DETERMINATION OF PHYSICAL PROPERTIES OF HIGH-ENERGY HADRONIC INTERACTIONS FROM THE $X_{\max}$-N_{μ} ANTICORRELATION

Zur Erlangung des akademischen Grades einer
Doktorin der Naturwissenschaften (Dr. rer. nat.)

von der KIT-Fakultät für Physik des

Karlsruher Instituts für Technologie (KIT)

angenommene DISSERTATION von

Mgr. Isabel Astrid Goos

Tag der mündlichen Prüfung: 11.02.2022

Referent: Dr. Xavier Bertou

Korreferent: Prof. Dr. Dr. h.c. Johannes Blümer

Betreuer: Dr. Tanguy Pierog

A todas las personas
que me acompañaron
en esta etapa

Contents

Figure Index

Table Index

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 the 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 that 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 of 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 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 componen. 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 reducir su número. Esto significa que existe una posición de máximo desarrollo $X_{\max}$. Las partículas que interactúan hadrónicamente producen muones que se pueden detectar a nivel del suelo. Este número 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_μ 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 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 hadrónica 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 Objekete 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 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, im Allgemeinen 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 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 because 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 made the detection of individual cosmic rays and the precise determination of their arrival times possible. Bothe and Kolhörster coupled the coincidence technique with the Geiger-Müller counter to study cosmic rays, marking thus 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 discoveries 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 the muon in cloud-chamber photographs [9], which, because of its appropriate mass, was at first wrongly

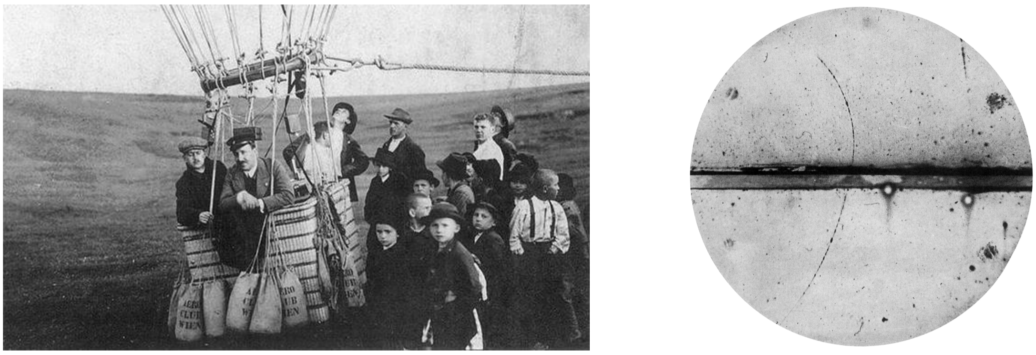

Figure 1.1: Left: Victor Hess, at the center, departing from Vienna about 1911, soon discovering the penetrating ionizing radiation from outer space. Right: Anderson's cloud chamber picture of cosmic radiation from 1932, showing for the first time the existence of the positron (picture taken from [4]).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

identified as 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 that 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-

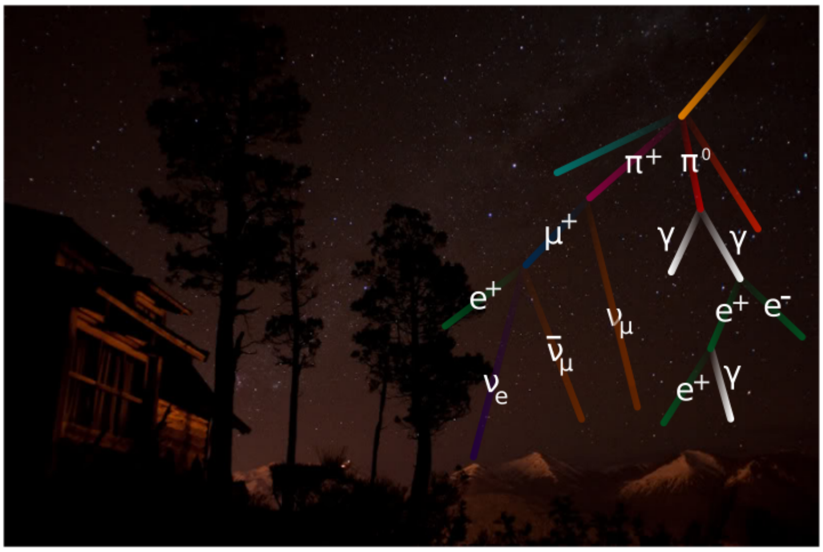

Figure 1.2: Simplified sketch of an extensive air shower generated by a cosmic ray. It can be appreciated that the generation of new particles and subsequent subshowers quickly results in a large number of particles. As the number of particles increases, the average individual energy carried by them 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]. Today, we understand that cosmic rays interact with molecules in the atmosphere, setting off 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 that 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 of coincidences between a pair of Geiger-Müller counters develops 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 the ground to infer the point of origin of the air shower that generated them, what we will call the *depth of the 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 the ground would carry, on average, the *critical energy*. The concept of critical energy will be used frequently in this work and defined in detail in section 3.4. The values obtained 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. However, since a huge portion of cascading processes is electromagnetic in nature, 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 proportionally 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 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 [25, 26, 27]. This is the largest observatory up to date for the 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 and complementary ways to measure air showers. The first method detects charged particles through their interaction with the water contained in the surface detector tanks, while the second one tracks the development of air showers by observing the 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 the mechanisms of cosmic ray acceleration and propagation. Figure 1.3 shows the “all-particle” spectrum as a combination of 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 otherwise. 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 that explain the physical origin of the knee feature [29]. 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 µG, 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 becomes significantly less efficient in our galactic disk, causing thus a steepening of the cosmic ray spectrum. This scenario also requires the existence of galactic cosmic ray sources that are capable of accelerating particles to energies well beyond the knee energy, in the first place. An alternative explanation to the knee feature is that the PeV energy scale actually is the highest energy achievable

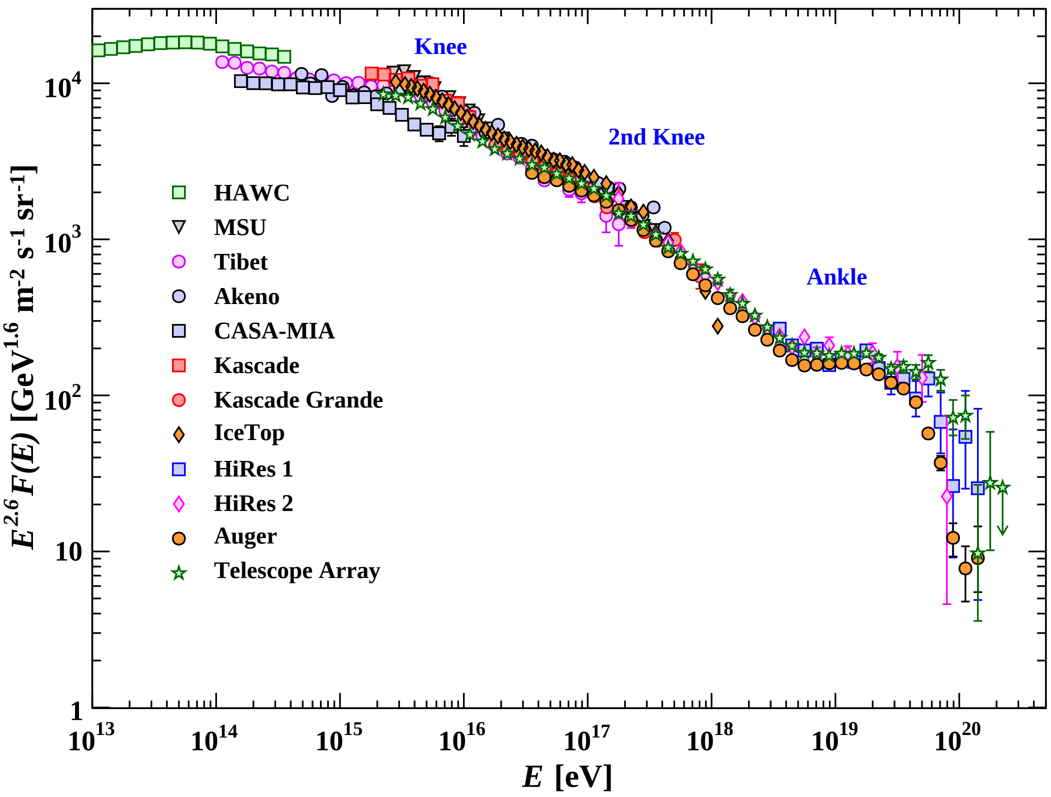

Figure 1.3: Cosmic ray differential energy spectrum reconstructed from air showers observed by a variety of experiments. The spectrum has been multiplied by $E^{2.6}$, in order to enhance the features that are present when the spectral index of the power law describing the spectrum changes. These features are the so-called *knee* between 10^{15} 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. Figure taken from [28].

in galactic cosmic ray sources, which are thought to be supernova remnants. The *Hillas criterion* can be used to study this possibility. It states that

$$E_{\mathrm{max}} = \eta^{-1}\beta_{\mathrm{sh}} qBR \tag{1.1}$$

is the maximum energy achievable by a particle of charge q at a source of characteristic size $R = l \cdot \Gamma$, surrounded by a magnetic field of strength B. l is the co-moving 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 the Larmor radius of the particle not to exceed the size of the acceleration region. It shows that a considerable magnetic field is needed in order to see feasible candidates in supernova remnants (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 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 [30].

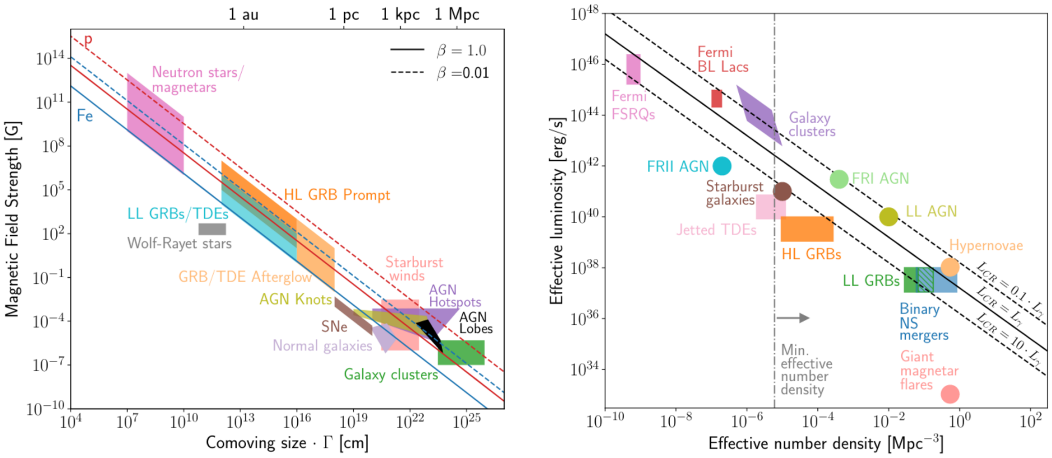

Figure 1.4: Left: Hillas diagram. It shows typical characteristic sizes R and magnetic fields B of different astrophysical sources [31]. Solid and dashed lines show for which values of B and R confinement of protons (red) and iron nuclei (blue) with an energy of 10^{20} eV are possible. Right: Source luminosity and number density of different astrophysical objects. Different black lines correspond to different values of 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, as detected at the Pierre Auger Observatory [32]. Both diagrams are taken from [33].

At energies just above that of the second knee, a steepening in the flux of heavy primary particles was measured by the KASCADE-Grande experiment [34]. The subsequent flattening of the spectrum means an onset of a new light component, which needs to be of extra-galactic origin. The flux suppression at the highest energies in the cosmic ray spectrum happens close to the energy threshold of about 6×10^{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:

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

or

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

However, a scenario solely based on the GZK-effect does not fit the flux measured at the Pierre Auger Observatory [35]. The cut-off could instead be explained by a combination of propagation effects and the maximum energy achievable at the source [36].

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 [33]) 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 an energy of 10^{20} eV are possible. Objects to the left of the lines do not satisfy the Hillas criterion to be able to accelerate cosmic rays up to this energy [31].

In addition to certain sizes and magnetic fields, cosmic ray accelerators also 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 [33]). The energy budget corresponding to each combination of values is compared to the ultra-high energy cosmic ray rate estimated in [38]. The regions to the left of the black lines do not satisfy the energy budget condition, while each black line corresponds to different values of 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, as detected at the Pierre Auger Observatory [32].

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 extra-galactic 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 to look for a cumulative flux from multiple close by objects. An anisotropy in the arrival directions was indeed detected at a more than 5.2σ level of significance by the Pierre Auger Collaboration [37]. This anisotropy can be described by a dipole, which is shown in figure 1.5. The direction of the dipole indicates an extra-galactic origin of ultra-high energy particles.

The Pierre Auger Collaboration also performed searches at smaller angular scales and correspondingly higher energies [39]. No statistically significant evidence of anisotropy was discovered, but some departure from isotropy was observed 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.

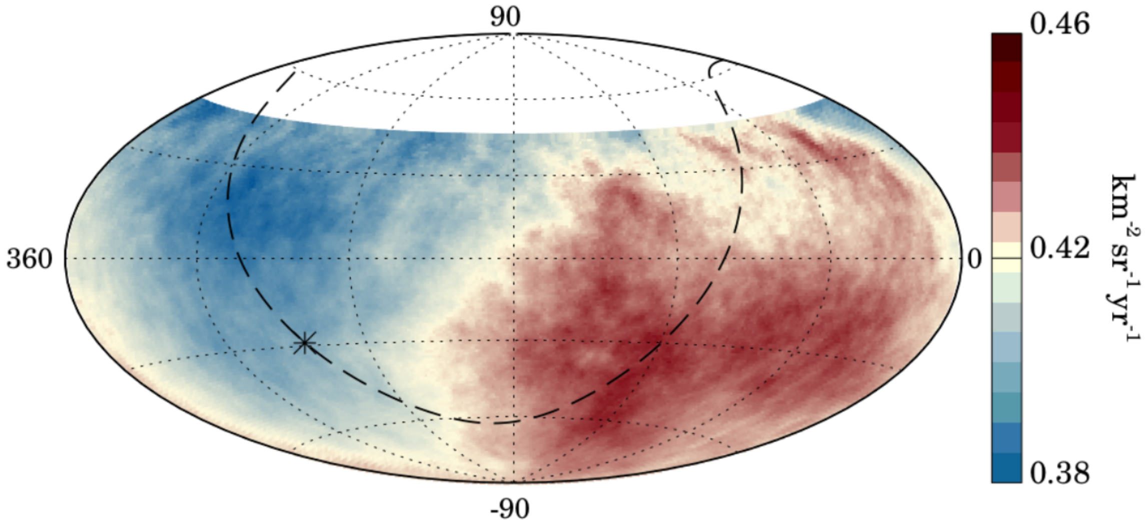

Figure 1.5: Smoothed cosmic ray flux for energies above 3 EeV, in equatorial coordinates, as measured at the Pierre Auger Observatory (figure taken from [37]). The dashed line follows the galactic plane, while the star indicates the center of the galaxy. The region with no values is not covered by the 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 [40], X_{max} distributions of simulated proton, helium, nitrogen and iron initiated showers were parameterized in various 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 in the cosmic ray flux, 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 primary 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

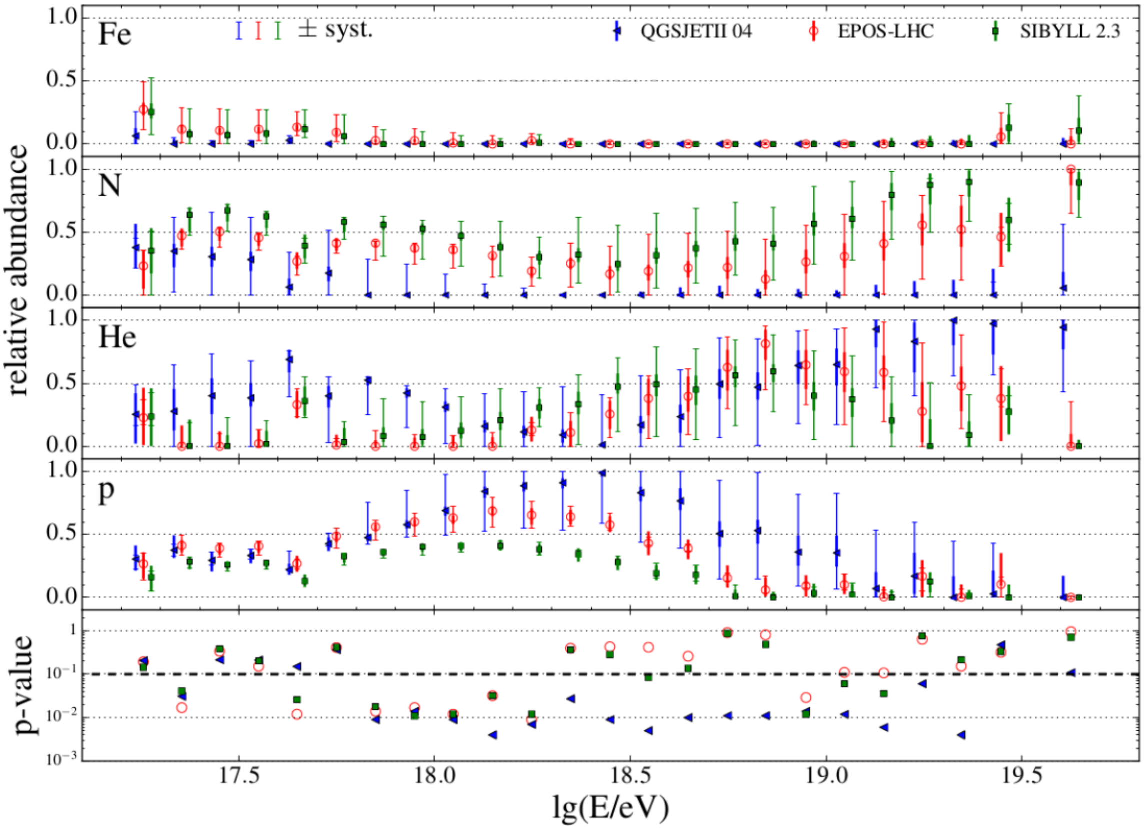

Figure 1.6: Mass fraction fits obtained by the Pierre Auger Collaboration (figure taken from [40]). The error bars indicate the statistic (smaller cap) and the systematic uncertainties (larger cap). The bottom panel indicates the goodness of the fits (p-values).

model used. However, 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 measured distribution of $X_{\max}$ at different energies with the corresponding values for sets of simulated proton showers and simulated iron showers, separately [40]. 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 QGSJETTII-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), which 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

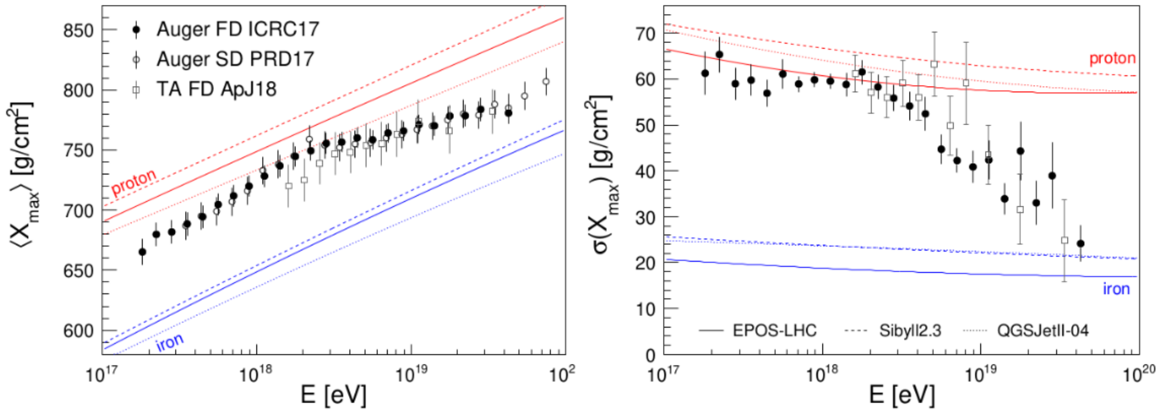

Figure 1.7: Measurements of the mean (left) and the standard deviation (right) of the distribution of $X_{\max}$ at different energies. The red and blue lines correspond to the values obtained from simulations considering pure proton and pure iron fluxes, respectively. Full, dashed and dashed-dotted lines correspond to the high-energy interaction models EPOS-LHC, SIBYLL-2.3 and QGSJETTII-04, respectively. These models are implemented in the simulation frameworks used and will be discussed in section 3.3. Both figures are taken from [40].

such study was carried out at the Pierre Auger Observatory in order to measure the proton-air cross-section [41]. This cross-section is of particular interest because it is not 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 [42] 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 relation will be used in the second step as a conversion function. The hadronic inelastic 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 that 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 an inelastic cross-section, using the aforementioned relation. 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 [41]. Knowing this cross-section is crucial for the present work.

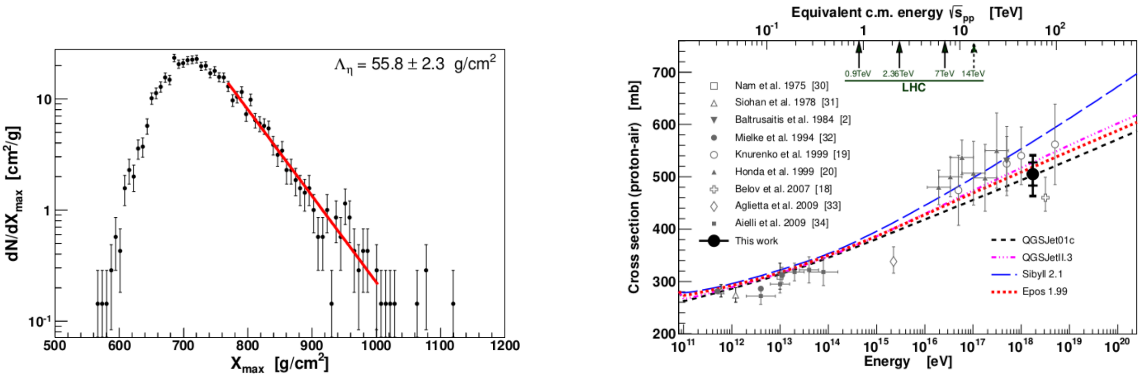

Figure 1.8: Left: X_{max} distribution together with the likelihood fit to obtain Λ_η, the slope of the tail of this distribution. This observable can be converted into an inelastic cross-section σ_{inel}. Right: Resulting cross-section $\sigma_{\mathrm{inel}}^{\mathrm{p-Air}}$ for proton-air interactions (black point) compared to measurements and predictions of simulations, using the aforementioned high-energy interaction models. Both figures are taken from [41].

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 [43]. 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 has been designed to study extensive air showers generated by cosmic rays of energies above 10^{18} eV [25], employing a hybrid detection technique, which was briefly introduced in section 1.1.

On the one hand, the *Surface Detector* (SD) [44] consists of an array of more than 1660 water-Cherenkov Detectors (WCDs) that measure the energy deposited by particles from air showers at the ground. In figure 2.1, each dot corresponds to one such

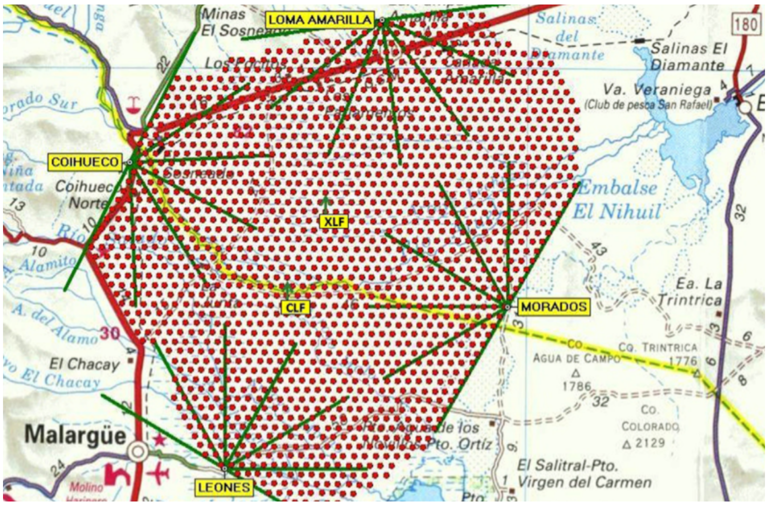

Figure 2.1: Layout of the Pierre Auger Observatory in the Pampa Amarilla near Malargüe in Mendoza, Argentina. Each dot corresponds to one of the 1660 surface detector stations. The four fluorescence detector enclosures are shown, each with the 30° field of view of its six telescopes. Also shown are the two laser facilities CLF and XLF, near the Observatory center, that are used for the calibration of the FD. This figure is obtained from [25].

detector station. On the other hand, the *Fluorescence Detector* (FD) [45] 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 km^2, 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 a 1500 m distance between stations. For this distance between stations, the energy threshold at which the array becomes fully efficient is 3×10^{18} eV [46]. Each surface detector station consists of a cylindrical polyethylene tank of about 3.6 m in diameter that 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 [26]). The inner surface of the tanks 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 that are 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 1.2 m from the center of the tank and look downwards into the water. The inner surface reflects diffusely, 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, a

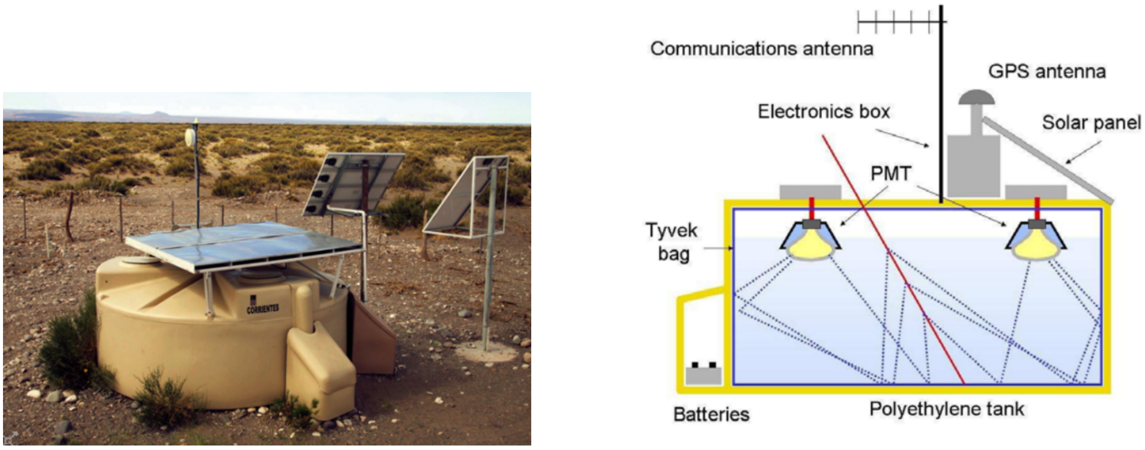

Figure 2.2: Left: Picture of a WCD. Right: Schematic representation of a WCD with its different components (taken from [26]).

