Géron, A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O’Reilly Media, 2019. 77
- [134] Teodorescu, L. Artificial neural networks in high-energy physics, 2008. 77
- [135] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*, **61**, Jan 2015. URL <http://dx.doi.org/10.1016/j.neunet.2014.09.003>. 77
- [136] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. Classification and Regression Trees. Belmont, CA: Wadsworth International Group, 1984. 85
- [137] Breiman, L. Bagging predictors. *Machine learning*, **24** (2), 123–140, 1996. 85
- [138] Breiman, L. Machine learning, volume 45, number 1 - springerlink. *Machine Learning*, **45**, 5–32, 10 2001. 85
- [139] Bergstra, J., Bengio, Y. Random search for hyper-parameter optimization. *Journal of machine learning research*, **13** (2), 2012. 85
- [140] Davison, A. C., Hinkley, D. V. Bootstrap methods and their application. 1. Cambridge university press, 1997. 92

- [141] Aab, A., Abreu, P., Aglietta, M., Ahn, E., Al Samarai, I., Albuquerque, I., *et al.* Depth of maximum of air-shower profiles at the pierre auger observatory. i. measurements at energies above $10^{17.8}$ eV. *Physical Review D*, **90** (12), 2014. [99](#), [100](#)
- [142] Ji, X., Gu, W., Qian, X., Wei, H., Zhang, C. Combined neyman–pearson chi-square: An improved approximation to the poisson-likelihood chi-square. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **961**, 2020. [99](#)
- [143] Demortier, L. Equivalence of the best-fit and covariance-matrix methods for comparing binned data with a model in the presence of correlated systematic uncertainties. *CDF*, 1999. [99](#)
- [144] Lista, L. Statistical methods for data analysis in particle physics, tomo 941. Springer, 2017. [99](#)
- [145] Flannery, B. P., Press, W. H., Teukolsky, S. A., Vetterling, W. Numerical recipes in c. *Press Syndicate of the University of Cambridge, New York*, **24** (78), 1992. [100](#)
- [146] Schmidt, F., Ave, M., Cazon, L., Chou, A. Applying extensive air shower universality to ground detector data. *arXiv preprint arXiv:0706.1990*, 2007. [103](#), [104](#)
- [147] Aab, A., Abreu, P., Aglietta, M., Ahn, E., Al Samarai, I., Albuquerque, I., *et al.* Depth of maximum of air-shower profiles at the pierre auger observatory. ii. composition implications. *Physical Review D*, **90** (12), 2014. [105](#)
- [148] Baur, S., Dembinski, H., Perlin, M., Pierog, T., Ulrich, R., Werner, K. Core-corona effect in hadron collisions and muon production in air showers. *arXiv preprint arXiv:1902.09265*, 2019. [112](#)
- [149] Pierog, T., Guiot, B., Karpenko, I., Sophys, G., Stefaniak, M., Werner, K. Epos 3 and air showers. En: EPJ Web of Conferences, tomo 210. EDP Sciences, 2019. [112](#)
- [150] Anchordoqui, L. A., Canal, C. G., Sciutto, S. J., Soriano, J. F. Through the looking-glass with alice into the quark-gluon plasma: A new test for hadronic interaction models used in air shower simulations. *Physics Letters B*, **810**, 2020. [112](#)

Acknowledgments

I would like to begin thanking my family. Manu: my companion, my love, thank you for being there by my side always, giving me strength, supporting me. You helped me gain confidence and believe in myself. Thank you for making this intense time such a happy and fun one! Claudia and Marcel: thank you for sharing your home with me, for every delicious meal, every fun moment, every time you reminded me of something I forgot to do. I will remember my time in Pfinztal as such a happy one! Eunise, Gustavo y Viqui: gracias por haberme adoptado y recibido en su hogar. Pude estar con ustedes en un momento complicado y gracias a todo el cariño que me dieron puedo recordar esas semanas con tanta felicidad. Gracias por todas las charlas, los buenos momentos, el ánimo y apoyo que me dieron! Leo, hermano: gracias por haberme dado un hogar en este año tan loco. Fuiste un gran apoyo moral, me ayudaste con tantas cosas que no tiene nombre. Gracias a Ema también. Entre los dos me regalaron un año muy feliz! Karin, Helene and Wim: thank you for all your support, happy moments and for knowing that I can count on you. Kitty, Buh, Pepe, Stinki: my furry little brothers and sisters. Thank you for all the cuddling! I am eternally grateful to all of you.

I would like to thank Tanguy and Xavier for guiding me through this process and reviewing my thesis. I would like to thank also Johannes Blümer and Ingo Allekotte for reviewing this thesis. Thanks to Matias Perlin, Ana Botti, Belén Andrada, Michael Unger and especially Ingo and Diego Ravignani, for all the great, interesting and fun discussions. I would also like to thank Sabine Bucher and Marie-Christine Mundt, for helping me with the bureaucracy. Thanks to Javier Fernandez for sharing the passion in education with me. I would like to thank all my colleagues at IAP and ITeDA. Thanks to the Grupo Partículas y Campos in Bariloche, my professional home.

I would like to thank my friends: Ana (Gramajo), Caro, Eve, Gaia, Raúl, Gaby, Ayelén, Mica, Matias (Roncoroni), Eugenio, Sam, Janis, Ale, Carmina, Hernán, Leandro, Mario (Solis), Lauri, Eli, Flor, Paola, Sergio, Geri, Max (Reininghaus), Kathrin, Martin, Max (Stadelmaier), Alex, Juli, Maritza, Edu, Naty, Mario (Pérez), Sima. With all of you I shared my feelings, problems and doubts and you helped me so much overcome bad feelings and get on the right track again. A huge thank you!