GPS receiver, a radio transceiver and a power controller.

An approximate estimate 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 that grows with the speed of light.

From this sample, an observable to estimate the shower size can be defined, too. 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 [31]. In figure 2.3 (right), the lateral distribution of the signals at the ground is depicted for an event produced by a cosmic ray of energy (104 ± 11) EeV and zenith angle (25.1 ± 0.1) degrees [27]. The function employed to fit the lateral distribution of the signals at the ground is a modified *Nishimura-Kamata-Greisen* function [23, 47]:

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

where r_{opt} is the optimal distance from the shower core for this calculation, $r_1 = 700\,\text{m}$ and $S(r_{\text{opt}})$ is an estimator of the shower size at r_{opt}. At this optimal core distance, the fluctuations in the expected signal $S(r_{\text{opt}})$, due to a lack of knowledge of the exact lateral distribution function, are minimized. The value for the optimal distance is $r_{\text{opt}} = 1000\,\text{m}$ [48].

The value of $S(1000)$ varies with the zenith angle θ, due to the attenuation of the shower particles and due to geometrical effects. A function $f_{\text{CIC}}(\theta)$, that describes this attenuation, can be used to convert $S(1000)$ to a new expression [49]:

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

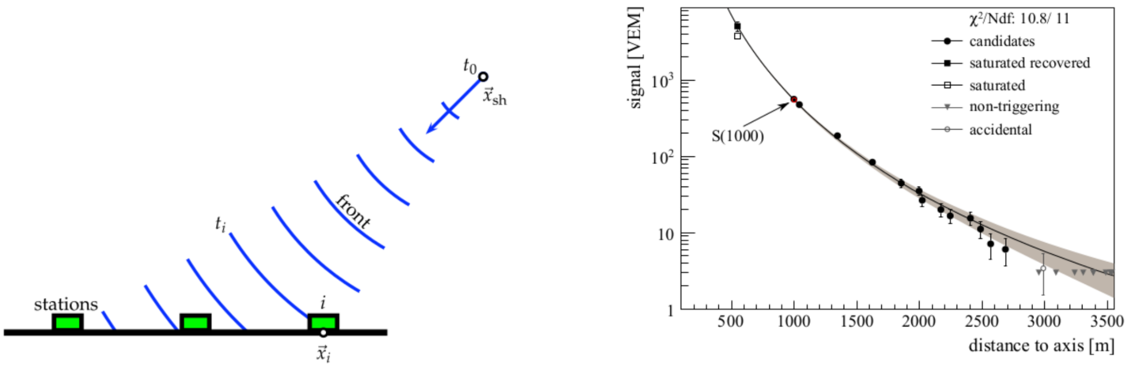

Figure 2.3: Left: Schematic representation of the evolution of the shower front, taken into account for the reconstruction of the shower geometry. Right: lateral distribution of the signals at the ground for an event produced by a cosmic ray of energy (104 ± 11) EeV and zenith angle (25.1 ± 0.1) degrees. Both figures are taken from [27].

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 the SD and the FD independently are used for cross-calibration [50]. 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:

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

The parameters from the data fit are $A = (1.9 \pm 0.05){\times}10^{17}\,\mathrm{eV}$ and $B = (1.025 \pm 0.007){\times}10^{17}\,\mathrm{eV}$. 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 of the energy calculated by the SD is then

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

2.2. The fluorescence detector

The basic FD consists of 24 telescopes that 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 that 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, were

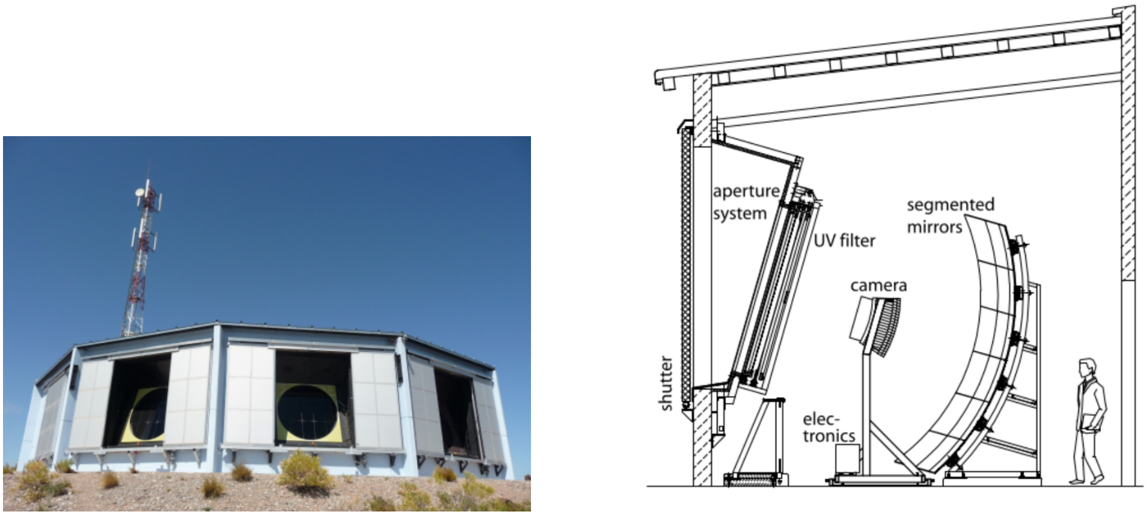

Figure 2.4: Left: Picture of the FD building at Los Leones. Right: Schematic representation of a fluorescence telescope with its different components (taken from [27]).

deployed near the FD site at Coihueco. The so-called High Elevation Auger Telescopes (HEAT) consist of similar telescopes that can be tilted up to a maximum angle of 60°. This extends the applicability of the Pierre Auger Observatory to the measurement of shallower showers initiated by lower energy primaries [51]. 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 a 750 m spacing, the so-called SD-750 [27].

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 and can be measured by the fluorescence telescopes (see figure 2.4). It enters the telescopes passing a UV transmitting filter that 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. With these considerations, 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 that is closest to it, as depicted in figure 2.5 (left, taken from [44]). Next, the timing information of the pixels is used to reconstruct the shower axis. There might be some degeneracy in this calculation that 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

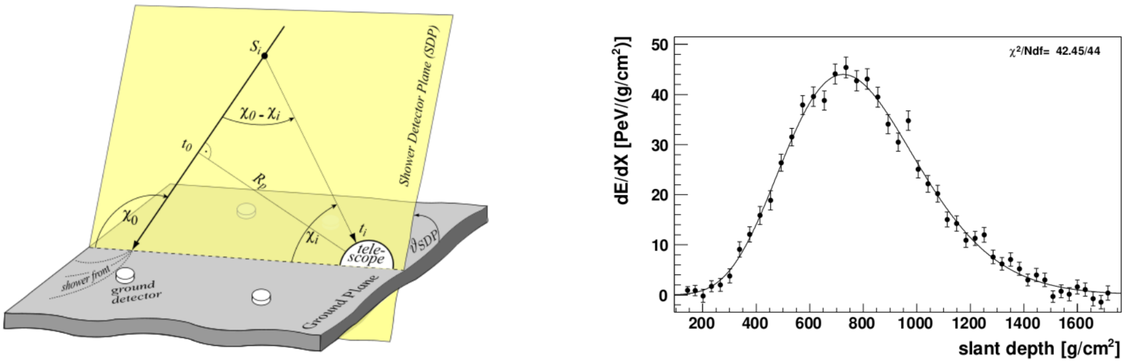

Figure 2.5: Left: Illustration of the geometrical shower reconstruction from the observables of the shower front. Right: Energy deposit profile as estimated using the FD. Both figures are taken from [44].

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 that will be explained in section 3.1 (see figure 2.5, right). After estimating the attenuation from the shower and disentangling all contributing light sources, one is left with the calorimetric energy [52]. 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 cosmic ray. Besides, from the energy deposit profile shown in figure 2.5 (right), an estimate of the position X_{max} of maximum development can be obtained [53].

2.3. AugerPrime

As will be explained in section 3.5, there is a discrepancy between 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 that introduce uncertainties in all analyses. In particular, the measurement of the mass composition of the cosmic ray flux, which is important in order 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 the muonic and electromagnetic components, when measurements are carried out [54]. 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 $2\,\text{m}^2$ [55, 54] (see figure 2.6). Each module consists of 24 fibers that 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} will estimated with a duty cycle of 100 %.

A so-called *matrix formalism*, as developed in [56] 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 direct measurement of the muonic component of extensive air showers [58]. 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 m^2 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 of the electromagnetic particles of the shower and imposes a cut-off for vertical muons of 1 GeV [59]. 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 a direct counting of the muons [54]. Muon counters sample scintillator signals at a frequency of 320 MHz, which

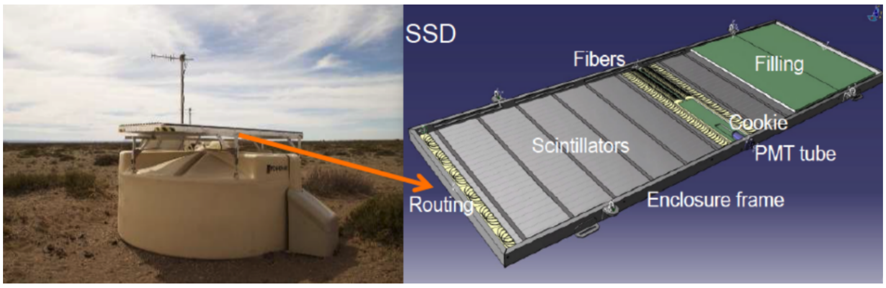

Figure 2.6: Left: Picture of one upgraded station. Right: Layout of the surface scintillator detector showing its different components. This picture is taken from [57].

means that every 3.125 ns 64 bits are acquired. Each bit stores the digitized value (either a “0” or a “1”, if the signal was below or above a predefined threshold, respectively) associated to one scintillator bar. Muon counting and the digitization of the integrated signal are 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 [55]. 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).

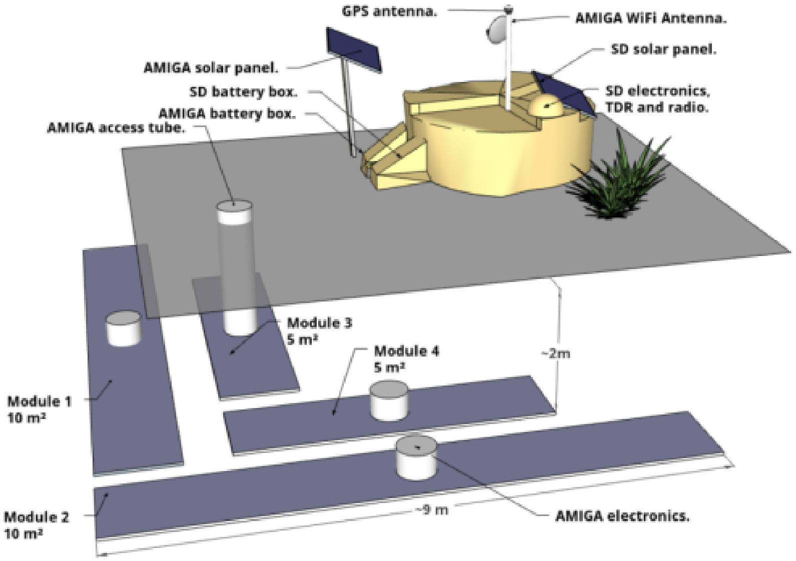

Figure 2.7: General overview of an underground muon detector from AMIGA. Both surface and underground detectors are shown in their arrangement during the prototype phase. In the final design, the $30\,\mathrm{m}^2$ detector is split into three $10\,\mathrm{m}^2$ modules. This figure is taken from [55].

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 [60, 61, 62]. Afterwards, we move on to a more detailed description of extensive air showers, following the semi-empirical model developed by Heitler [63] and Matthews [64, 65]. Those works concentrate on showers of lower energy 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 [66] 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. [67, 68, 68], 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 [69, 70]. 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 kind of 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 $0\,\mathrm{g\,cm^{-2}}$ and the sea level to $1030\,\mathrm{g\,cm^{-2}}$. In figure 3.1, an example of such a longitudinal profile of a simulated shower is shown. The numbers of particles are displayed separated into different components that 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 $c\tau = 7.8\,\mathrm{m}$ [71], 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 the ground. For this reason, they carry information about hadronic interactions in extensive air showers. In this work, the total number of muons at the ground is used and denoted by N_μ. In figure 3.1, the ground level of the Pierre Auger Observatory is represented by the dashed violet line at $\sim 880\,\mathrm{g\,cm^{-2}}$. 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 $c\tau = 25.5\,\mathrm{nm}$ [71]). These photons originate electromagnetic subshowers. All these subshowers together build up the electromagnetic component. Due to the short interaction length of electromagnetic particles [71], any information on the distribution of the neutral pions is lost. In figure 3.1, the electromagnetic component is further sepa-

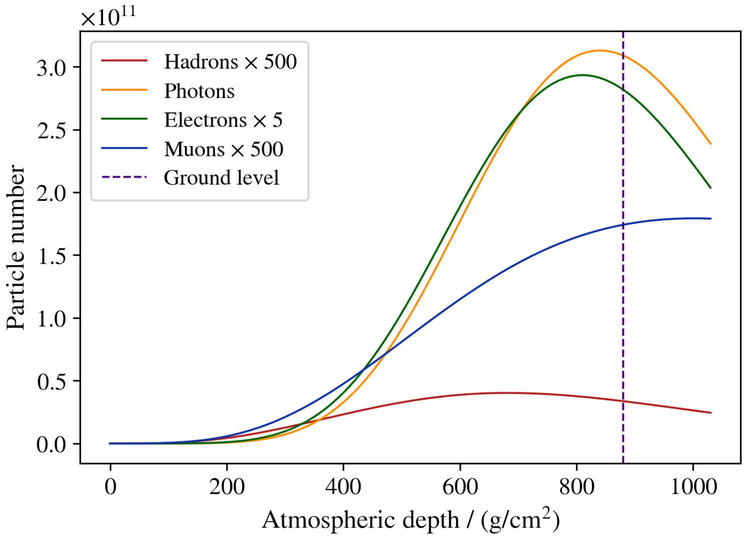

Figure 3.1: Longitudinal profiles of the hadrons, photons, electrons and muons of a simulated shower initiated by a vertical proton of 10^{20} eV and simulated using the high-energy interaction model EPOS-LHC. Some components are scaled for visualization purposes. The ground level represents the depth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

rated 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 *electrons*. Note that the vast majority of the particles in a shower are of electromagnetic nature. This component grows in number until the energy per particle is low enough ($\sim 85\,\mathrm{MeV}$) for these to be absorbed by the atmosphere. A maximum size of the shower is thus reached at a certain depth in the atmosphere X_{max}, 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 [72]. For the case of a vertical trajectory and the *standard isothermal atmosphere*, the following simple expression can be obtained:

$$X(h) = X_0 \cdot e^{-(h/h_{\mathrm{s}})}. \tag{3.1}$$

$X_0 \simeq 1030\,\mathrm{g\,cm^{-2}}$ 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 8400 m at sea level. In this case, the *barometric equation* is given by

$$\rho(h) = \rho_0 \cdot e^{-(h/h_{\mathrm{s}})},$$

where $\rho_0 = 0.001\,07\,\mathrm{g\,cm^{-3}}$ 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 the ground N_μ is 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 distribution 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 anticorrelation. 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 the 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 elaborate a model that predicts $X_{\max}$ and N_μ as a function of these parameters, in

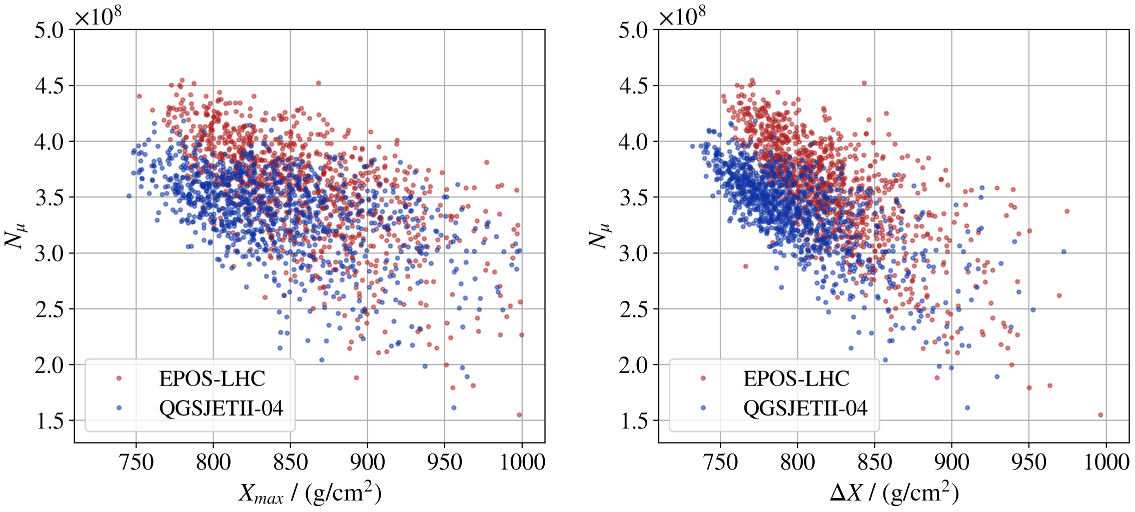

Figure 3.2: Anticorrelation between the depth $X_{\max}$ of maximum shower development and the number N_μ of muons at the ground (left). Sets of 1000 showers generated by proton primaries of 10^{20} eV are used. One set is generated using EPOS-LHC (red points) and the other one using QGSJETII-04 (blue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

a way that their anticorrelation is reproduced.

We also investigate the anticorrelation for different primary masses. As is shown 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. Furthermore, the anticorrelation decreases 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 that 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 [60, 61, 62]. 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 the 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

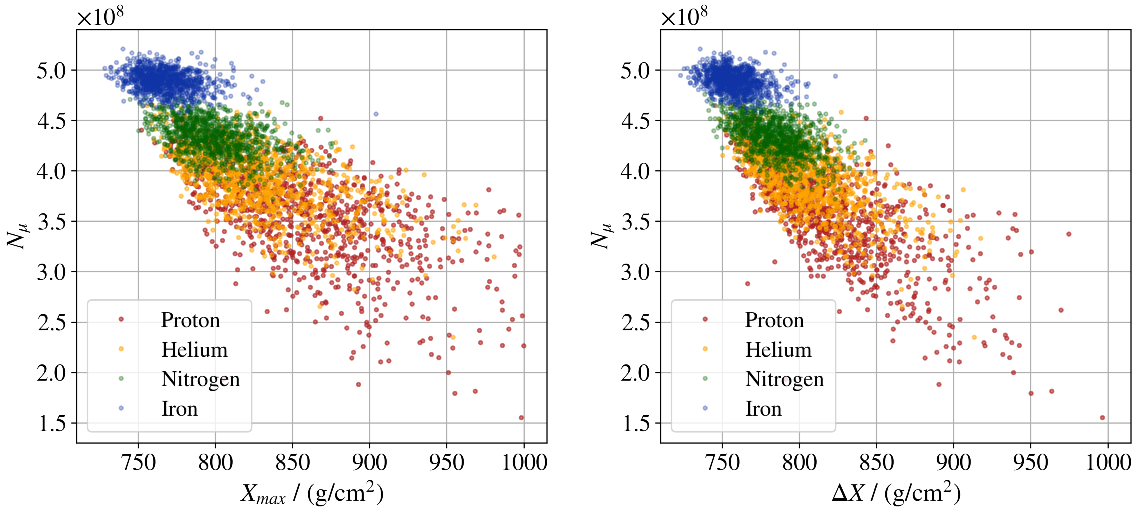

Figure 3.3: Anticorrelation between the depth $X_{\max}$ of maximum shower development and the number N_μ of muons at the ground (left). Sets of 1000 vertical showers, generated by different primaries of 10^{20} eV, are used. These simulations are carried out using the high-energy interaction model EPOS-LHC.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

be required for the primary energies we are interested in. Even applying weighted sampling algorithms included in the software, like the *thinning method* [73], are no solution. Here, the computing time is reduced by treating only a small portion of representative particles explicitly, 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 [74].

In CONEX, this problem is solved combining explicit Monte Carlo simulation of the highest energy portion of the air shower (first few interactions), with numerical expressions for the lower energy part [75, 76, 77, 78, 79]. 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 [62]. Thus, average extensive air shower parameters are accurate as well [80].

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 the ground N_μ, which we need for the present work. The depth of maximum development X_{max} we need for our study is calculated and output by CONEX, after performing a fit of the Gaisser-Hillas function [81]

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

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 properties is an extension that enables the output of ROOT files [82] containing detailed information pertaining all the 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 the 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 the 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 the 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 [83]. 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 and argon molecules in the atmosphere) is specific to generators designed for high energies. In the first part of the work, we will analyze showers generated using EPOS-LHC [84, 85, 86] and QGSJETTII-04 [87, 88, 89]. It is interesting to use these two models because, as we shall see in later sections, distributions of certain physical parameters are noticeably 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 will also include simulations performed using SIBYLL-2.3d [90, 91, 92]. Even though associated results are quite similar to those obtained implementing EPOS-LHC, the small differences make it an interesting alternative to use. Each model describes physical processes in a different way.

The generators QGSJET and SIBYLL only implement phenomena that are important for the 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 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 [93]. It is not clear, as of today, which model represents physical processes at the highest energies more truthfully. Each modeling of the physical processes leads to slight differences in the parameter and observable distributions. Therefore, it makes sense to use the three high-energy interaction models. 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 simple 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 (see chapter 5), special care needs to be taken. Since neural networks behave mostly 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 favorable 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 [63] and Matthews [64, 65]. The calculations of formulas presented in this section are reproduced from [64, 65, 1, 2].

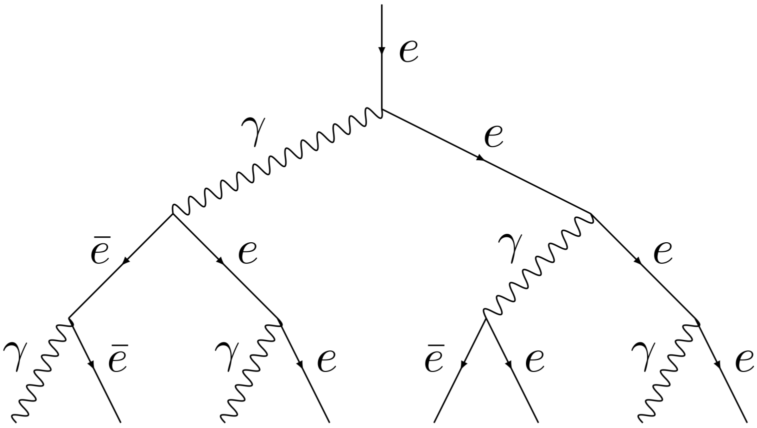

Figure 3.4: Schematic representation of the Heitler splitting model of electromagnetic cascades [63]. e represents an electron or positron and $\bar{e}$ the corresponding anti-particle. An electromagnetic subshower is mainly driven by electron-positron pair production induced by photons and bremsstrahlung experienced by electrons.

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 [63]. The initiating particle can be a photon or an electron. In this model, particles are assumed to travel one splitting length $X_\mathrm{r} = \lambda_\mathrm{r} \ln 2$, where λ_r is the *radiation length*, and create then new particles. Photons produce an electron-positron pair and electrons undergo bremsstrahlung (see figure 3.4). Since the energy loss through radiation for electrons is described by $E(X) = E_0 e^{-X/\lambda_\mathrm{r}}$, 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 of 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 $\lambda_\mathrm{r} = 37\,\mathrm{g\,cm^{-2}}$ and $\xi_\mathrm{c}^e = 85\,\mathrm{MeV}$.

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

$$N_\mathrm{max}^\gamma = 2^{n_\gamma} = \frac{E_0}{\xi_\mathrm{c}^e}.$$

Consequently, the maximum number of generations can be expressed as

$$n_\gamma = \frac{\ln\left(E_0/\xi_\mathrm{c}^e\right)}{\ln 2}$$

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

$$X_\mathrm{max}^\gamma = n_\gamma \lambda_\mathrm{r} \ln 2 = \lambda_\mathrm{r} \ln\left(\frac{E_0}{\xi_\mathrm{c}^e}\right). \tag{3.3}$$

The *elongation rate* Λ is defined as the rate at which X_max changes with respect to the primary energy E_0:

$$\Lambda = \frac{\mathrm{d}X_\mathrm{max}}{\mathrm{dlog}_{10} E_0}.$$

Inserting the concrete values for this setting, one obtains:

$$\Lambda_\gamma = \lambda_\mathrm{r} \ln 10.$$

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 proportionally 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 is shown, for example, in figure 3.1). In practice, this model for showers initiated by electrons is 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 in the next section.

3.4.2. Proton initiated showers

A semi-empirical model of hadronic showers was presented by Matthews in 2001 [64, 65] (a schematic representation is shown in figure 3.5). The case for primary protons has some similarities with electron induced showers. The interacting primary proton is assumed to generate a set of charged and neutral pions that traverse a splitting length $X_\mathrm{I} = \lambda_\mathrm{I} \ln 2$, where λ_I is the *interaction length* of strongly interacting particles (in [1], it is discussed that it makes more sense to take $X_\mathrm{I} = \lambda_\mathrm{I}$ without the factor $\ln 2$ because 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 that originate electromagnetic subshowers as the ones 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 the ground. In this model, the values $\lambda_\mathrm{I} = 120\,\mathrm{g\,cm^{-2}}$ and $\xi_\mathrm{c}^\pi = 20\,\mathrm{GeV}$ are used for interactions in air. Besides, multiplicities are reduced to the constant values

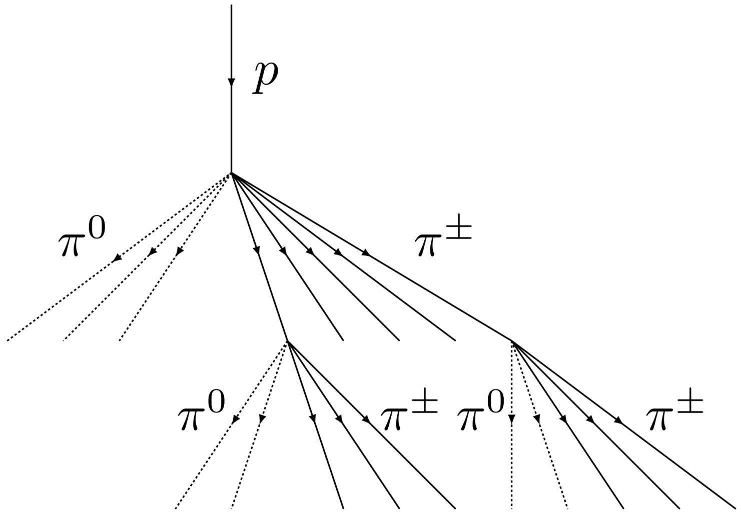

Figure 3.5: Representative scheme of the model of an extensive air shower generated by a proton [64, 65]. Hadronic interactions are simplified and considered to generate only neutral and charged pions. The former decay almost immediately to a pair of photons that subsequently generate electromagnetic 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.

$N_{\text{ch}} = 10$ and $N_0 = 5$, which means that the fraction of charged particles out of the total multiplicity N_{tot} is $f_{\text{ch}} = 2/3$. 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.

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

$$\xi_{\text{c}}^\pi = E_\pi^{\text{ind}}(n_p) = \frac{(f_{\text{ch}})^{n_p}\,E_0}{(N_{\text{ch}})^{n_p}}. \tag{3.4}$$

Here, $(f_{\text{ch}})^{n_p}$ 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 $(N_{\text{ch}})^{n_p}$ particles in the last generation, when all the charged pions decay to muons. As a consequence,

$$n_p = \frac{\ln\left(E_0/\xi_{\text{c}}^\pi\right)}{\ln\left(N_{\text{ch}}/f_{\text{ch}}\right)}$$

and the number of muons at the ground is

$$N_\mu^p = N_\pi(n_p) = (N_{\text{ch}})^{n_p} = \left(\frac{E_0}{\xi_{\text{c}}^\pi}\right)^\beta, \tag{3.5}$$

where

$$\beta = \frac{\ln\left(N_{\text{ch}}\right)}{\ln\left(N_{\text{ch}}/f_{\text{ch}}\right)}. \tag{3.6}$$

This time, the depth X_{max}^p, at which the maximum number of particles is obtained, needs to be calculated differently compared to what is done with the electromagnetic subshowers 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

$$X_{\text{max}}^p = X_0 + \lambda_{\text{r}} \ln\left(\frac{E_0/(3\cdot 2N_0)}{\xi_{\text{c}}^e}\right) = X_0 + \lambda_{\text{r}} \ln\left(\frac{E_0/(3\cdot N_{\text{ch}})}{\xi_{\text{c}}^e}\right), \tag{3.7}$$

where $X_0 = \lambda_{\text{I}}^p(E_0)\ln 2$ 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:

$$E_0 = \xi_{\mathrm{c}}^{e} N_{\mathrm{max}}^{p} + \xi_{\mathrm{c}}^{\pi} N_{\mu}^{p}. \tag{3.8}$$

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

$$\Lambda_p = \Lambda_\gamma + \frac{\mathrm{d}}{\mathrm{d}\log_{10} E_0} \left(X_0 - \lambda_{\mathrm{r}} \ln\left(3N_{\mathrm{ch}}\right)\right). \tag{3.9}$$

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 results obtained from simulations quite well. 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

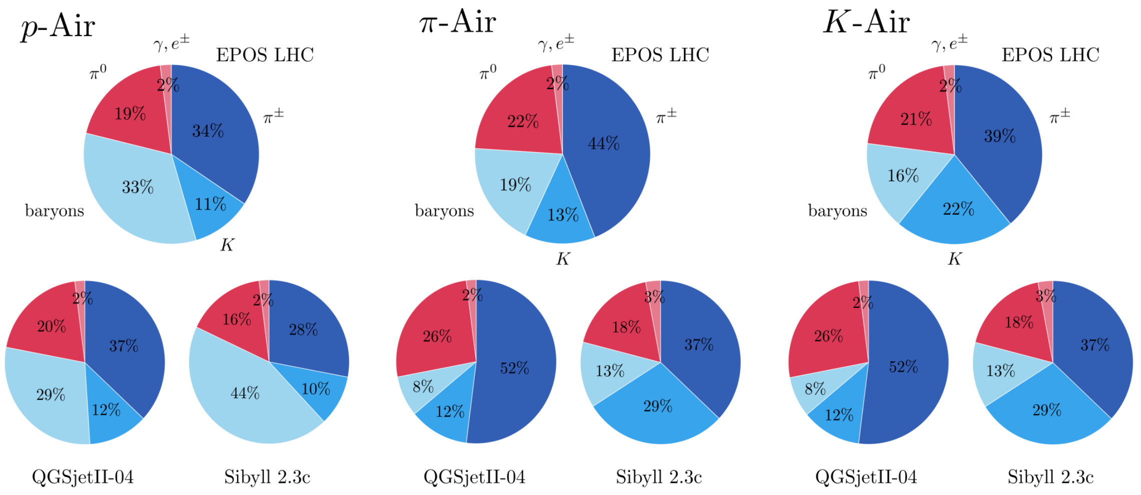

Figure 3.6: Fractions of the principal groups of particles (charged pions, kaons, baryons, neutral pions and electromagnetic particles) created in p-Air, π-Air and K-Air interactions at 10^{19} eV using different high-energy interaction models (EPOS-LHC, QGSJETII-04, SiIBYLL-2.3c, reproduced from [67]). The blue slices represent groups of particles that mainly contribute to the hadronic channel. 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.

[71]. 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 the interpretation of experimental results. It also reflects Linsley's *elongation rate theorem*, which states that the elongation rate for electromagnetic showers is an upper limit to the elongation rate of hadronic showers [94].

We now explore the just presented parameters at energies relevant in the context of the present work, using our sets of simulations. Concerning the values of the effective multiplicities N_{ch} and N_0, it is most senseful to assign them geometric mean values. The geometric mean of the numbers $x_1, x_2, \ldots, x_n$ is

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

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 because 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 [67]). We need to define the multiplicities N_{ch} and N_0, taking into account that there are particles other than pions, a fact that 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 to the electromagnetic component or they keep energy in the hadronic channel.

Eta mesons (η) decay extremely rapidly, mainly to neutral pions or directly to

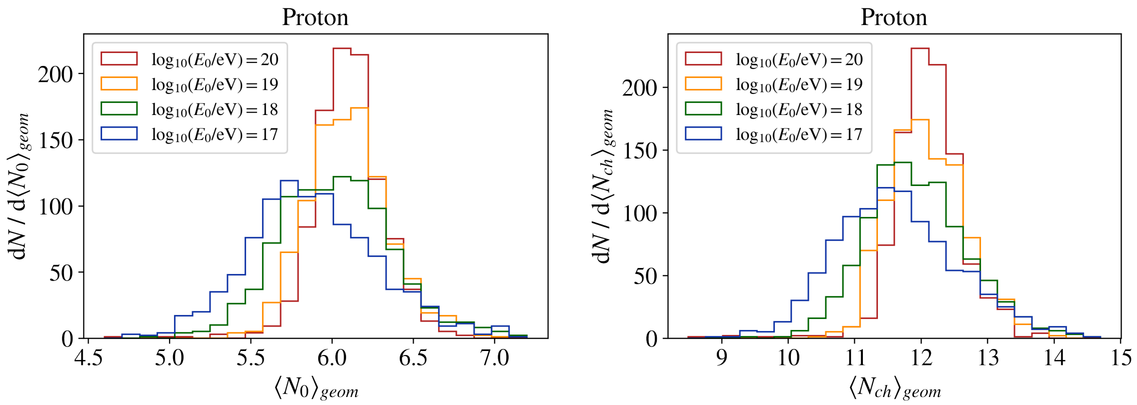

Figure 3.7: Distributions of the geometric mean of the numbers N_0 (left) and N_{ch} (right), as defined in the text, for sets of 1000 showers initiated by vertical protons and simulated using the high-energy interaction model EPOS-LHC and different primary energies.

photons [71]. Therefore, we add them to the number N_0. Kaons behave similarly to pions. The decay lengths for charged ($K^{\pm}$), short-lived (K^0_S) and long-lived kaons (K^0_L) are $c\tau = 3.7\,\mathrm{m}$, $c\tau = 2.7\,\mathrm{cm}$ and $c\tau = 15.3\,\mathrm{m}$, respectively [71], while 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 ($c\tau = 7.9\,\mathrm{cm}$ [71]) 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 mean of the numbers N_0 (left) and N_{ch} (right) are shown for different energies of interest in the present work. In each case, 1000 simulations initiated by vertical protons are considered, where the high-energy interaction model used was EPOS-LHC. The distribution's mean grows with energy (5.89, 6.02, 6.09, 6.07 for $\langle N_0\rangle_{\mathrm{geom}}$ and 11.64, 11.93, 12.08, 12.03 for $\langle N_{\mathrm{ch}}\rangle_{\mathrm{geom}}$), while the peaks get narrower (standard deviations 0.4, 0.37, 0.33, 0.41 for $\langle N_0\rangle_{\mathrm{geom}}$ and 0.94, 0.82, 0.77, 0.94 for $\langle N_{\mathrm{ch}}\rangle_{\mathrm{geom}}$). Consequently, the values are a bit higher than those typically used for the knee region (constant values 5 and 10 are used in [64]). 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 [95]). In addition, it becomes clear that within a 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. This could lead then to very low multiplicities very soon.

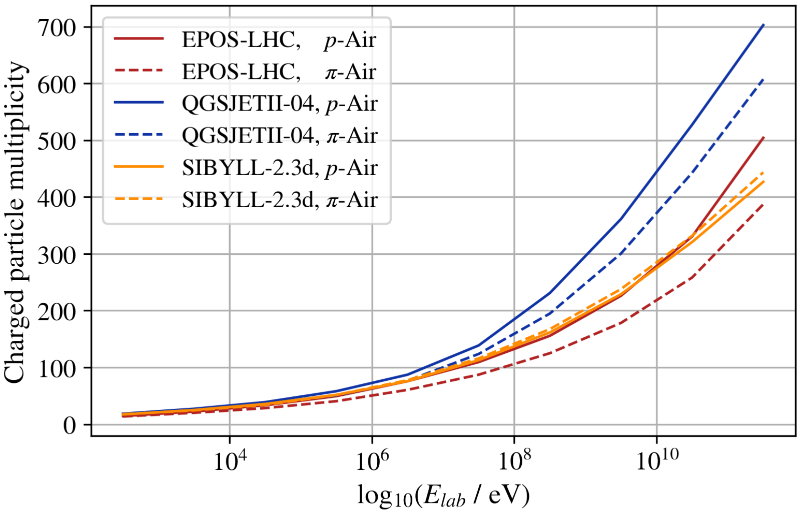

Figure 3.8: Mean charged particle multiplicity in p-air and π-air collisions, as calculated with EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, as a function of energy in the laboratory system (reproduced from [95]). The differences in the predicted secondary particle multiplicities increase with energy. The multiplicity of neutral pions is closely linked to that of charged particles and hence shows qualitatively the same behavior.

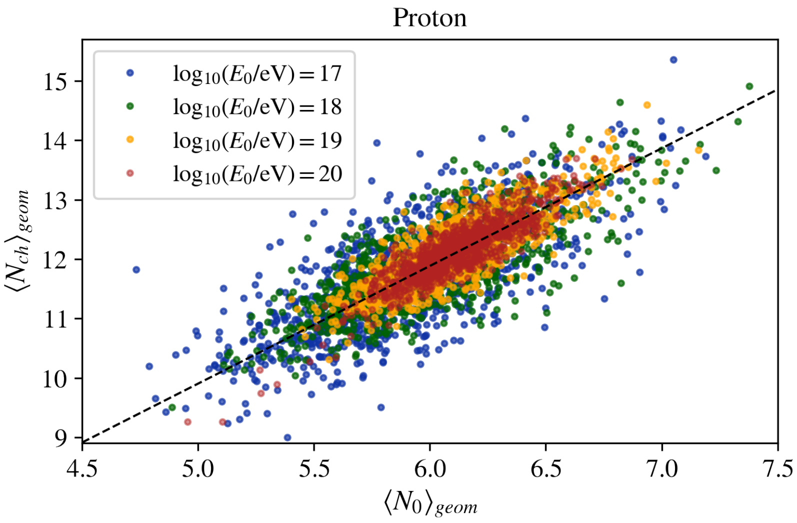

Figure 3.9: Correlation between the geometric means of N_0 and N_{ch} from figure 3.7. The black line represents $y = 1.98x$. Therefore, in certain contexts, it is safe to assume that there are 2 times more hadronically interacting secondaries than secondaries delivering electromagnetic particles.

The correlation between $\langle N_0 \rangle_{\mathrm{geom}}$ and $\langle N_{\mathrm{ch}} \rangle_{\mathrm{geom}}$ for the same sets of simulations is shown in figure 3.9, where the black line represents $y = 1.98x$. Here, the factor $c_0^{\mathrm{ch}} = 1.98$ 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 for 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 predicts that the multiplicity of neutral pions is equal to the average of those of positively and negatively charged pions [96]. 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

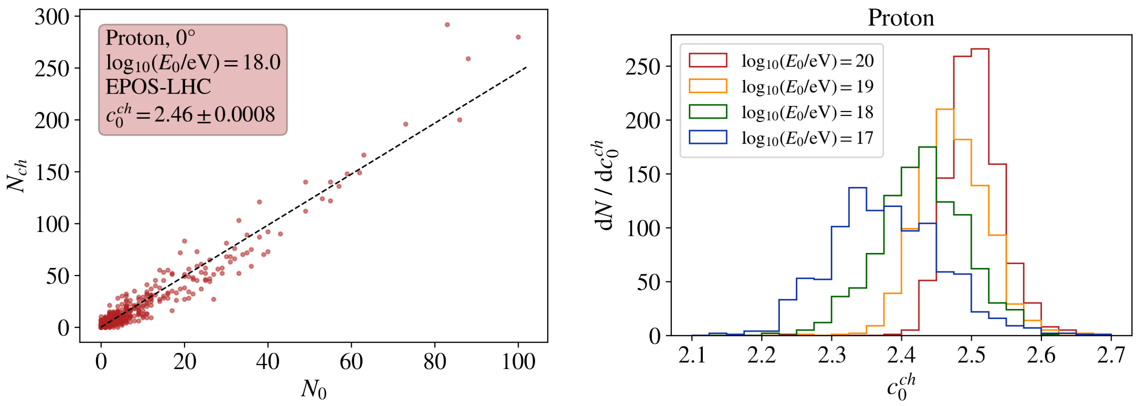

Figure 3.10: Left: Relationship between the multiplicities N_0 and N_{ch} from all the interactions (with energy above $E_{\mathrm{thr}} = 0.005 \times E_0$) present in a typical proton initiated shower of $E_0 = 10^{18}$ eV, simulated using the high-energy interaction model EPOS-LHC. It has, on average,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

the π^0 is the same as the average of the contents of the π^+ and the π^-:

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

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

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 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 $E_{\text{thr}} = 0.005 \times E_0$ for this set of showers), a factor of $c_0^{\text{ch}} = 2.46$ is found, which is close to but higher than 2. 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 [97]. In figure 3.10 (right), the distributions of the factor c_0^{ch}, obtained the same way as for figure 3.10 (left) for the showers used in figure 3.7, are shown. The mean of this factor grows with energy (2.38, 2.43, 2.47, 2.50), as is expected from the behavior of the kaon-to-pion ratio, 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.

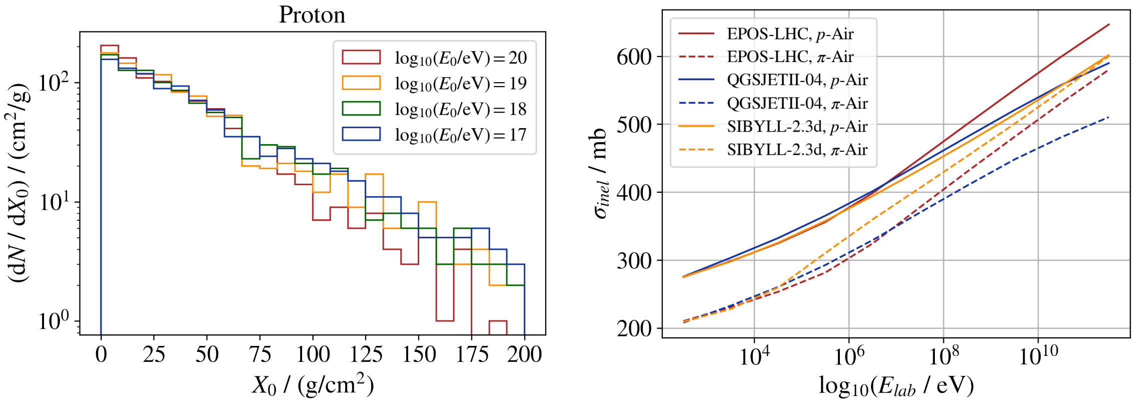

Figure 3.11: Left: Distributions of the depth of the first interaction for vertical showers initiated by protons, simulated using the high-energy interaction model EPOS-LHC and different primary energies. Right: Proton- and pion-air inelastic cross section as a function of energy in the laboratory system (reproduced from [95]). The inelastic cross section is defined as the cross section of all collisions in which at least one new particle is produced. It can be written as $\sigma_{\text{inel}} = \sigma_{\text{tot}} - \sigma_{\text{el}} - \sigma_{\text{qel}}$, where σ_{tot} is the total cross section and σ_{el} and σ_{qel} are the elastic and quasi-elastic cross sections, respectively.

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 $X_0 = \lambda_{\mathrm{I}}^{p}(E_0)$ (figure 3.11, left), one sees that a large range of values is possible for fixed values of primary 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 of 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 [95]). Using that the interaction length λ_{I} is related to the particle inelastic cross-section σ_{inel} via the expression

$$\lambda_{\mathrm{I}} = \frac{\langle m_{\mathrm{air}} \rangle}{\sigma_{\mathrm{inel}}}$$

[98] and that the effective mass number of air is 14.8, we obtain that the interaction length can vary from $\lambda_{\mathrm{I}} = 41\,\mathrm{g\,cm^{-2}}$ for the highest energies to $\lambda_{\mathrm{I}} = 124\,\mathrm{g\,cm^{-2}}$ 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]:

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

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

$$\lambda_{\mathrm{I}}(E) = 200\,\mathrm{g\,cm^{-2}} - 3.3\,\mathrm{g\,cm^{-2}} \cdot \ln{(E/\mathrm{eV})}.$$

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

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

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

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

Then, the depth

$$X_n(E_0) = X_0(E_0) + \sum_{i=1}^{n} \lambda_{\mathrm{I}}(E_i),\ n \geq 1$$

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

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

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 numerically from this expression. The critical energy is then

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

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.

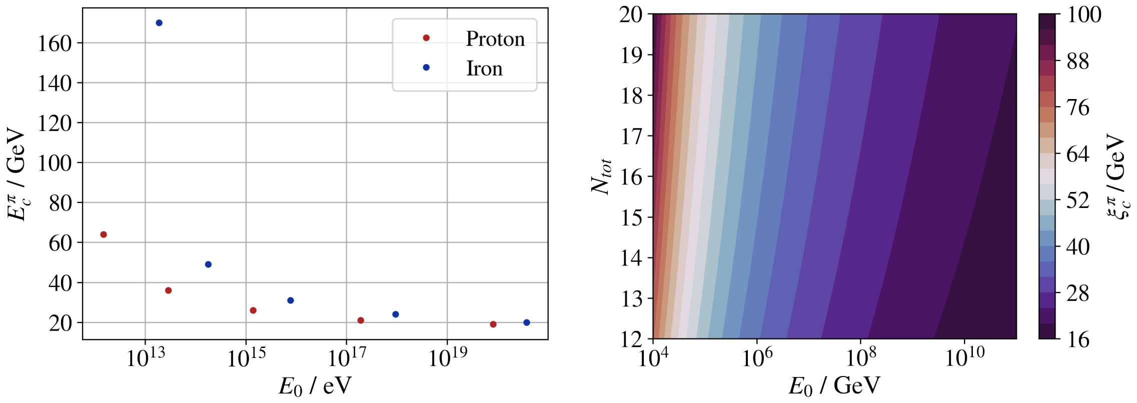

Figure 3.12: Critical energy as a function of primary energy, as calculated in [1] for proton and iron primaries (left), and as a function of primary energy and effective total multiplicity, as calculated in [2] for proton primaries (right). The first approach is more detailed and complex than the second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

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:

$$\lambda_{\mathrm{I}} = \lambda_{\mathrm{dec}}. \tag{3.10}$$

The decay length is

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

where ρ is the height-dependent density of air and

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

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

$$\cos{(\theta)} = \frac{\rho(h)h_{\mathrm{s}}}{X} = \frac{\rho(h)h_{\mathrm{s}}}{n\lambda_{\mathrm{I}}}. \tag{3.12}$$

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

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

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

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

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

$$n_d = -\frac{\mathrm{W}_{-1}\left(-\frac{h_{\mathrm{s}}}{c\tau_{\pi^\pm}}\frac{m_{\pi^\pm}}{E_0}\frac{\ln{(N_{\mathrm{tot}})}}{\cos{(\theta)}}\right)}{\ln{(N_{\mathrm{tot}})}},$$

where W_{-1} denotes the lower branch of the *Lambert-W* function [99].

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 is easy to see 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 $-n_d\ln{(N_{\mathrm{tot}})}e^{-n_d\ln{(N_{\mathrm{tot}})}} \in [-e^{-1}, 0]$ (since $n_d\ln{(N_{\mathrm{tot}})} < e^{n_d\ln{(N_{\mathrm{tot}})}-1} = 1 + (n_d\ln{(N_{\mathrm{tot}})} - 1) + ...$). The critical energy is obtained as before via

$$\xi_{\mathrm{c}}^{\pi} = \frac{E_0}{(N_{\mathrm{tot}})^{n_d}}$$

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 $f_{\mathrm{ch,en}}$ 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 significantly 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 of the critical energy in chapter 4, where a similar approach to that of Kampert and Unger will be followed.

3.4.3. Hadronic showers

When a heavy nucleus enters the atmosphere, it interacts rather quickly. The median of the depth of the first interaction for showers initiated by iron primaries is close to $7\,\mathrm{g\,cm^{-2}}$ for all energy groups (see figure 3.14, for proton showers this value is around $30\,\mathrm{g\,cm^{-2}}$, 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 to nuclear primaries with atomic number $A > 1$, the *superposition model* is applied, where an even distribution of energy is assumed [64, 65]. As a result, the distribution of the depth of the 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

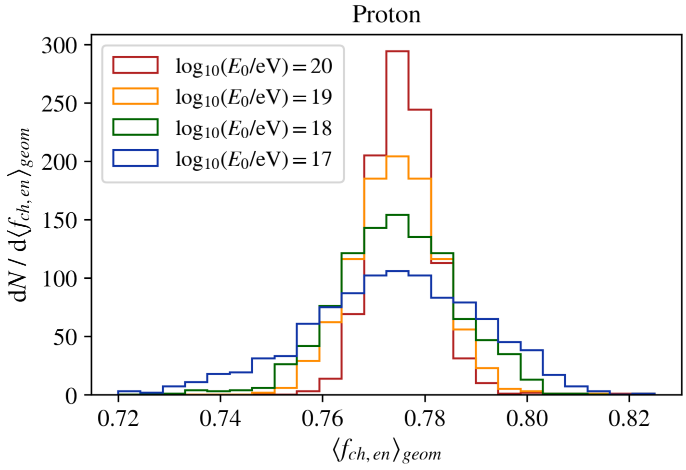

Figure 3.13: Distributions of the geometric mean of the fractions of energy $f_{\mathrm{ch,en}}$ taken by hadronically interacting particles for the different groups of simulations used for figure 3.7. More than two thirds of the available energy is taken by these particles in every energy group. This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

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:

$$X^A_{\max}(E_0) = X^p_{\max}(E_0) - \lambda_\mathrm{r} \ln(A), \tag{3.13}$$

$$N^A_{\max}(E_0) = N^p_{\max}(E_0), \tag{3.14}$$

$$N^A_\mu(E_0) = N^p_\mu(E_0) \cdot A^{1-\beta}, \tag{3.15}$$

$$\Lambda_A(E_0) = \Lambda_p(E_0). \tag{3.16}$$

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 $\langle N_0\rangle_\mathrm{geom}$ and 10.02, 11.04, 11.53, 11.73 for $\langle N_\mathrm{ch}\rangle_\mathrm{geom}$. 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 $(10^{20}\,\mathrm{eV})/56$. An equivalent 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

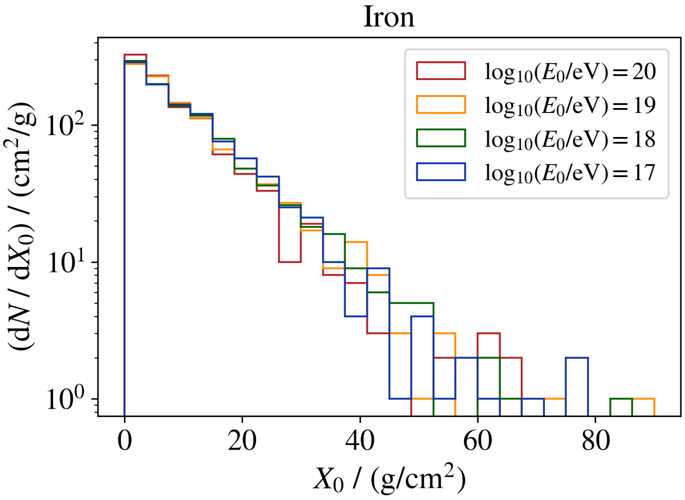

Figure 3.14: Distributions of the depth of the first interaction for sets of 1000 showers initiated by vertical iron primaries and simulated using the high-energy interaction model EPOS-LHC and different primary energies. Even though showers generated by lower energy particles have, on average, a depth of the 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).

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 $\log_{10}(E_0/\mathrm{eV}) = 17, 18, 19, 20$ iron showers, respectively. We suggest median values because of the skewness of the distributions. Also here, the distributions are 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 the first interaction for 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 the 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

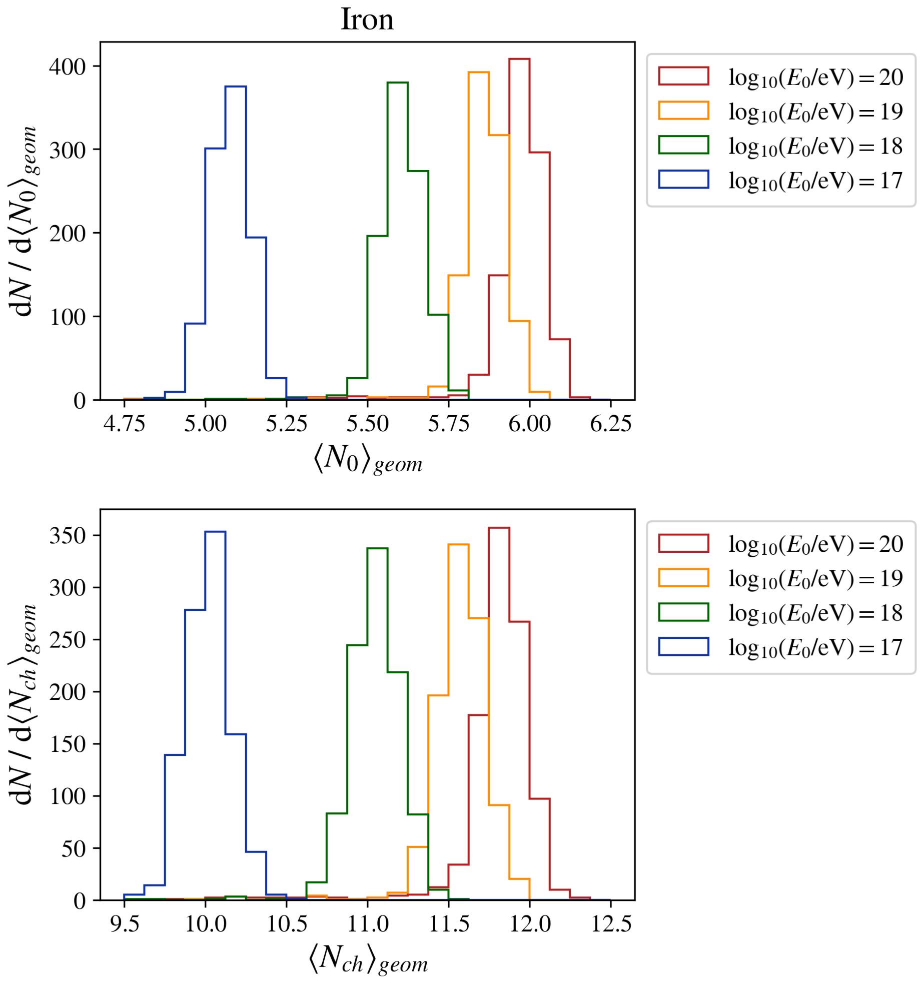

Figure 3.15: Distributions of the geometric means of N_0 (top) and N_{ch} (bottom) for the showers from figure 3.14. These values are calculated the same way as was done for figure 3.7. It is, as for proton primaries, and in certain contexts, safe to assume that in iron showers there are two times more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

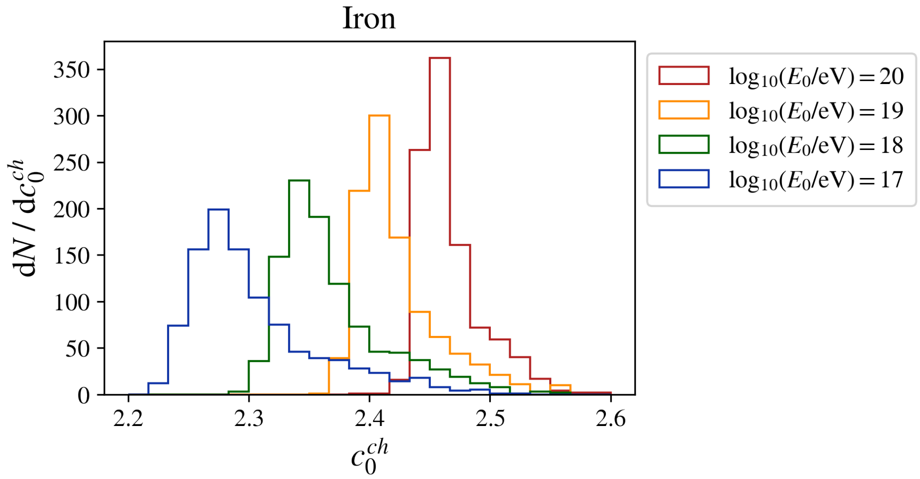

Figure 3.16: Distributions of the factor c_0^{ch} for the different groups of simulations from figure 3.14, as calculated for figure 3.10. As for proton primaries, these values are slightly above 2 and can be used when values representative of individual showers are needed.

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 and read 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 $\lambda_{\mathrm{r}} \ln{(A)}$. On the other hand, the multiplicity of the first interaction is higher for iron primaries than for proton primaries. This can be deduced from the following expression [1]:

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

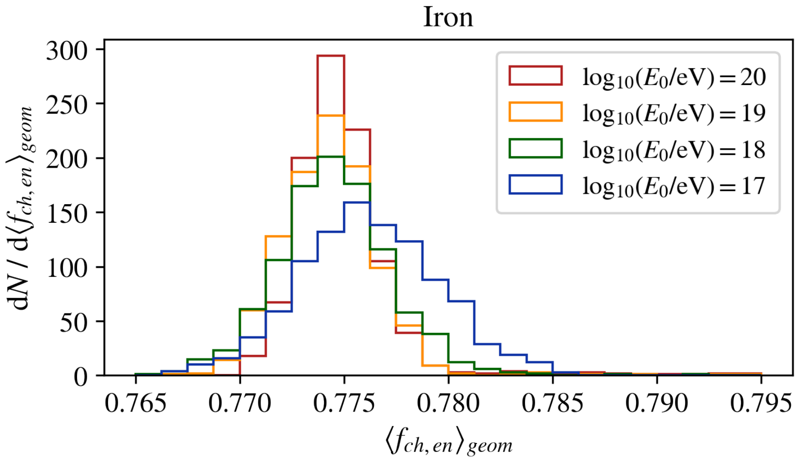

Figure 3.17: Distributions of the geometric mean of the fractions of energy $f_{\mathrm{ch,en}}$ taken by hadronically interacting particles for the simulations from figure 3.14. Also for iron primaries, more than two thirds of the available energy are taken by these particles.

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 values of X_{max} calculated 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 $X^{p}_{\mathrm{max}}(E_0)$. 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 the respective distributions and the bars the standard deviations for the calculated distributions. The agreement is quite good for some combinations of energy and primary mass, 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, while for iron primaries it is 58.8 g cm^{-2} per decade of energy.

From equation (3.6) and since $\ln(f_{\mathrm{ch}}) \leq 0$, we can conclude that $0 < \beta \leq 1$. So, following equation (3.15), the number of muons is higher for higher atomic number and fixed primary energy E_0. This can also be observed in figure 3.19, where the muon numbers at the ground, for our simulated showers and calculated using equation (3.5), are shown. In addition to the individual effective multiplicities, this time, the critical

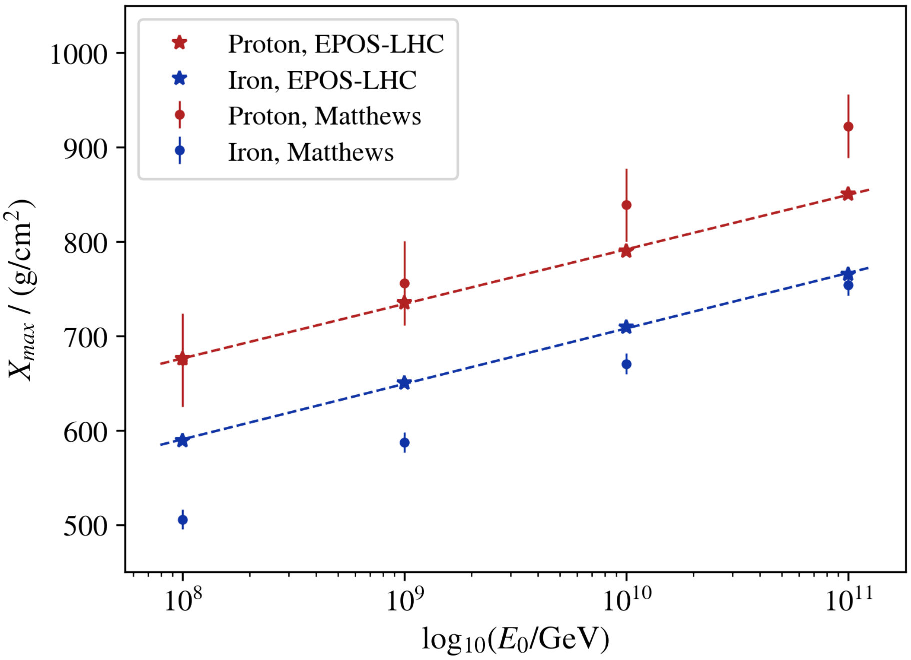

Figure 3.18: Comparison between X_{max} calculated using Matthews' model (filled circles) and its values from our simulations (stars). Sets of 1000 simulations of vertical protons are used. The dashed lines are fitted to the simulated median values. They make the elongation rate, which is very similar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

energy is needed, as well. We calculate it with the method from [2]. Qualitatively, the observed behavior 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 we shall see, 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 employing expression 3.8 are presented together with 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 from simulations and calculated values for different primary types carrying the same energy are very close to one another. However, there is a gap of more than one 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

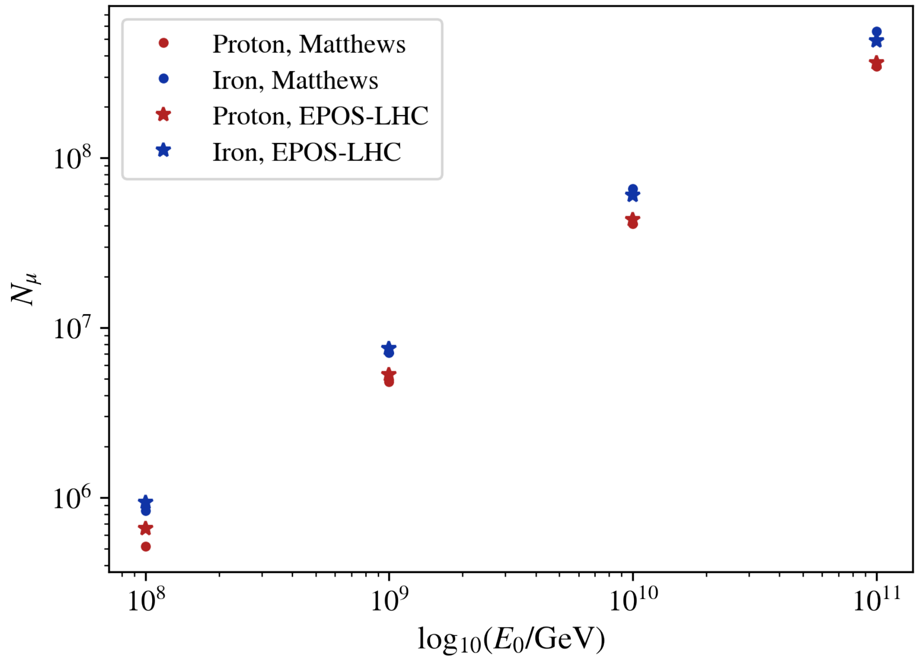

Figure 3.19: Comparison between N_μ calculated using Matthews’ model (filled circles) and its values from our simulations (stars). The same simulations as in figure 3.18 are used. According to calculations, there are $A^{1-\beta}$ times more muons in showers generated by primaries with atomic number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

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. Instead, there are groups of particles that are created by different combinations of leading particles and "generic" ones. The mean elasticities in EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d are summarized in figure 3.21 (reproduced from [95]) for a wide range of energies.

Considering 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 $\left(\frac{1}{3}\kappa E\right)$ among N_0 neutral pions, as a consequence, one gets a slightly modified expression for β:

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

This expression is obtained almost the same way as for (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:

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

Here, $(1+N_{\mathrm{ch}})$ 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. This approximation for

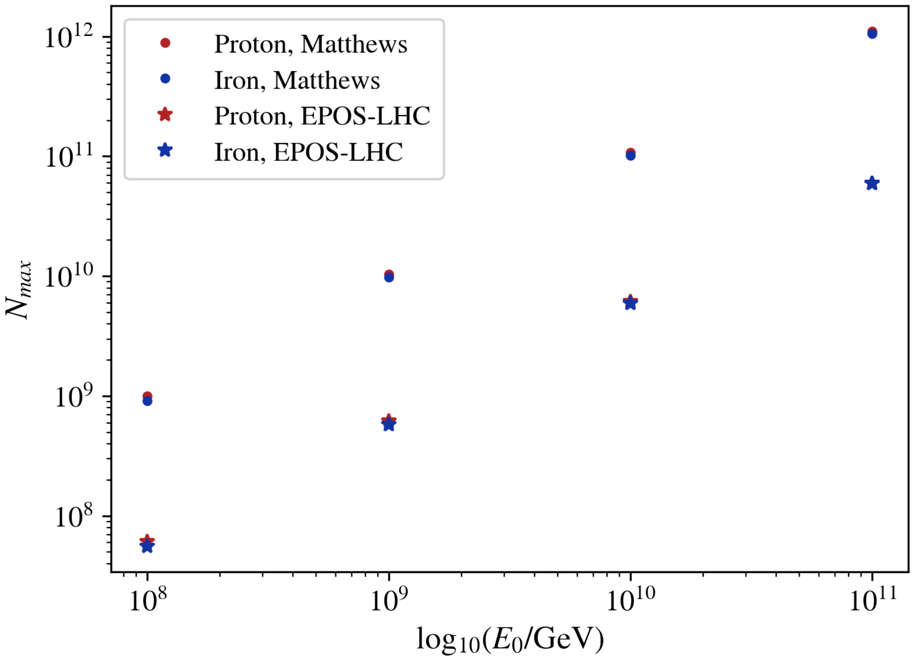

Figure 3.20: Comparison between N_{max} calculated using Matthews' model (filled circles) and its values from our simulations (stars). The same simulations as in figure 3.18 are used. According to calculations, the maximum shower size does not depend on the primary mass.

n_p^{lp} is still adequate because simulations reveal 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 has already been reported in the year 2000 by the HiRes/MIA collaboration [100] 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° [101]. 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 the air showers, lateral distance of the muon density measurement, energy threshold of the muons that

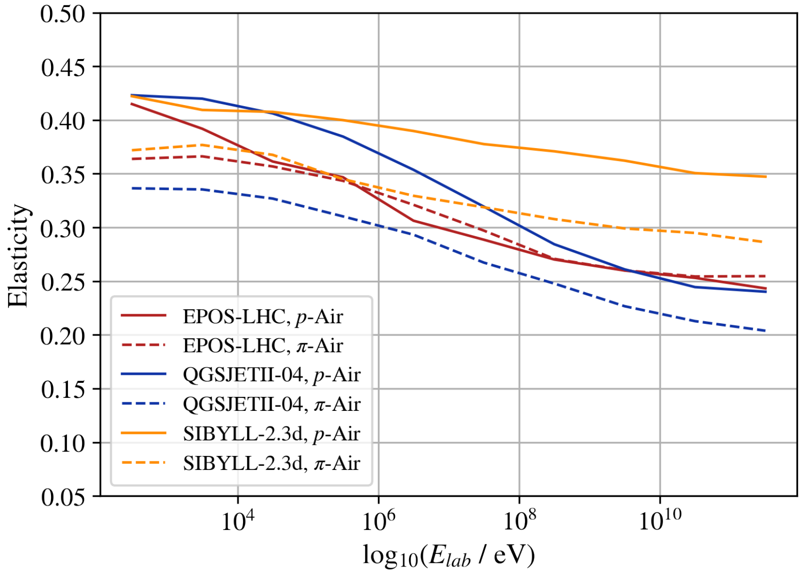

Figure 3.21: Mean elasticity in proton-air and pion-air collisions, as predicted by EPOS-LHC, QGSJETII-04 and SIBYLL-2.3.d (reproduced from [95]).

are measured), the z-scale is introduced [66]:

$$z = \frac{\ln\left(N_{\mu}^{\text{det}}\right) - \ln\left({N_{\mu}}_{p}^{\text{det}}\right)}{\ln\left({N_{\mu}}_{\text{Fe}}^{\text{det}}\right) - \ln\left({N_{\mu}}_{p}^{\text{det}}\right)}.$$

Here, N_{μ}^{det} is the muon density measured by the detector, while ${N_{\mu}}_{p}^{\text{det}}$ and ${N_{\mu}}_{\text{Fe}}^{\text{det}}$ 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

$$z_{\text{mass}} = \frac{\langle \ln(A) \rangle}{\ln(56)} \quad [102].$$

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 $\Delta z = z - z_{\text{mass}}$ is studied in [66]. Here, z_{mass} is computed from the *Global Spline Fit* model [103]. 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 becomes visible. It is noticeable that the results for different experiments are fairly consistent in this behavior.

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 [98]. 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 the inference of $\langle \ln(A) \rangle$ 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 the inferred value of $\langle \ln(A) \rangle$ 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 $\sim 13\,\mathrm{TeV}$, which corresponds to an incident cosmic ray energy of almost $10^{17}\,\mathrm{eV}$. 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

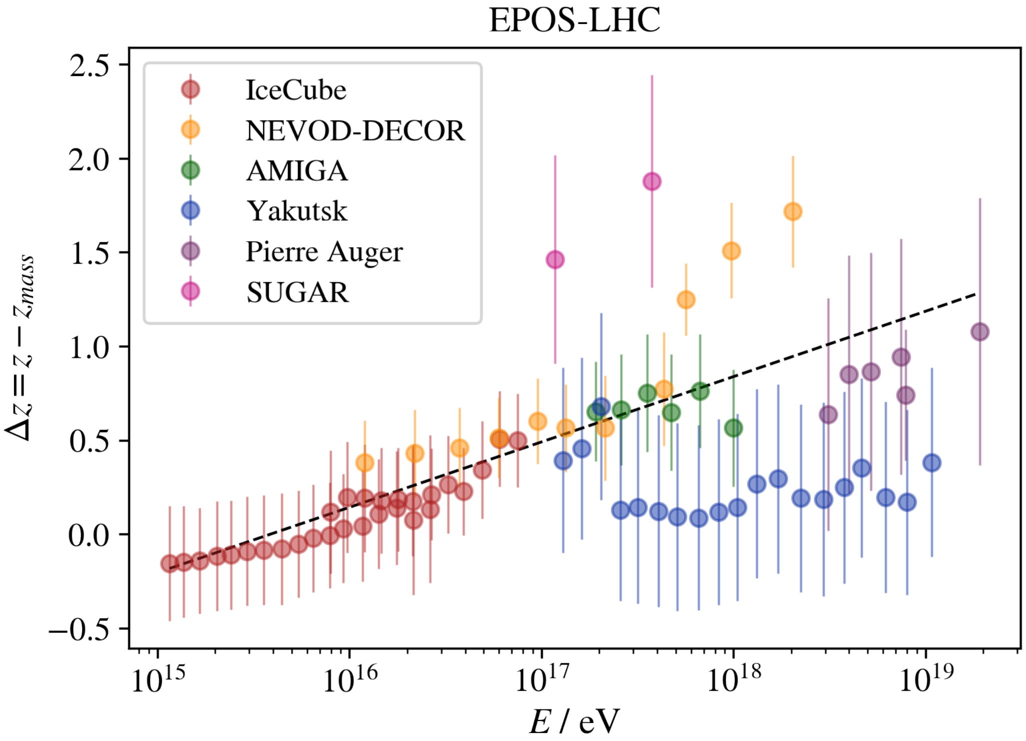

Figure 3.22: $\Delta z = z - z_{\mathrm{mass}}$ for EPOS-LHC and different air shower experiments, after applying cross-calibration (reproduced from [66]). A consistency between measurements and simulations is observed up to about $10^{17}\,\mathrm{eV}$. At higher energies, a growing muon deficit in the simulations is evident.

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 $E_0 = 10^{15}\,\mathrm{eV}$, the most probable energy of the *grandmother* particle of any muon arriving at the ground is within the range of beam energies of fixed experiments [104]: 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 [105], where they show 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 specialized in 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* [106]. In this approach, the nucleus-nucleus scattering amplitude is defined by the sum of the contributions of the diagrams corresponding to multiple scattering processes between parton constituents of the projectile and target nucleons. In SIBYLL-2.3d, this is accomplished by means of the *Semi-superposition Model* [107]. 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 [108]. Here, the primary energy is equally divided among nucleons, while the proton-B cross-section is based on a Glauber calculation [109]. This means that the depth of the 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 [110].

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) $f_{\text{em}} = N_0/N_{\text{tot}}$ turned out to be interesting parameters. Ulrich et al. introduced an *ad-hoc* model for air shower simulations, in which these parametes are modified after each individual interaction, and implemented it in the CONEX framework [112]. 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

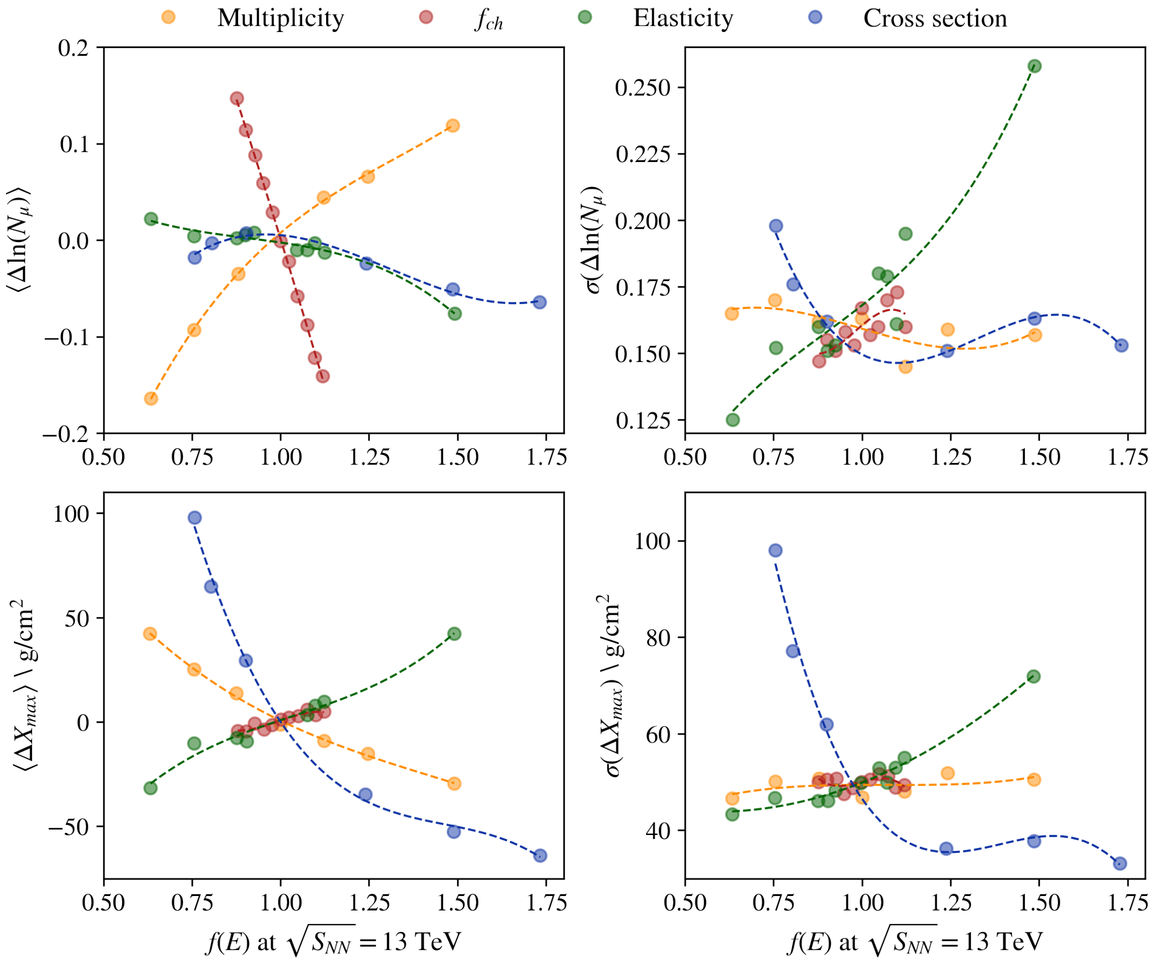

Figure 3.23: Impact of changing individual hadronic interaction features on the means and standard deviations of the logarithm of N_μ (top) and $X_{\max}$ (bottom) (reproduced from [111]). $10^{19.5}$ eV proton showers simulated with CONEX using SIBYLL-2.1 as the baseline model are used. In the left 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 $E = \sqrt{S_{\text{NN}}}$. This factor is extrapolated logarithmically towards higher energies.

is at rest and where

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

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 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 deviation of the muon number N_μ and the depth of shower maximum X_{max} are presented in figure 3.23 (reproduced from [111]). 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. These results reveal 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 is observed concerning the standard deviation of the muon number and using EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d. This constrains 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.

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

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

that goes into the electromagnetic channel on X_{max} and $\ln(N_\mu)$ in full air shower simulations is explored in [113]. The result for 10^{19} eV simulations performed using EPOS-LHC in CONEX is presented in figure 3.24 (reproduced from [113]). 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 $f_{\mathrm{em,en}}$ by up to ±20 % are also shown. These are compared to a measurement done at the Pierre Auger Observatory [101]. Changes in N_{tot} move the line almost horizontally. Thus, no answer regarding the Muon Puzzle can be found here. However, a change in $f_{\mathrm{em,en}}$ 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 [115]. 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 has also been shown in [116] that the forward production of ρ^0 mesons plays an important role in increasing $f_{\mathrm{ch,en}}$. Forward ρ^0 production is namely an alternative to the charge exchange reaction $\pi^- + p \to \pi^0 + n + X$. 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 [117]. As a consequence, more energy goes into hadronic subshowers, leading to more hadronic generations and, in the end, to more muons. Yet, there is a lack of data on both identified hadron spectra and on strangeness production in the forward region.

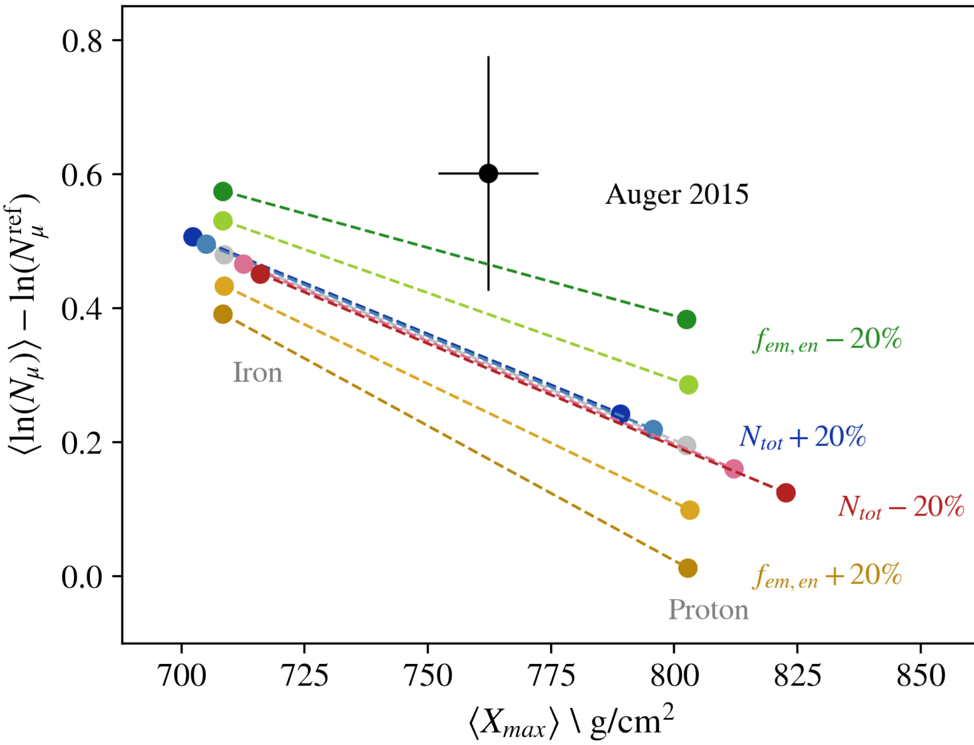

Figure 3.24: Impact of modifying the hadronic multiplicity N_{tot} (dashed lines) and the electromagnetic energy ratio $f_{\mathrm{em,en}}$ (dotted lines) in collisions at 13 TeV on EPOS-LHC predictions of the air shower observables X_{max} and $\ln(N_\mu)$. This result is reproduced from [113]. The datum is from the Pierre Auger Observatory [114]. The model lines represent all values that can be obtained for any combination of nuclei from a pure proton scenario (bottom right) to a pure iron scenario (top left).

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 $f_{\mathrm{em,en}}$ 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 [67, 118, 68, 119, 120]. There are many arguments in favor of a close relationship between the fluctuations of the muon content and the 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):

$$N_\mu = \Pi_{i=1}^{n_p} N_{\mathrm{ch}}^i,$$

where N_{ch}^i is the 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 $\sigma(N_{\mathrm{ch}})$, the dispersions of N_{ch}^i can be calculated as

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

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, while 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 $\sigma(N_{\mathrm{ch}})$ actually decreases with energy [121].

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 originating from the first interaction:

$$N_\mu^1(E_0) = \sum_{i=1}^{N_{\mathrm{ch}}^1} \langle N_\mu(E_i^{\mathrm{had}})\rangle = \sum_{i=1}^{N_{\mathrm{ch}}^1} \left(\frac{E_i^{\mathrm{had}}}{\xi_{\mathrm{c}}^\pi}\right)^\beta, \tag{3.19}$$

where E_i^{had} is the energy of the ith 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 $x_i = E_i^{\mathrm{had}}/E_0$, equation (3.19) can be rewritten as

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

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

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

defined.

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

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

where $E_{ij}^{\mathrm{had}} = E_0 x_i x_{ij}$ 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, is defined. The interactions of the second generation are 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 (eventually) uneven way energy is distributed among particles is equal for all interactions happening in that generation. 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:

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

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 is evident from figure 3.25, where 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) are used. The accumulation of events along the vertical line at $\alpha_1 = 1$ (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 ($\rho_{X,Y} = \mathrm{cov}(X,Y)/(\sigma_X \cdot \sigma_Y)$) is close to 0.8 for all cases.

The relative fluctuations of N_μ can be estimated through the quadratic sum of the

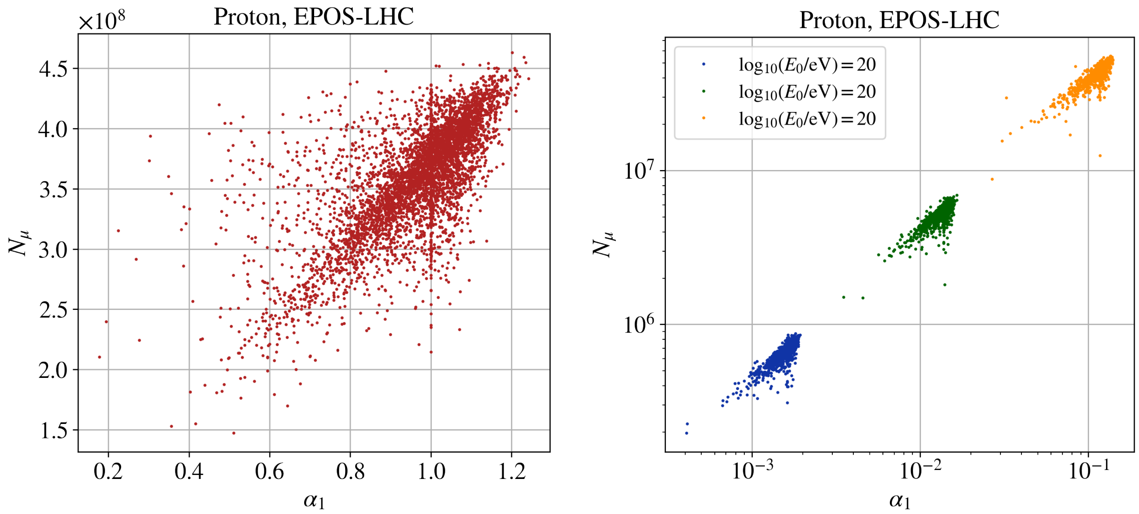

Figure 3.25: Correlation between the number of muons and α_1 for a set of 5000 showers initiated by vertical protons of 10^{20} eV (left) and sets of 1000 showers initiated by vertical protons of lower energies (right). All simulations were performed using the high-energy interaction model EPOS-LHC.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

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 there are. As a consequence, $\sigma(\alpha_k)$ becomes smaller as k increases. This means that the strongest 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_μ 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 that depends on a small set of parameters that 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:

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

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 $\xi_c^\pi = 20\,\text{GeV}$, which we discussed in section 3.4.2. E_i^{had} is the energy of the ith particle that 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 behavior. One important comment to make here is that it is once again confirmed (like in the works from Cazon et al., [67, 118, 68, 119, 120]) 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

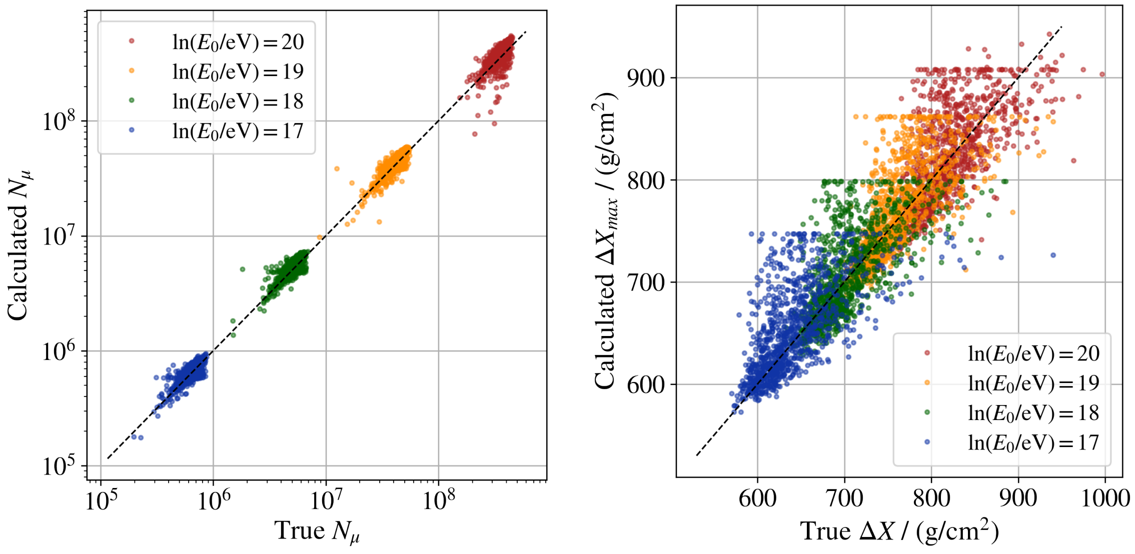

Figure 4.1: Comparison between the true number of muons at the ground (left) and the true values of $\Delta X = X_{\text{max}} - X_0$ (right) with the corresponding calculated values. The *true* values are extracted from showers simulated with CONEX using EPOS LHC. Vertical protons of different energies 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 $E_j^{\mathrm{em}}/2$. E_j^{em} is the energy of the jth π^0 or η. For all the other secondaries, we also use expression (3.3) for the calculation of $X_{\max}$, with the exception that E_0 is replaced by $E_i^{\mathrm{had}}/(2 \cdot N_{\mathrm{tot}}^*)$. Here, N_{tot}^* is 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 and 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 keeps a considerable amount of energy in many successive interactions, 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 distribution 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 arise from quasi-elastic events, where the information useful for this simple model actually would be found in the second interaction.

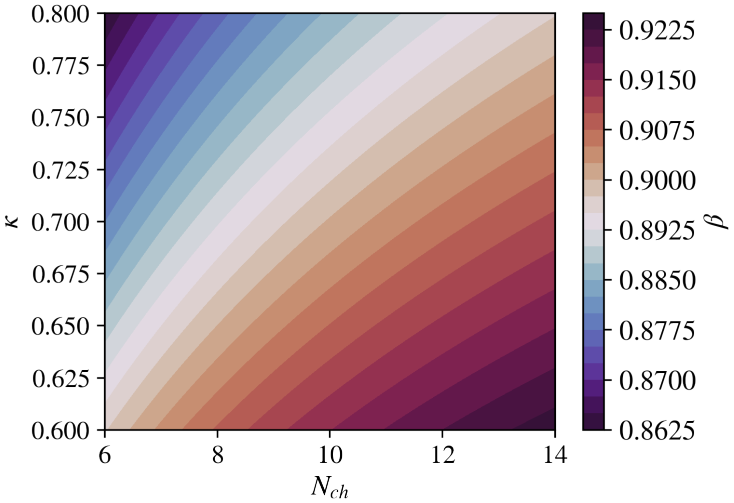

Figure 4.2: β as a function of the multiplicity N_{ch} of hadronically interacting particles and the inelasticity κ. Expression (3.17) is used, which involves the leading particle effect.

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 the energy and identity of each particle being created there, in order to assign it the correct role in the model. It is crucial 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 ΔX-N_μ 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^π_{c}, as calculated by Kampert and Unger, which only depends on the total multiplicity N_{tot}. We calculate $N_\mu = (E_0/\xi^\pi_{\text{c}})^\beta$ with the corresponding expressions from chapter 3 for proton showers of different energies. For the parameters needed, we insert realistic values spanning intervals of the effective multiplicity N_{tot} and the effective inelasticity κ, motivated by the corresponding simulated distributions shown in figures 3.7 and 4.3 (left). This means that we use realistic values of $\langle N_0\rangle_{\text{geom}} + \langle N_{\text{ch}}\rangle_{\text{geom}}$ 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

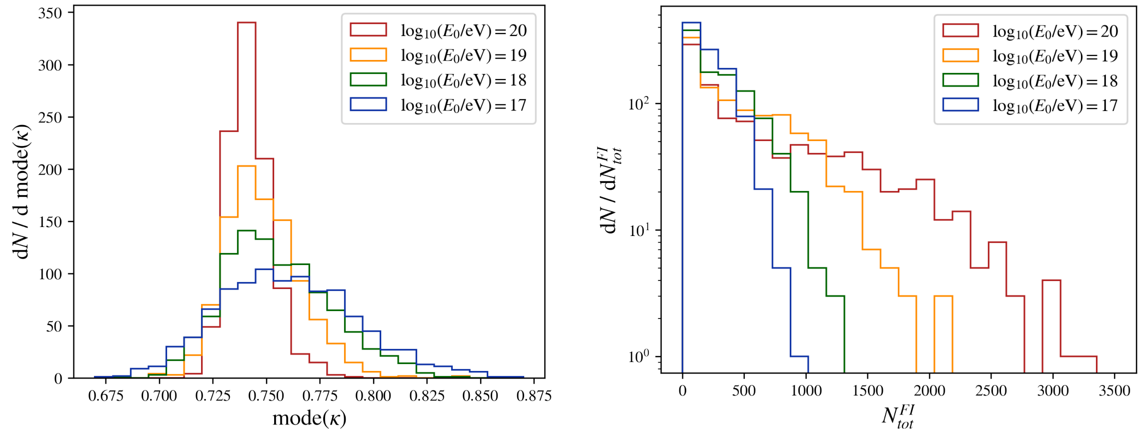

Figure 4.3: Distributions of the mode of the inelasticity κ (left) and the total multiplicity $N^{\text{FI}}_{\text{tot}}$ of the first interaction (right) for sets of 1000 simulations done with CONEX using EPOS-LHC with vertical protons of different energies as primaries.

representative value for the 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 mode(κ) to replace the effective inelasticity. The distribution of these values is shown for showers of different primary energy in figure 4.3 (left).

We calculate $\Delta X = X_{\mathrm{max}} - X_0$ 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 treat them as if they were all equal) with the following expression:

$$\Delta X = \lambda_{\mathrm{r}} \ln\left(\frac{(1-\tilde{f}^{\mathrm{FI}}_{\mathrm{ch,en}})\,\kappa^{\mathrm{FI}} E_0}{2\,(N^{\mathrm{FI}}_{\mathrm{tot}}/3)\,\xi^{e}_{\mathrm{c}}}\right), \tag{4.1}$$

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 $N^{\mathrm{FI}}_{\mathrm{tot}}$ are the inelasticity and the total multiplicity of the first interaction, respectively. $\tilde{f}_{\mathrm{ch,en}}$ is the fraction of the hadronic energy out of the energy that is available for production of new particles, in contrast to $f_{\mathrm{ch,en}}$, which was the same fraction but out of the total energy of the interaction. Ranges of realistic values for $N^{\mathrm{FI}}_{\mathrm{tot}}$, κ^{FI}, $\tilde{f}^{FI}_{\mathrm{ch,en}}$ are taken from figures 4.3 (right) and 4.4 (left and right). For $N^{\mathrm{FI}}_{\mathrm{tot}}$, we count explicitly how many particles are created in the first interaction of

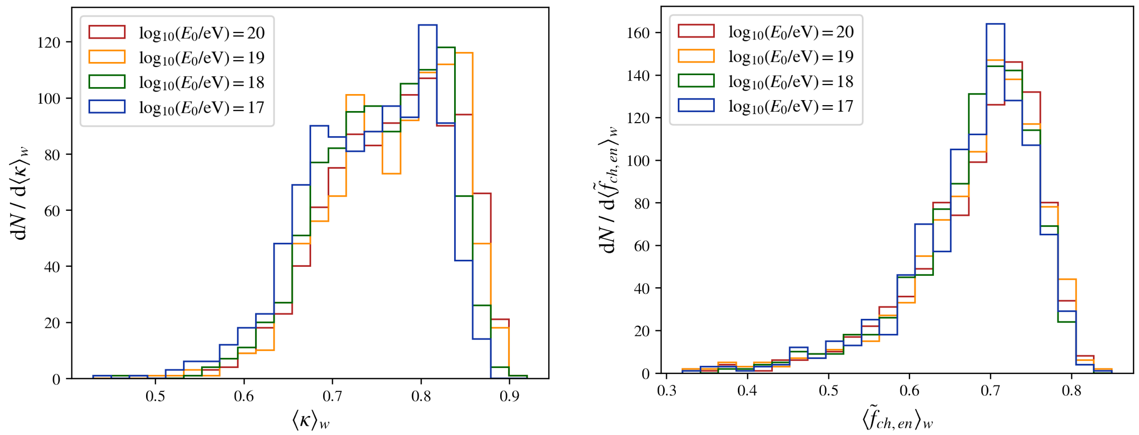

Figure 4.4: Distributions of the inelasticity $\langle\kappa\rangle_{\mathrm{w}}$ (left) and the hadronic energy fraction of energy $\langle\tilde{f}_{\mathrm{ch,en}}\rangle_{\mathrm{w}}$ (right). These weighted averages are computed for each shower taking into account all the interactions that occur (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 of simulations used to obtain the plots in figure 4.3.

each shower, disregarding electrons, photons and muons. They, most probably, will not influence the development of the shower, considering that they typically take a negligible amount of energy. We could 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 $\langle\cdot\rangle_\mathrm{w}$. 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 at 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 $N_\mathrm{tot}^\mathrm{FI}$, $\langle\kappa\rangle_\mathrm{w}$ and $\langle\tilde{f}_\mathrm{ch,en}\rangle_\mathrm{w}$ to be inserted in $N_\mathrm{tot}^\mathrm{FI}$, κ^FI and $\tilde{f}_\mathrm{ch,en}^\mathrm{FI}$, respectively. Note that, since the highest weights w happen in the first few interactions, 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 can argue that this component is fed by neutral pions, which arise from the hadronic component. This motivates probing

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

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 $\langle\kappa\rangle_\mathrm{w}$ and $\langle N_0\rangle_\mathrm{geom}+\langle N_\mathrm{ch}\rangle_\mathrm{geom}$, respectively. The interaction length $\lambda_\mathrm{I} = 120\,\mathrm{g\,cm^{-2}}$ is used, as in Matthews' work. We only vary the other parameters because the effect of varying a multiplicative factor is easy to interpret.

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

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

The concrete ranges we used to test the expressions from Matthews, Kampert and Unger are summarized in table 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 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

$$\frac{\rho(h)h_\mathrm{s}}{n\cos(\theta)} = \lambda_\mathrm{I} = \lambda_\mathrm{dec} = \rho(h)c\tau_{\pi^\pm}\gamma, \tag{4.2}$$

with the difference being in how

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

is calculated. E_π^ind is the energy of individual pions in the n-th generation. n_g will be 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 $\tilde{f}_\mathrm{ch,en}$, the effective inelasticity κ and the corresponding values of the first interaction $N_\mathrm{ch}^\mathrm{FI}$, $\tilde{f}_\mathrm{ch,en}^\mathrm{FI}$ and κ^FI, instead of only the

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 values are compared to the corresponding ranges obtained from simulations done with CONEX using EPOS-LHC and vertical protons of different energies as primaries. N_μ is calculated combining the approaches from Matthews, Kampert and Unger (M., K., U.). ΔX is calculated in two ways, using Matthews' approach (M.) and using Kampert and Ungers' approach (K., U.).

total multiplicity N_{tot}:

$$\xi_{\mathrm{c}}^{\pi} = \frac{\left(1-(1-\tilde{f}^{\mathrm{FI}}_{\mathrm{ch,en}})\kappa^{\mathrm{FI}}\right)\left(1-(1-\tilde{f}_{\mathrm{ch,en}})\kappa\right)^{n_{\mathrm{g}}-1} E_0}{\left(1+N^{\mathrm{FI}}_{\mathrm{ch}}\right)\left(1+N_{\mathrm{ch}}\right)^{n_{\mathrm{g}}-1}}. \tag{4.3}$$

$\left(1-(1-\tilde{f}^{\mathrm{FI}}_{\mathrm{ch,en}})\kappa^{\mathrm{FI}}\right)\left(1-(1-\tilde{f}_{\mathrm{ch,en}})\kappa\right)^{n_{\mathrm{g}}-1}$ is the fraction of the primary energy that stays in the hadronic channel after n_{g} interactions, including the leading particle. $(1-\tilde{f}_{\mathrm{ch,en}})\kappa$ is namely the energy that is "lost" to the electromagnetic component. This way, we describe 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

$$N_\mu = \left(1+N^{\mathrm{FI}}_{\mathrm{ch}}\right)\left(1+N_{\mathrm{ch}}\right)^{n_{\mathrm{g}}-1} \tag{4.4}$$

muons that reach the ground. The first interaction has been separated from the consecutive ones, because the multiplicity $N^{\mathrm{FI}}_{\mathrm{ch}}$ 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 $\left(1-(1-\tilde{f}^{\mathrm{FI}}_{\mathrm{ch,en}})\kappa^{\mathrm{FI}}\right)$ 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 energy fraction can 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

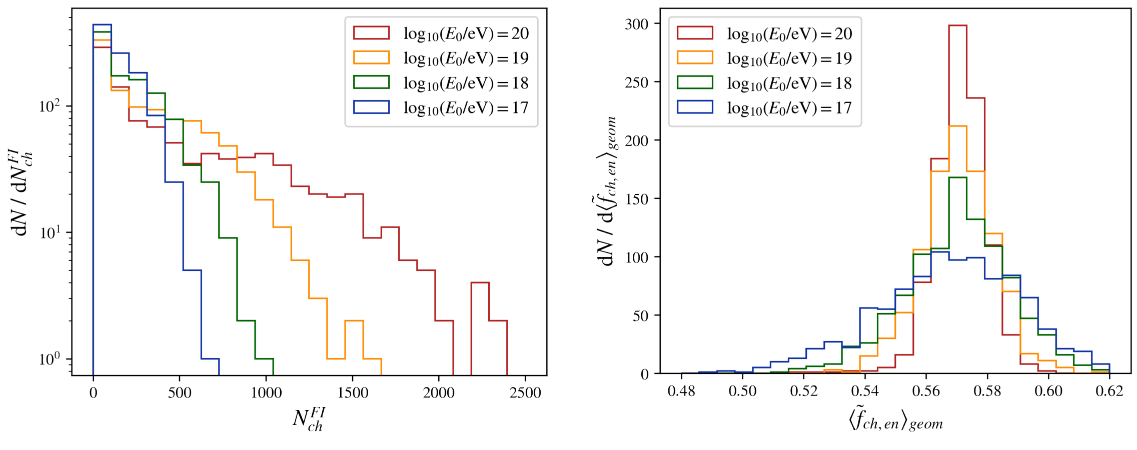

Figure 4.5: Distributions of the charged multiplicity $N^{\mathrm{FI}}_{\mathrm{ch}}$ of the first interaction (left) and of the geometric mean $\langle \tilde{f}_{\mathrm{ch,en}} \rangle_{\mathrm{geom}}$ of the hadronic energy fraction (right) for the sets of simulations used in figure 4.3.

(4.2), we obtain the following:

$$
\begin{aligned}
\frac{\rho(h)h_\mathrm{s}}{n_\mathrm{g}\cos(\theta)} &= \frac{\rho(h)c\tau_{\pi^\pm}\xi_\mathrm{c}^\pi}{m_{\pi^\pm}c^2} \\
\Rightarrow \frac{h_\mathrm{s}}{n_\mathrm{g}\cos(\theta)} &= \frac{c\tau_{\pi^\pm}}{m_{\pi^\pm}c^2}\frac{\left(1-(1-\tilde{f}^\mathrm{FI}_\mathrm{ch,en})\kappa^\mathrm{FI}\right)\left(1-(1-\tilde{f}_\mathrm{ch,en})\kappa\right)^{n_\mathrm{g}-1}E_0}{(1+N^\mathrm{FI}_\mathrm{ch})\,(1+N_\mathrm{ch})^{n_\mathrm{g}-1}} \\
\Rightarrow \frac{h_\mathrm{s}}{\cos(\theta)}\frac{m_{\pi^\pm}c^2}{c\tau_{\pi^\pm}E_0}\frac{1+N^\mathrm{FI}_\mathrm{ch}}{1+N_\mathrm{ch}}\frac{1-(1-\tilde{f}_\mathrm{ch,en})\kappa}{1-(1-\tilde{f}^\mathrm{FI}_\mathrm{ch,en})\kappa^\mathrm{FI}} &= n_\mathrm{g}\left(\frac{1-(1-\tilde{f}_\mathrm{ch,en})\kappa}{1+N_\mathrm{ch}}\right)^{n_\mathrm{g}} \\
&= n_\mathrm{g}\exp\left(n_\mathrm{g}\ln\left(\frac{1-(1-\tilde{f}_\mathrm{ch,en})\kappa}{1+N_\mathrm{ch}}\right)\right) \\
\Rightarrow n_\mathrm{g} = W_{-1}\left(\ln\left(\frac{1-(1-\tilde{f}_\mathrm{ch,en})\kappa}{1+N_\mathrm{ch}}\right)\cdot\frac{h_\mathrm{s}}{\cos(\theta)}\frac{m_{\pi^\pm}c^2}{c\tau_{\pi^\pm}E_0}\frac{1+N^\mathrm{FI}_\mathrm{ch}}{1+N_\mathrm{ch}}\frac{1-(1-\tilde{f}_\mathrm{ch,en})\kappa}{1-(1-\tilde{f}^\mathrm{FI}_\mathrm{ch,en})\kappa^\mathrm{FI}}\right)& \\
\Big/\ln\left(\frac{1-(1-\tilde{f}_\mathrm{ch,en})\kappa}{1+N_\mathrm{ch}}\right). & \qquad (4.5)
\end{aligned}
$$

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 typical values for 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^FI, while allowing for a variety of values of $N^\mathrm{FI}_\mathrm{ch}$ and $\tilde{f}^\mathrm{FI}_\mathrm{ch,en}$. In order to understand the effect of the parameters of the first interaction $N^\mathrm{FI}_\mathrm{ch}$, $\tilde{f}^\mathrm{FI}_\mathrm{ch,en}$ and κ^FI on ΔX and N_μ, we show in the first row of figure 4.6 the result of varying $N^\mathrm{FI}_\mathrm{ch}$ and $\tilde{f}^\mathrm{FI}_\mathrm{ch,en}$ for the extreme upper value $\kappa^\mathrm{FI} = 0.88$, while the result for the extreme lower value $\kappa^\mathrm{FI} = 0.62$ is shown in the second row. For this, we need the distribution of $N^\mathrm{FI}_\mathrm{ch}$, which is shown in figure 4.5 (left).

$\langle N_\mathrm{ch}\rangle_\mathrm{geom}$ and $\langle\kappa\rangle_\mathrm{w}$ 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 $\tilde{f}_\mathrm{ch,en}$, for which we use the geometric mean $\langle\tilde{f}_\mathrm{ch,en}\rangle_\mathrm{geom}$, whose distribution is shown in figure 4.5 (right). The product of the $\tilde{f}_\mathrm{ch,en}$ values from 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 $\tilde{f}_\mathrm{ch,en}$, 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 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 $N_{\mathrm{ch}}^{\mathrm{FI}}$ (left column) and $\langle \tilde{f}_{\mathrm{ch,en}} \rangle_{\mathrm{w}}$ (right column) in the ΔX-N_μ plane, we can observe that the true distribution from simulations is fairly well reproduced (see figure 4.6, bottom, and figure 4.7, top, for comparison). The higher

Parameter	Replacement	Values	New name	Short description
$N_{\mathrm{ch}}^{\mathrm{FI}}$	$N_{\mathrm{ch}}^{\mathrm{FI}}$	$1-2300$	$M_0 = \ln(N_{\mathrm{ch}}^{\mathrm{FI}}+1)$	Hadronic multiplicity of the first interaction
N_{ch}	$\langle N_{\mathrm{ch}} \rangle_{\mathrm{geom}}$	12.03	M_{eff}	Effective multiplicity of the rest of the shower
κ^{FI}	$\langle \kappa \rangle_{\mathrm{w}}$	$0.62, 0.88$	K_0	Inelasticity, mainly representing the first interaction
κ	mode(κ)	0.75	K_{eff}	Effective inelasticity of the rest of the shower
$\tilde{f}_{\mathrm{ch,en}}^{\mathrm{FI}}$	$\langle \tilde{f}_{\mathrm{ch,en}} \rangle_{\mathrm{w}}$	$0.3-0.8$	F_0	Hadronic energy fraction, mainly representative of the first interaction
$\tilde{f}_{\mathrm{ch,en}}$	$\langle \tilde{f}_{\mathrm{ch,en}} \rangle_{\mathrm{geom}}$	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 with expressions (4.1), (4.4) and (4.5) that depend on the six parameters $N_{\mathrm{ch}}^{\mathrm{FI}}$, N_{ch}, κ^{FI}, κ, $\tilde{f}_{\mathrm{ch,en}}^{\mathrm{FI}}$ and $\tilde{f}_{\mathrm{ch,en}}$. In this table, we summarize how we calculate these parameters (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 $N_{\mathrm{ch}}^{\mathrm{FI}}$ and $\tilde{f}_{\mathrm{ch,en}}^{\mathrm{FI}}$, 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 $\ln(N_{\mathrm{ch}}^{\mathrm{FI}}+1)$, 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 multiplicity $N_{\mathrm{ch}}^{\mathrm{FI}}$ (see left column of the first two rows in figure 4.6), 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

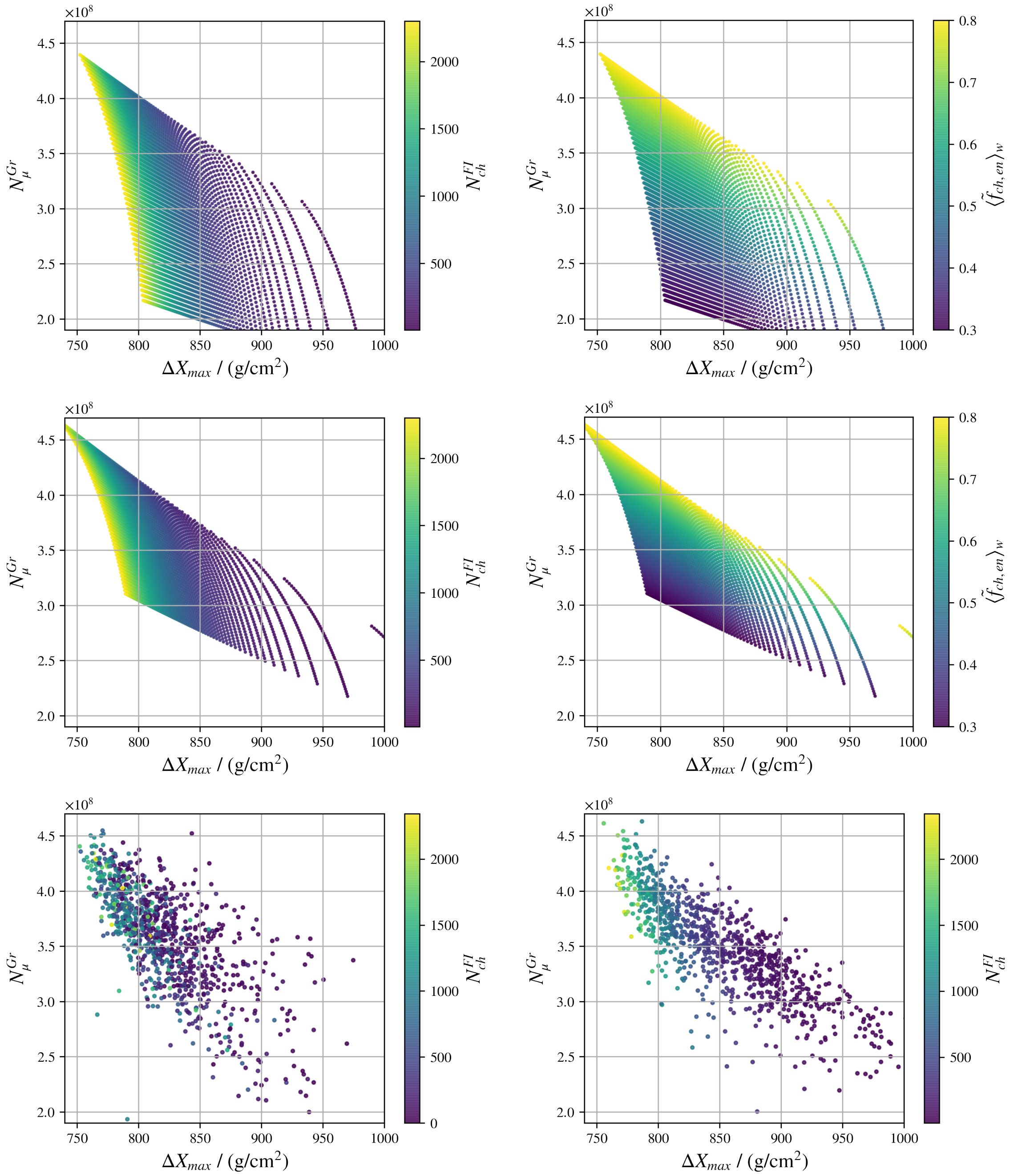

Figure 4.6: The first and second rows show the ΔX-N_μ distributions, as calculated using our model given by expressions (4.1), (4.4) and (4.5) and inserting the values summarized in table 4.3 into the parameters. The first row is calculated with a fixed value of $\langle\kappa\rangle_{\mathrm{w}} = 0.88$,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

of muons. This is also and more efficiently achieved by a higher fraction $\langle \tilde{f}_{\text{ch,en}} \rangle_{\text{w}}$ (see right column of the first two rows in figure 4.6).

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 into equations (4.1), (4.4) and (4.5). The same corrections as before are applied to the pre-

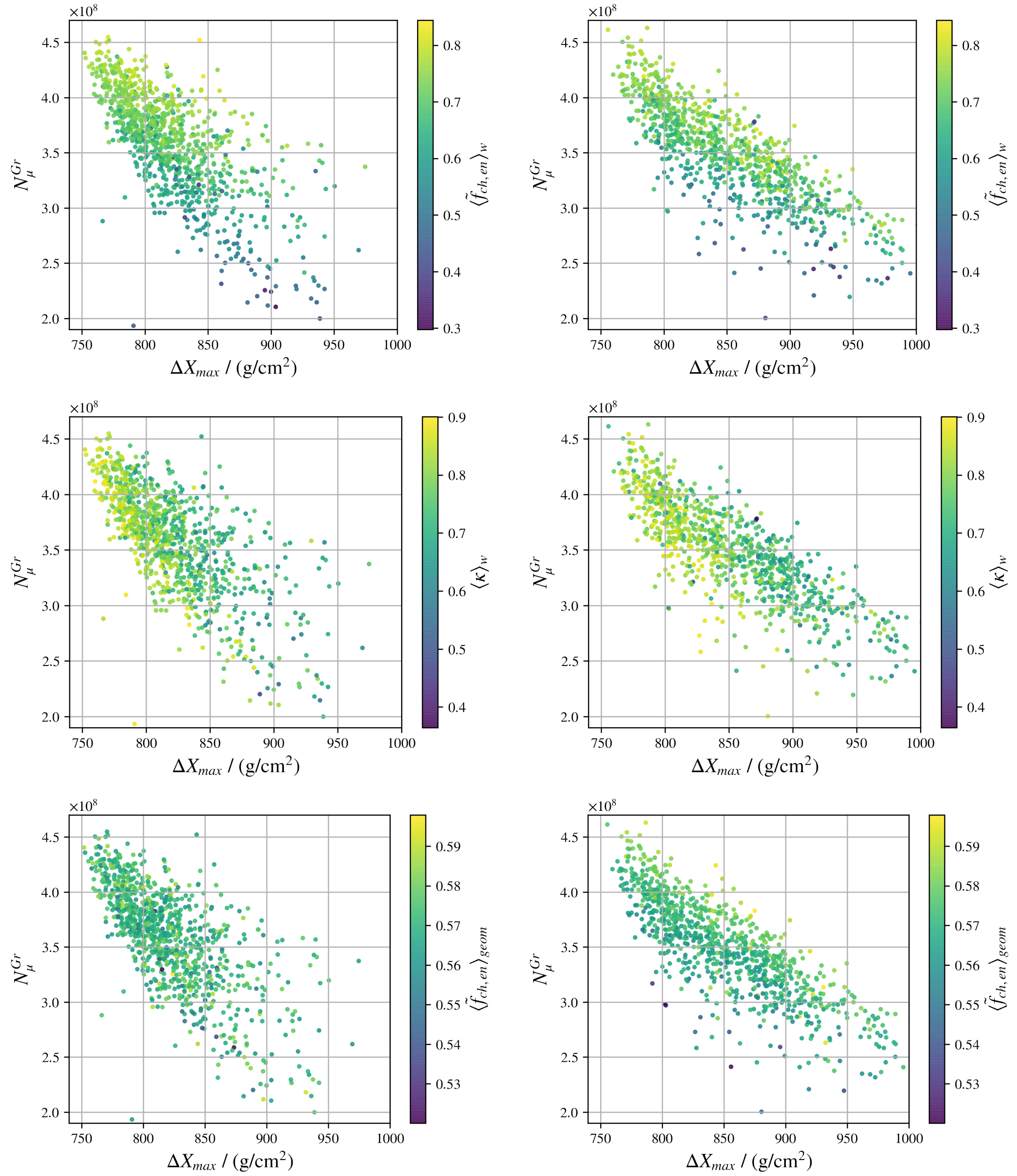

Figure 4.7: As in the last row of figure 4.6, we compare the behavior of the parameters $\langle \tilde{f}_{\text{ch,en}} \rangle_{\text{w}}$, $\langle \kappa \rangle_{\text{w}}$ and $\langle \tilde{f}_{\text{ch,en}} \rangle_{\text{geom}}$ on the distributions from simulations (left) with that on the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

dictions of ΔX and N_μ. The only difference to the previous analysis is that we also insert the individual values of $\langle \tilde{f}_{\mathrm{ch,en}} \rangle_{\mathrm{geom}}$, while for the other two effective parameters mean values are used. Out of the three effective parameters, $\langle \tilde{f}_{\mathrm{ch,en}} \rangle_{\mathrm{geom}}$ is the one with the highest impact on the predicted values. So, we want to see how the model works taking its individual values into consideration, as well.

The results are shown in figure 4.7 and the third row of figure 4.6. In each row, we compare the behavior of a different parameter in the ΔX-N_μ plane for the simulations, obtained using EPOS-LHC with vertical protons of 10^{20} eV (left), with the distribution obtained through our model (right). The ΔX-N_μ 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 ΔX-N_μ plane represent quite well those from simulations means that the behavior of these parameters is qualitatively and even numerically well captured in the model. $\langle \tilde{f}_{\mathrm{ch,en}} \rangle_{\mathrm{geom}}$ follows the correct behavior for lower values only.

This model is adequate to describe the ΔX-N_μ 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 fraction 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 [122] (Numpy and Matplotlib), Scikit-learn [123] and Keras [124, 125] using Tensorflow [126], all interpreted by Python 3.6 [127]. 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 $\langle X_{\text{max}} \rangle$, $\langle N_\mu \rangle$, $\sigma(X_{\text{max}})$ 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 [128, 129]. 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 the 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 the ground (at a vertical atmospheric depth of $880\,\mathrm{g\,cm^{-2}}$), 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 $1\,\mathrm{GeV}$. 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 right away taking into account the features that are suggested by our model presented in section 4.2 and 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 the 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 $\ln(N_{\mathrm{tot}}^{\mathrm{FI}}+1)$ instead of $N_{\mathrm{tot}}^{\mathrm{FI}}$ 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 $N_{\mathrm{tot}}^{\mathrm{FI}}$ 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 $\ln(X_0)$. 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 will be 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 value of 20 % for the test set but only 10 % out of the remaining set for the validation set, since otherwise our training set would get too small.

It has been 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 [130]. 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 that 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 datasets of 1000 instances from then on.

When separating the dataset into training, validation and test sets, another care has to be taken. It is always favorable 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 that 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 is shown 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

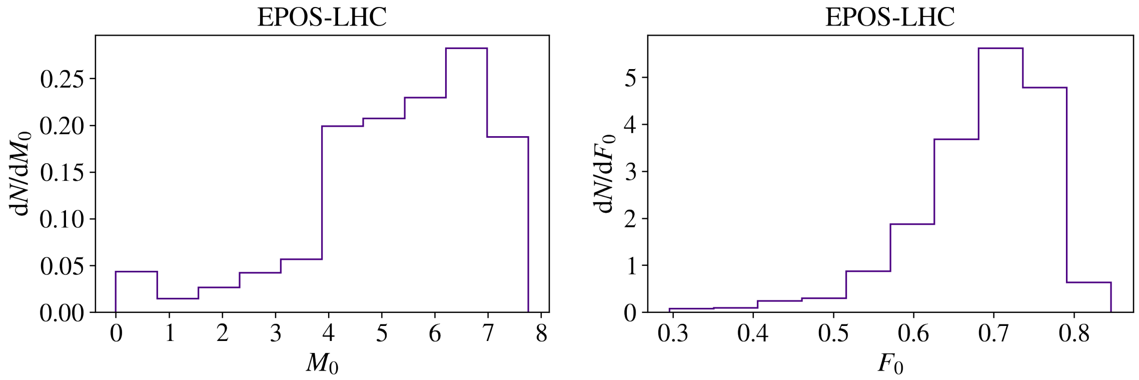

Figure 5.1: Correlations between features and targets suggest that M_0 (left) has the highest impact on X_{max} and F_0 (right) on N_μ. Since the distribution of $N_{\mathrm{tot}}^{\mathrm{FI}}$ presents a very long tail, which typically cannot be properly dealt with in neural networks, we use $\ln(N_{\mathrm{tot}}^{\mathrm{FI}} + 1)$ for the parameter we call M_0 from now on. 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.

layer (red circle) returns the target value. All the layers in between (grey circles) are 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* that 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_{μ} distribution has quite some outliers, we decide to work with the l_1 *norm*, which is also called *mean absolute error*:

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

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 figure 5.2, each arrow

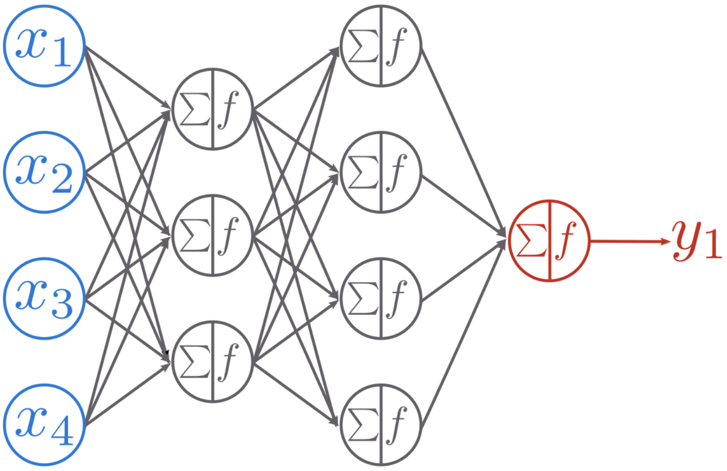

Figure 5.2: Outline of a neural network structure. The input layer (blue circles) introduces the input features. Nodes in hidden layers (grey circles) compute weighted sums of the outputs from previous layers and pass them on to the next layer through an activation function. After calculating optimal 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).

corresponds to a weight parameter, the biases are not shown there for simplicity, but 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 hyperparemeters 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 this process. 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 [131]. 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 favorable for convergence of the loss function to a global minimum [132]. 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 the 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 [133]. 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 $2/n_{\text{in}}$, 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 has 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 the literature [134, 135, 136], 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 the datasets we are interested in modeling, we resort to randomized parameter optimization, a method that will be explained in section 5.2. With this algorithm, we get a rough idea of promising architectures. Then, we fine-tune by training neural networks using many combinations close to these values.

The first dataset we are interested in modeling is the set of simulations obtained for 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 the 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 datastes.

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 sets. As is usual, the model performs better on the training set. The early stopping method prevents the gap between the two curves from 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 $20\,\mathrm{g\,cm^{-2}}$. 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 described very well by the seven 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 anticorrelation is less pronounced and X_{max}

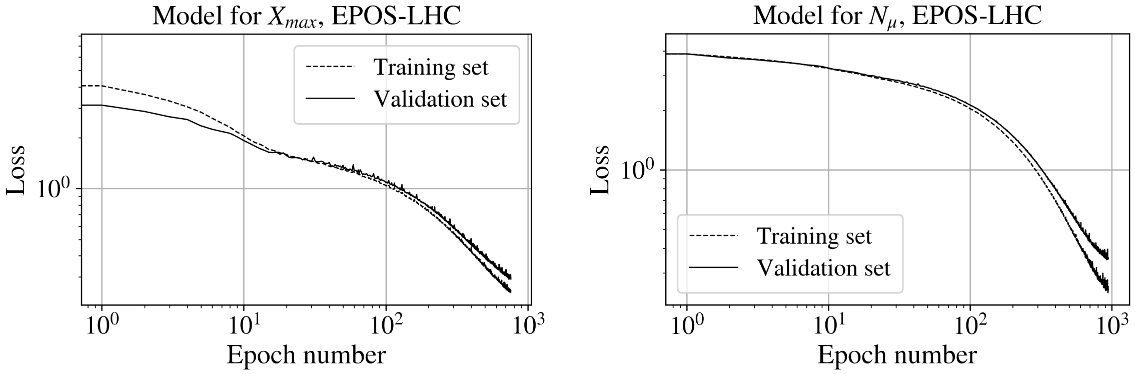

Figure 5.3: Learning curves of the neural networks trained to predict X_{max} (left) and N_μ (right), respectively. These networks are trained on the dataset of vertical proton primaries of 10^{20} eV simulated with CONEX using EPOS-LHC.

and N_μ reach lower values.

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 the three high-energy interaction models considered in this work. Even though these models 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 develop a neural network that is trained on the joint ensemble consisting of the 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

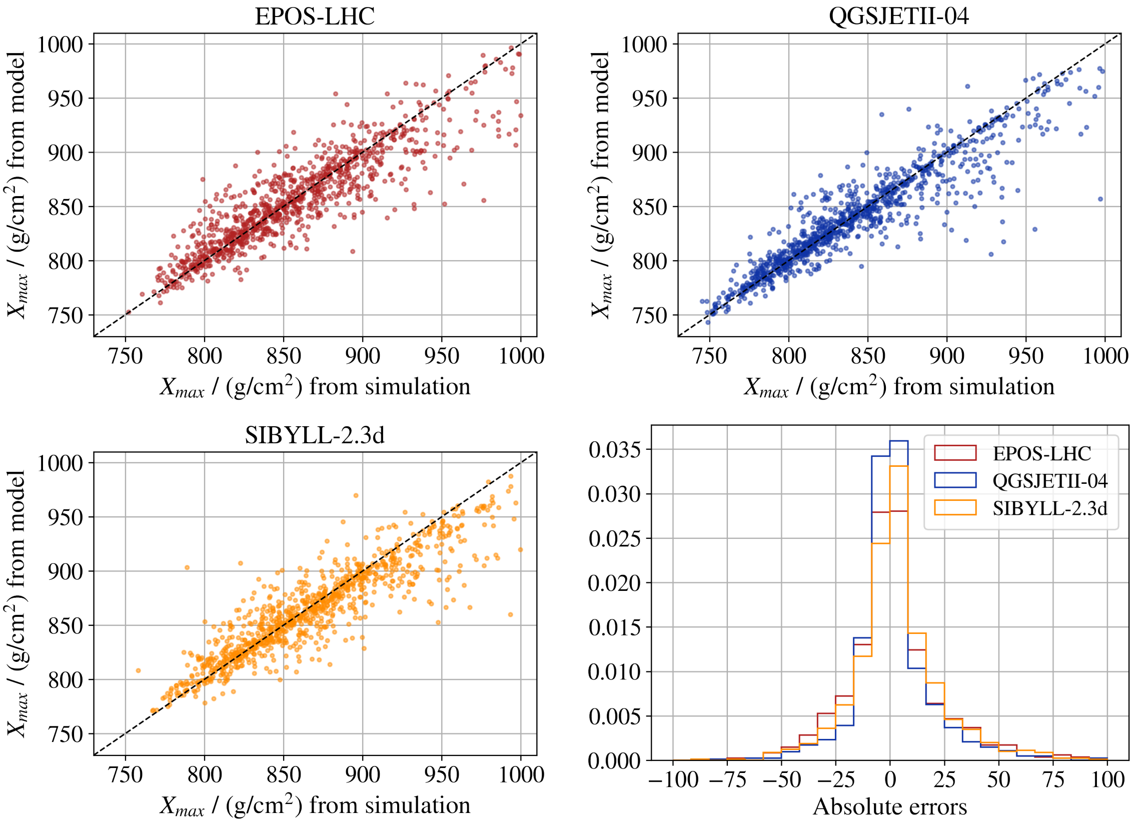

Figure 5.4: Initially, we investigate how well neural networks perform on showers generated by vertical protons of 10^{20} eV because the anticorrelation is strongest for this primary group and it is desirable to see it explained through these models. The first approach is to consider the datasets simulated using 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 simulations 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.

mass composition depends on the primary energy and the high-energy interaction model 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 $X_{\mathrm{max}} - N_\mu$ distributions almost don't overlap with that of 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

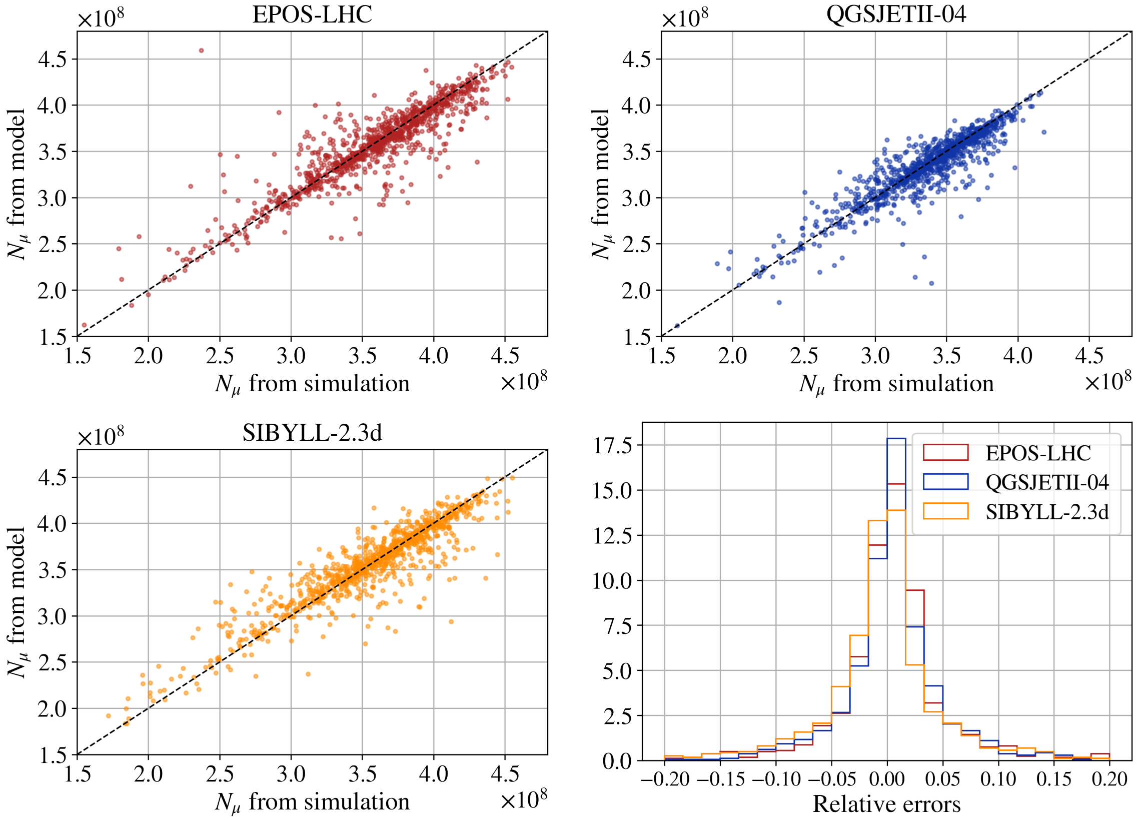

Figure 5.5: As presented in figure 5.4, a neural network is trained on each training set simulated using the different high-energy interaction models separately, but this time in order to predict the number of muons N_μ. The comparison between the true values from simulations and the values obtained 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.

to obtain enough events in this energy range for which we can obtain X_{max} and N_{μ}.

With that in mind, we decide on the energy range $\ln(E_0) = 18.098$ to 18.198 (see figure 1.6). 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 obtain the best unified network for the prediction of X_{max} 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 each gave the

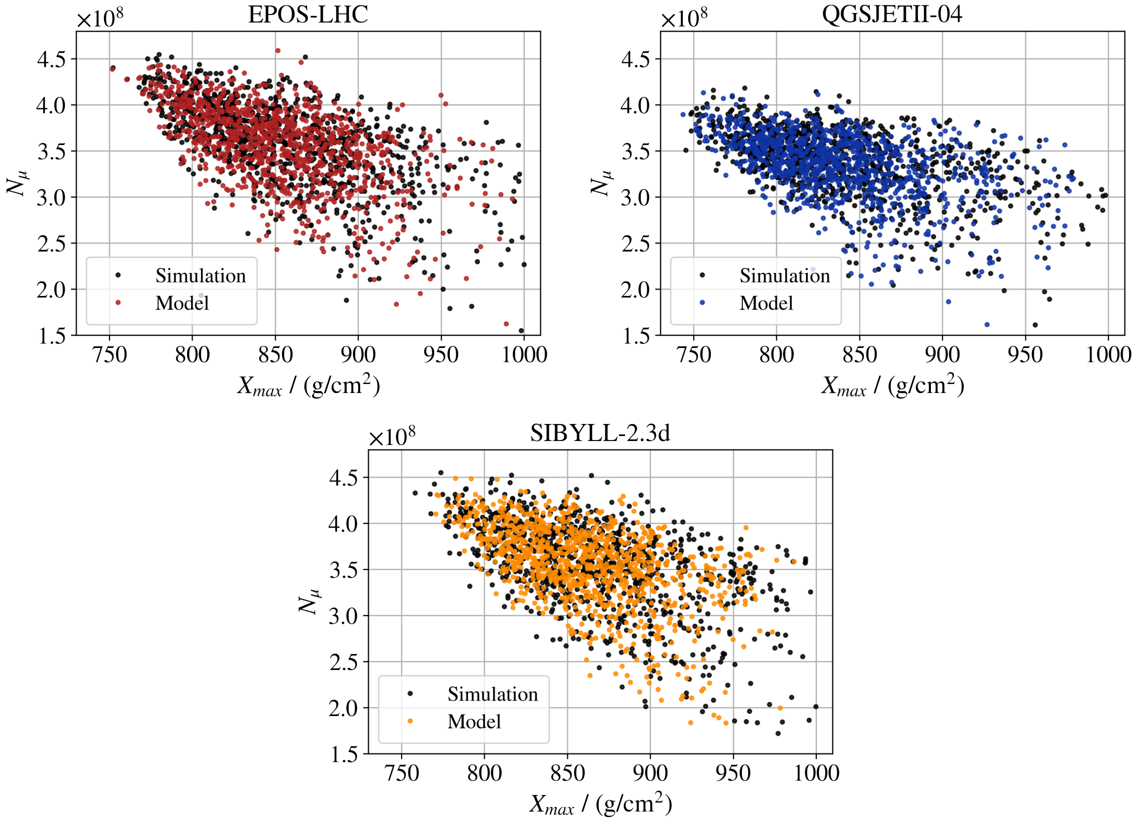

Figure 5.6: The results shown in figures 5.4 and 5.5 are combined in order to compare the X_{max}-N_{μ} distributions from simulations (black dots) to the distributions obtained through our models (colored dots), for each high-energy interaction model. The anticorrelation is well reproduced in all cases.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

best result (see figure 5.8). Also here, the relative errors are in general slightly higher than before, but also distributed very similarly. Only a small bias for EPOS-LHC and QGSJETTII-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 the 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

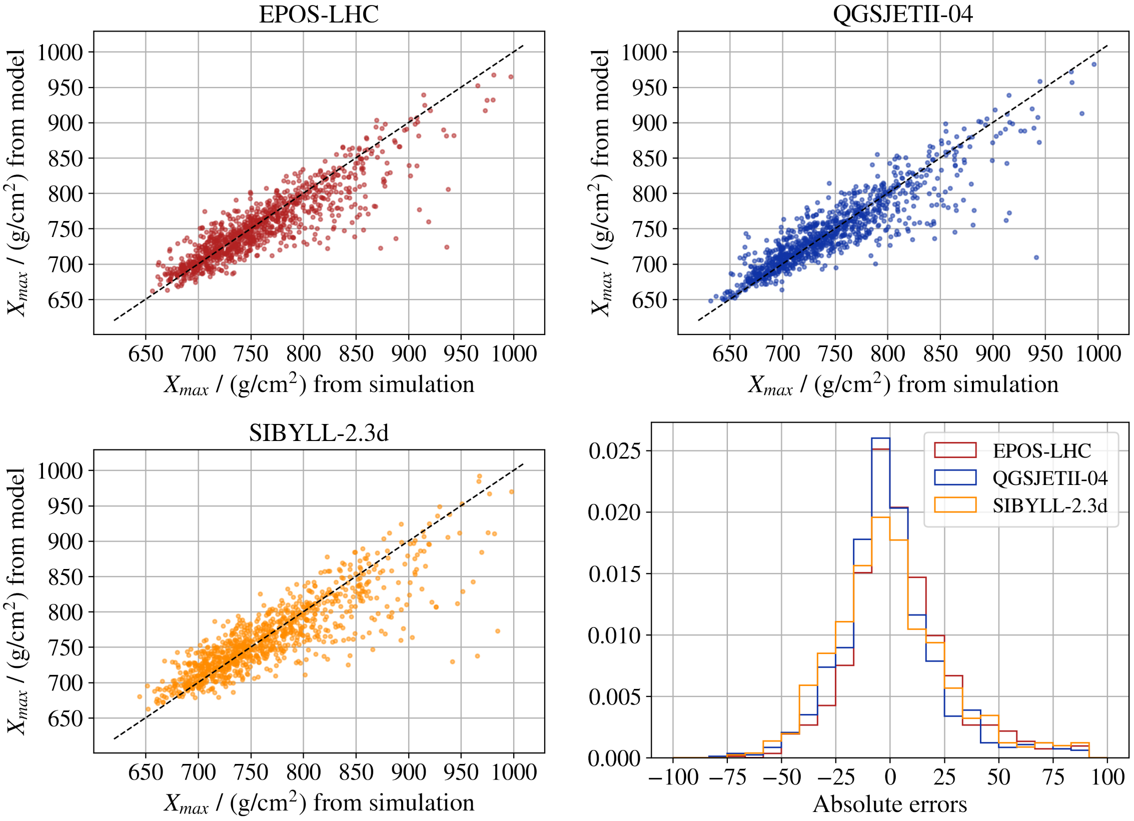

Figure 5.7: We now use the complete dataset of simulated showers generated by protons of primary energy between $\ln(E_0) = 10^{18.098}$ eV and $10^{18.198}$ eV, comprising all three high-energy interaction models. The goal is to obtain a model that predicts the $X_{\max}$-N_μ anticorrelation independently 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 neural network trained to obtain $X_{\max}$ is then tested on each subdataset separately and the results are shown here. The absolute errors are slightly higher than for the neural networks trained on each dataset separately: 27 g cm^{-2}, 26 g cm^{-2} and 32 g cm^{-2} 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.

that the model we obtained is independent of the high-energy interaction model used.

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

In the previous section, we used seven features and included then the primary energy as an input to our model and obtained a very good predictability. However, we know that real data is affected by uncertainties in the measurement and the reconstruction.

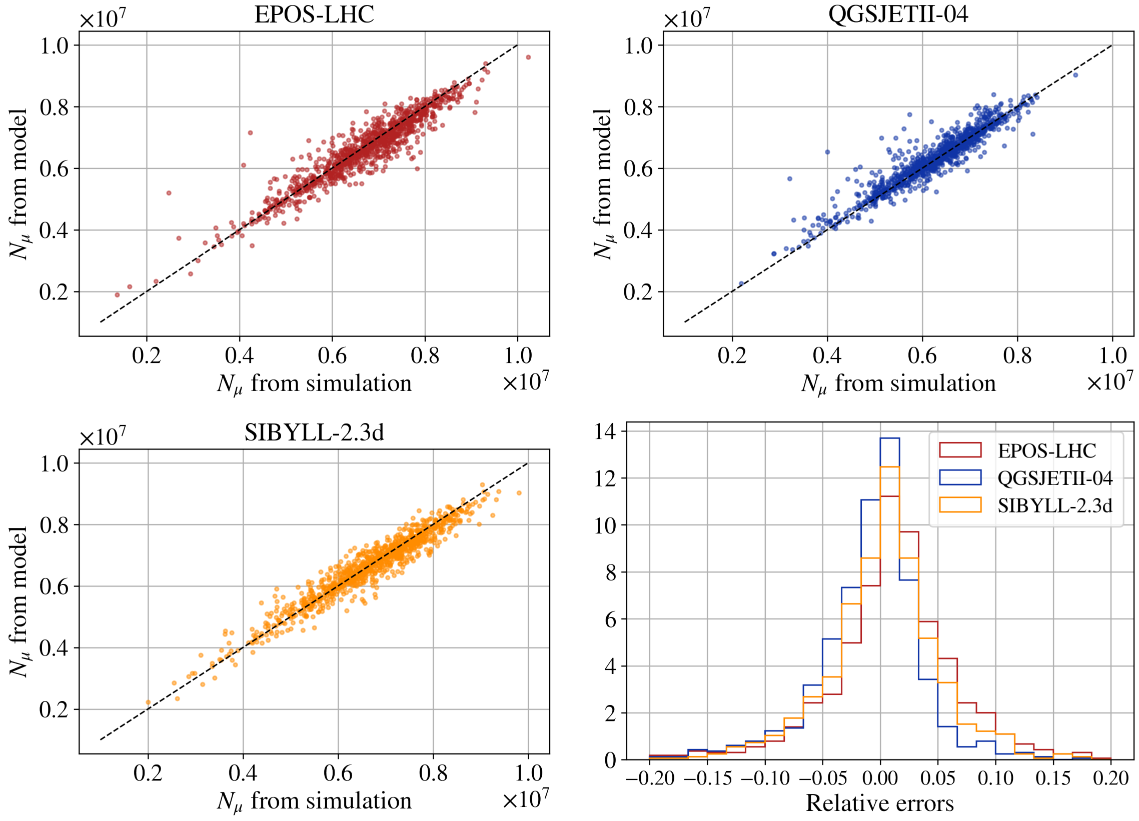

Figure 5.8: The same procedure as presented in figure 5.7 is carried out with N_μ as the target. The relative errors are again slightly higher (except for SIBYLL-2.3d) than for separately trained models: 7.4 %, 6.3 % and 5.1 % for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively. However, these errors 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 and in opposite directions can be observed.

Furthermore, the cosmic ray flux comprises all types of primaries ranging from proton 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. 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

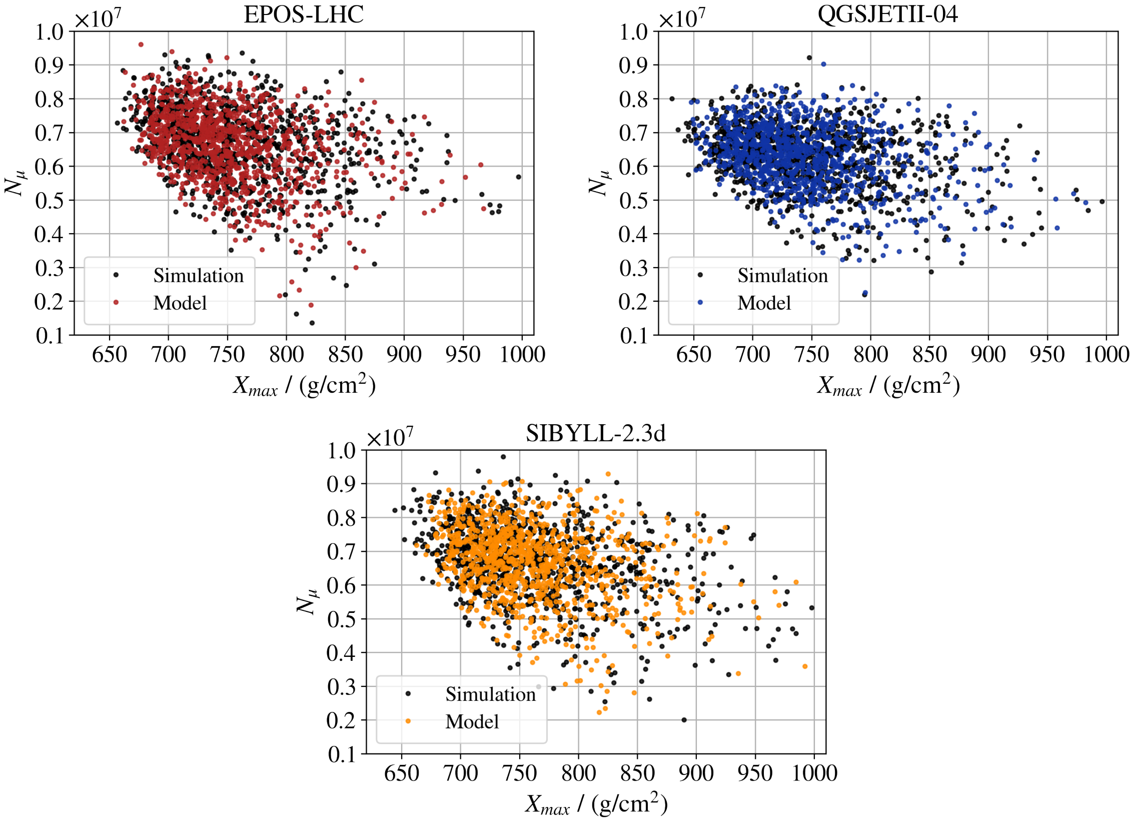

Figure 5.9: The results shown in figures 5.7 and 5.8 are combined in order to compare the $X_{\max}$-N_μ distributions from simulations (black dots) to the distributions obtained through our unified model (colored dots), for each high-energy interaction model. Also for this energy group and using a single 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.

between the distributions of the features and those of the targets. In order to characterize 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 to discard, 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 tasks [125, 137]. 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 thought of 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 variations in the training data.

A way to improve the algorithm substantially is by considering an *ensemble* of decision trees. The algorithm, in which each decision tree is trained on a different subset of the training set and in which then all the predictions are combined into a mean value as the final prediction, is called a *random forest*. The multiple subsets are formed 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 [138, 139]. 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 [138].

In order to use the attribute from the random forest regressor that 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 be considered 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 hyperparameter optimization than trials on a complete grid [140]. 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 each of the K models on the corresponding validation set. Individual performances are measured via the mean squared errors in the corresponding validation set. For the hyperparameter combination that performs best, we obtain the relative importance 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 be considered 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. From analyzing the correlations, we already had a few ideas of which are the most important features for each model 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 [67], 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. 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 set of features comprising E_0, X_0, M_0 and F_0, a large variety of neural network architectures is evaluated for the N_μ prediction. 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 rates between 0.001 and 0.0001, hyperparameters of l_2-regularization between 0.001 and 0.01 and patience hyperparameters between 10 and 25 epochs are tested. After this broad search, another more localized randomized search is performed.

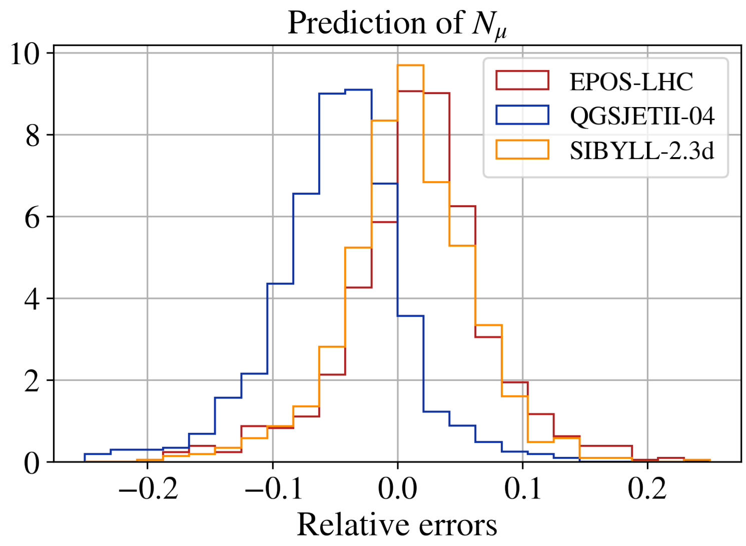

Figure 5.10: Relative errors for the N_μ prediction from a neural network trained using the same basic configuration as described in section 5.1, combined with 4 layers of 50 nodes each. The features used are E_0, X_0, M_0 and F_0. These are only close to sufficient to capture all the differences between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

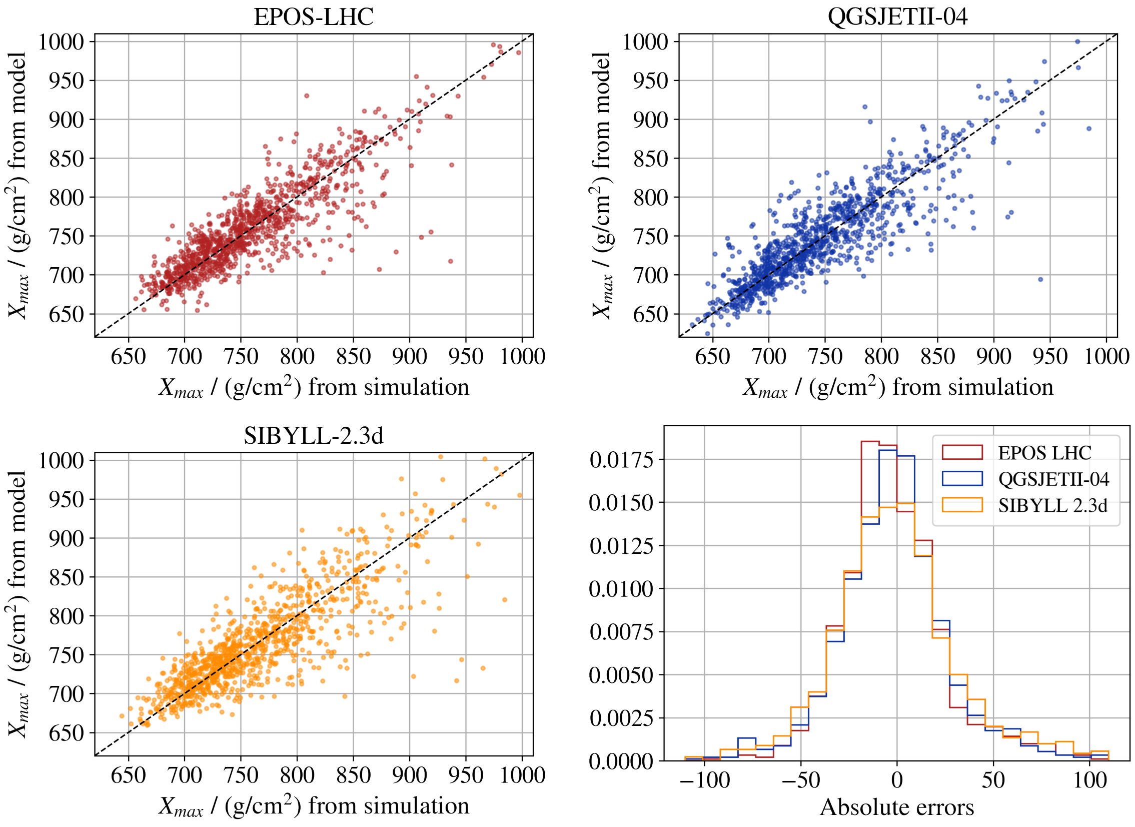

Figure 5.11: The same dataset of simulated showers as in figure 5.7 is used here. This time, a reduced set of features comprising E_0, X_0, M_0, F_0 and K_{eff} is implemented. The neural network trained to obtain X_{max} is then tested on each subdataset and the result is shown here. The absolute errors are slightly higher than for the unified neural network trained using the complete feature set: 30 g cm^{-2}, 31 g cm^{-2} and 37 g cm^{-2} for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively. However, the errors are still small enough and the neural network makes predictions in a model independent way, as is needed.

When analyzing the relative errors between the real values and the predictions of N_μ, the result is always qualitatively similar to the distributions shown in figure 5.10, which correspond to one particular case. The muon number is overestimated in showers generated with QGSJETII-04 and underestimated in showers generated with EPOS-LHC and SIBYLL-2.3d, by a very similar magnitude. This means that there is some mechanism missing in this reduced number of features that takes energy from the hadronic channel in QGSJETII-04 or that keeps energy in it for the other two models (or that does both to some extent). The best the neural network can do is find an "intermediate" model, which has simultaneously for all high-energy interaction 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} to model N_μ, the missing mechanism is still not covered and results very similar to those observed in figure 5.10 are obtained. 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

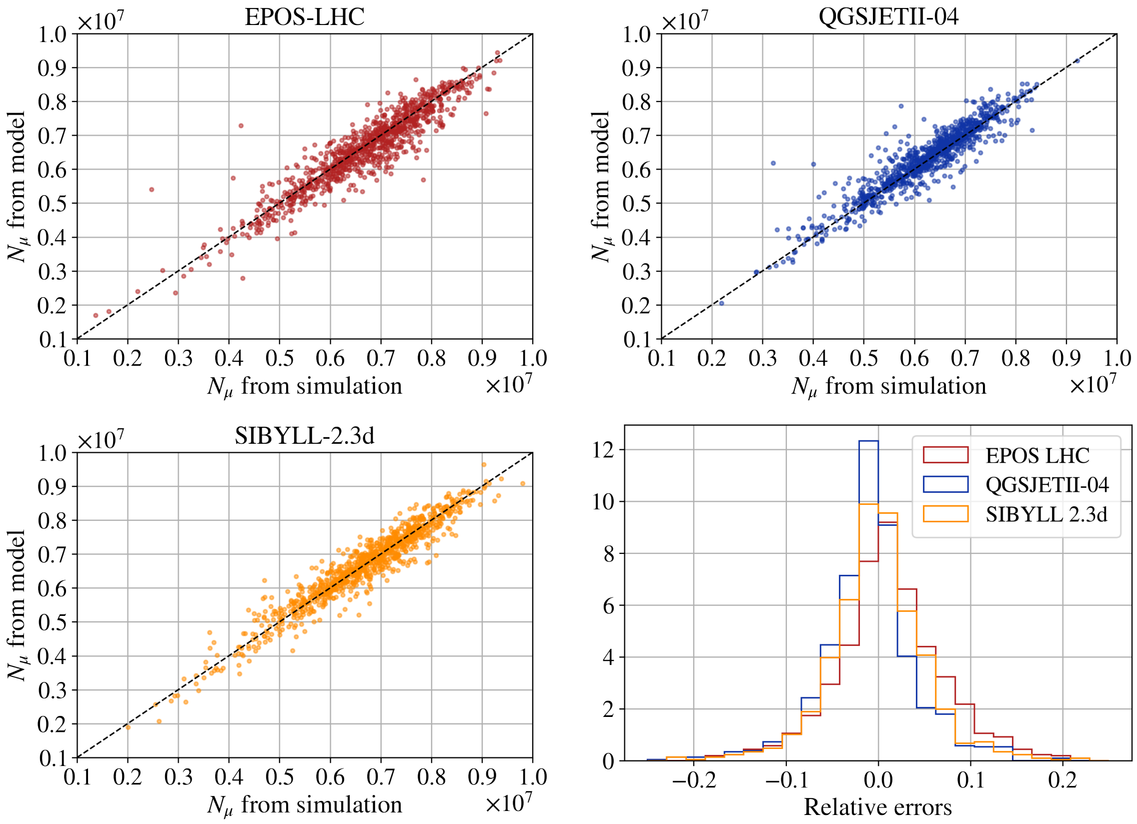

Figure 5.12: The same dataset of simulated showers and the same reduced set of features as implemented in figure 5.11 are used to model the number of muons. The relative errors are slightly higher than for the unified neural network trained on the complete feature set: 7.7 %, 6.5 % and 5.4 % for EPOS-LHC,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

nodes each, gives the best result for predicting $X_{\max}$. The results are shown in figure 5.11. The absolute errors are around $30\,\mathrm{g\,cm^{-2}}$, $31\,\mathrm{g\,cm^{-2}}$ and $37\,\mathrm{g\,cm^{-2}}$ for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively. They are a bit higher than the errors obtained when using the complete feature set, which is expected because leaving features out can lead to some ambiguity. However, these errors are still small enough and the neural network gives predictions in a model independent way, as is needed.

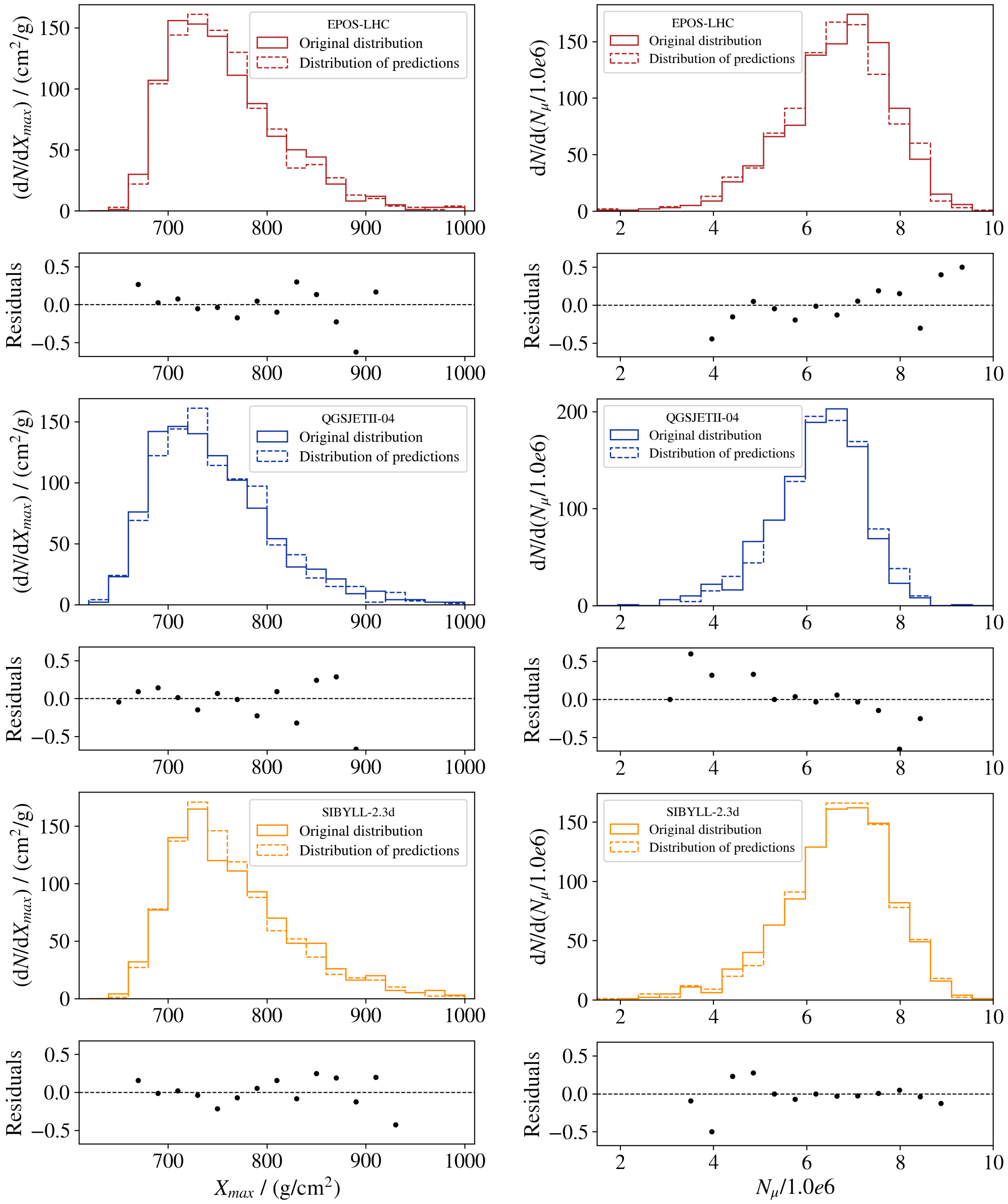

Figure 5.13: Comparison between the original distributions of $X_{\max}$ (left) and N_μ (right) and the distributions of the corresponding predictions obtained via the neural networks discussed in figures 5.11 and 5.12. The residuals for all the combinations are shown, as well.

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 of 4 layers with 50 nodes each. The results are shown in figure 5.12. 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. 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 then, 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 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 $1\,\text{g}\,\text{cm}^{-2}$, in all cases. The dispersion of X_{max} is generally underestimated by around $1\,\text{g}\,\text{cm}^{-2}$. 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 from simulations. 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 the hadronically interacting particles of the first interaction 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 in 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 $\lambda = 47\,\mathrm{g\,cm^{-2}}$, 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) that 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 (location and scale) parameters, it would be impossible to know which relation between such new parameters is the correct one.

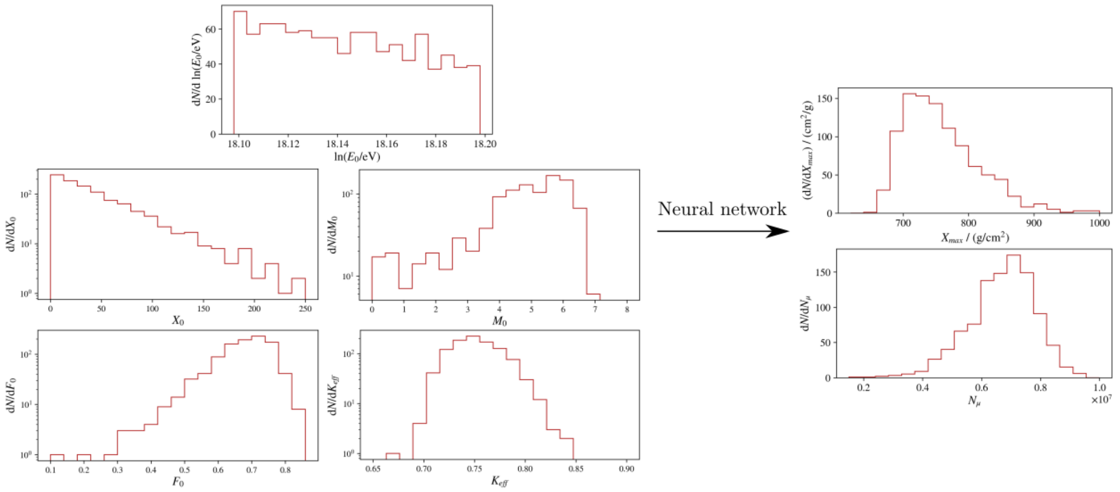

Figure 5.14: Depiction of our method. In sections 5.1 and 5.2, we developed a neural network that yields, on a shower-to-shower basis, $X_{\max}$ as a function of E_0, X_0, F_0, M_0 and K_{eff} and another network that returns N_μ. Even though the model was trained on individual instances, we now switch 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.

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

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

where a is the location parameter and b is 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 of the small impact this parameter has on the observables that we will use (see explanation to figure 5.17), we will fix it at this mean value: $M_0^{\rm scale} = 1.05$. Fitting the Gumbel distribution (see figure 5.15, right), with the scale parameter fixed, gives $M_0^{\rm loc}(E) = 5.32$, $M_0^{\rm loc}(Q) = 5.78$ and $M_0^{\rm loc}(S) = 5.40$ (E, Q and S in the parenthesis stand for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, respectively). In the case of the hadronic energy fraction F_0, there are no coincidences, so we need to leave both parameters free. The fitting parameters we obtain are $F_0^{\rm loc}(E) = 0.725$, $F_0^{\rm loc}(Q) = 0.708$, $F_0^{\rm loc}(S) = 0.714$, $F_0^{\rm scale}(E) = 0.066$, $F_0^{\rm scale}(Q) = 0.058$ and $F_0^{\rm scale}(S) = 0.068$ (see figure 5.16, left). $K_{\rm eff}$ follows approxima-

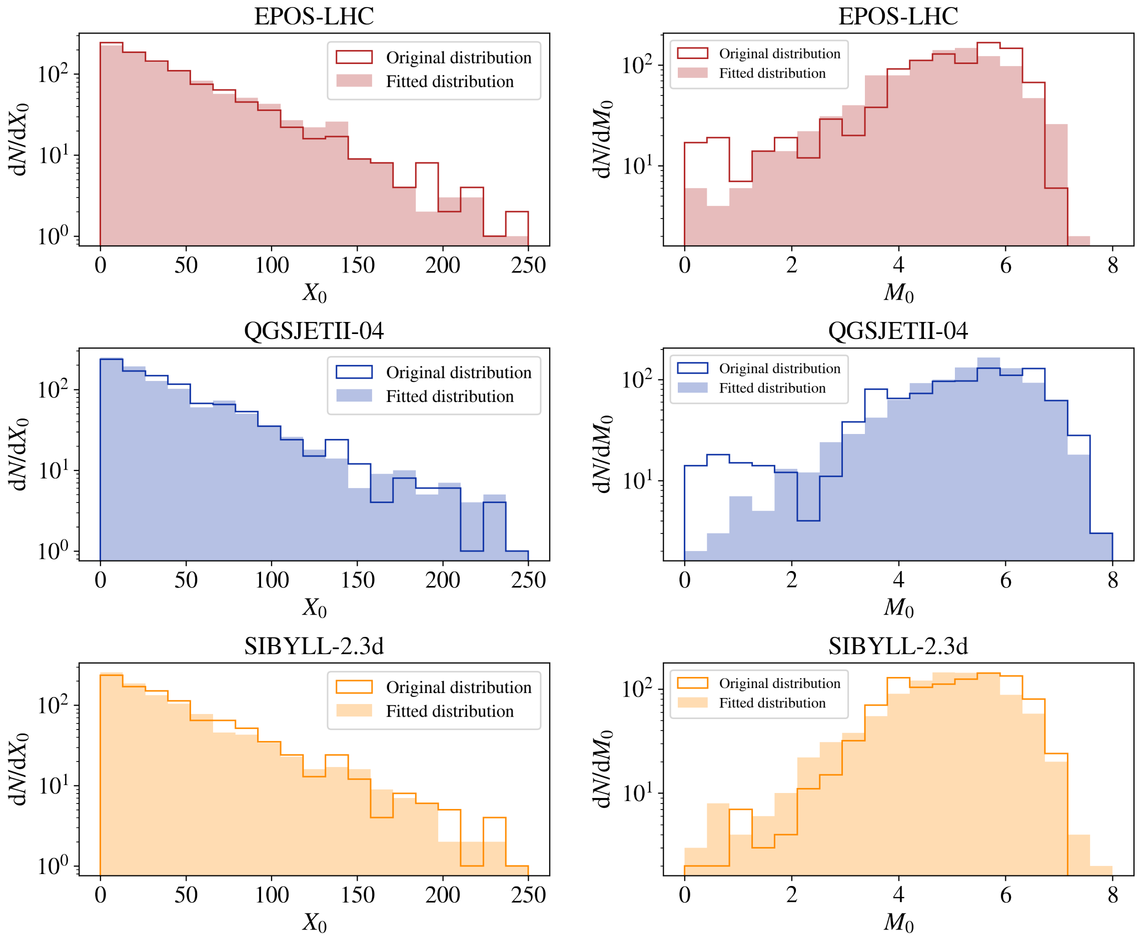

Figure 5.15: Distributions of X_0 (left) and M_0 (right). The depth of the first interaction follows an exponential distribution with $\lambda = 47\,\mathrm{g\,cm^{-2}}$. The hadronic multiplicity can be described with the pdf of the Gumbel minimum distribution. The scale parameter is very close to 1.05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

tely a normal distribution (see figure 5.16, right) whose width is very similar among all high-energy interaction models and is thus fixed at the mean value $K_{\text{eff}}^{\text{scale}} = 0.024$. In figure 5.17, it becomes evident that changing this parameter has no significant effect on the observables. The mean values we obtain are $K_{\text{eff}}^{\text{loc}}(E) = 0.752$, $K_{\text{eff}}^{\text{loc}}(Q) = 0.813$ and $K_{\text{eff}}^{\text{loc}}(S) = 0.732$.

We have just seen that the three features with unknown distributions can be approximately described using six 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 [112], we vary these parameters by multiplying them by a factor. We do this for each parameter, while leaving the other five 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

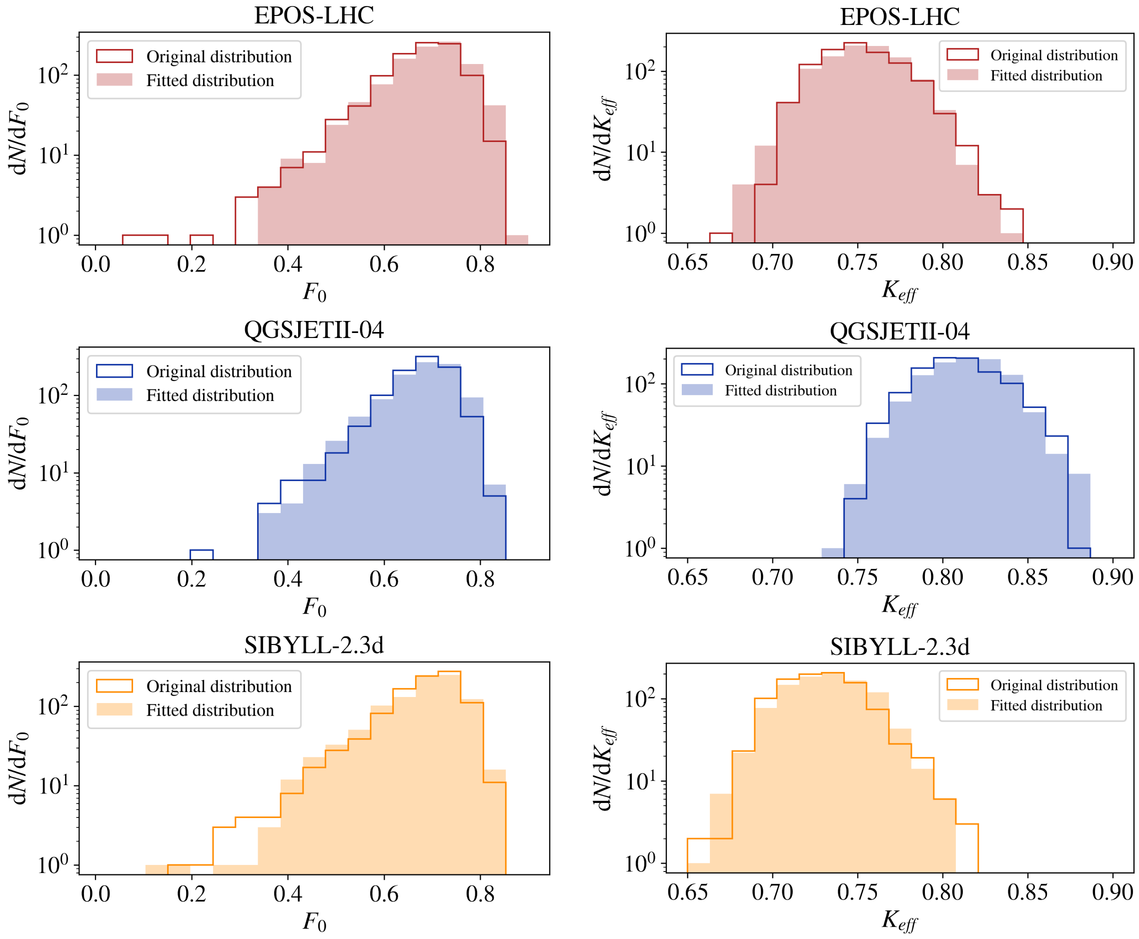

Figure 5.16: Distributions of F_0 (left) and K_{eff} (right). The hadronic energy fraction can be described with the probability distribution function of the Gumbel minimum distribution. The location and scale parameters vary for the hadronic interaction models considered and are thus left free. The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

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 networks were trained.

We also show the true values of the corresponding observables as dashed black horizontal lines and the 1σ error in the measurement of the values as dashed grey lines. These estimates are calculated using the *Bootstrap Method* [141]. 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 extracted and represented by black and grey lines, respectively.

The results are shown in figures 5.17, 5.18 and 5.19 for the different high-energy

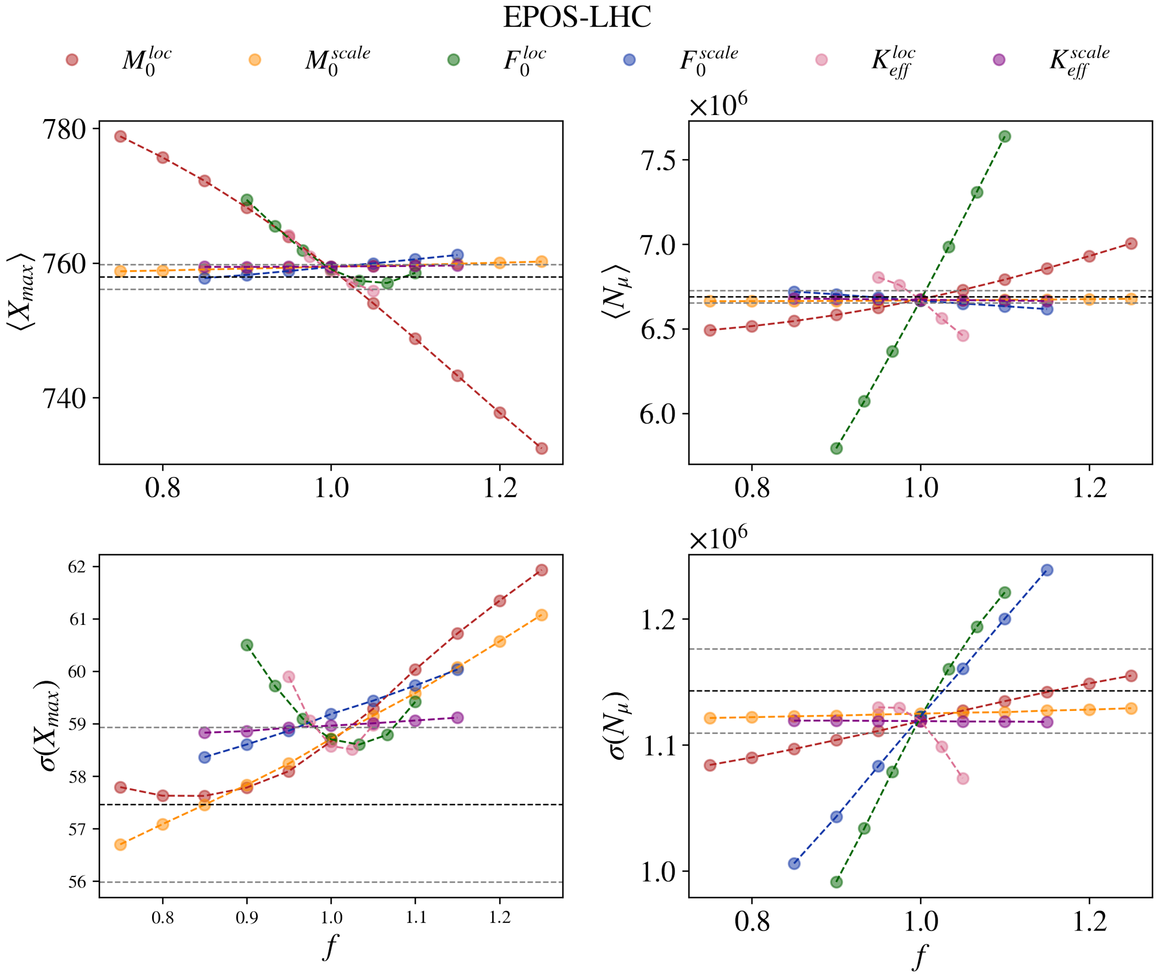

Figure 5.17: Impact of changing the basic parameters, used to describe the distributions of the features, on the means and standard deviations of $X_{\max}$ and N_μ for EPOS-LHC.

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. This way, 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 $K_{\text{eff}}^{\text{scale}}$ (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 $\sigma(X_{\text{max}})$, it is the one parameter that has the least impact of all, after $K_{\text{eff}}^{\text{scale}}$. 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 $\sigma(X_{\text{max}})$ is very small.

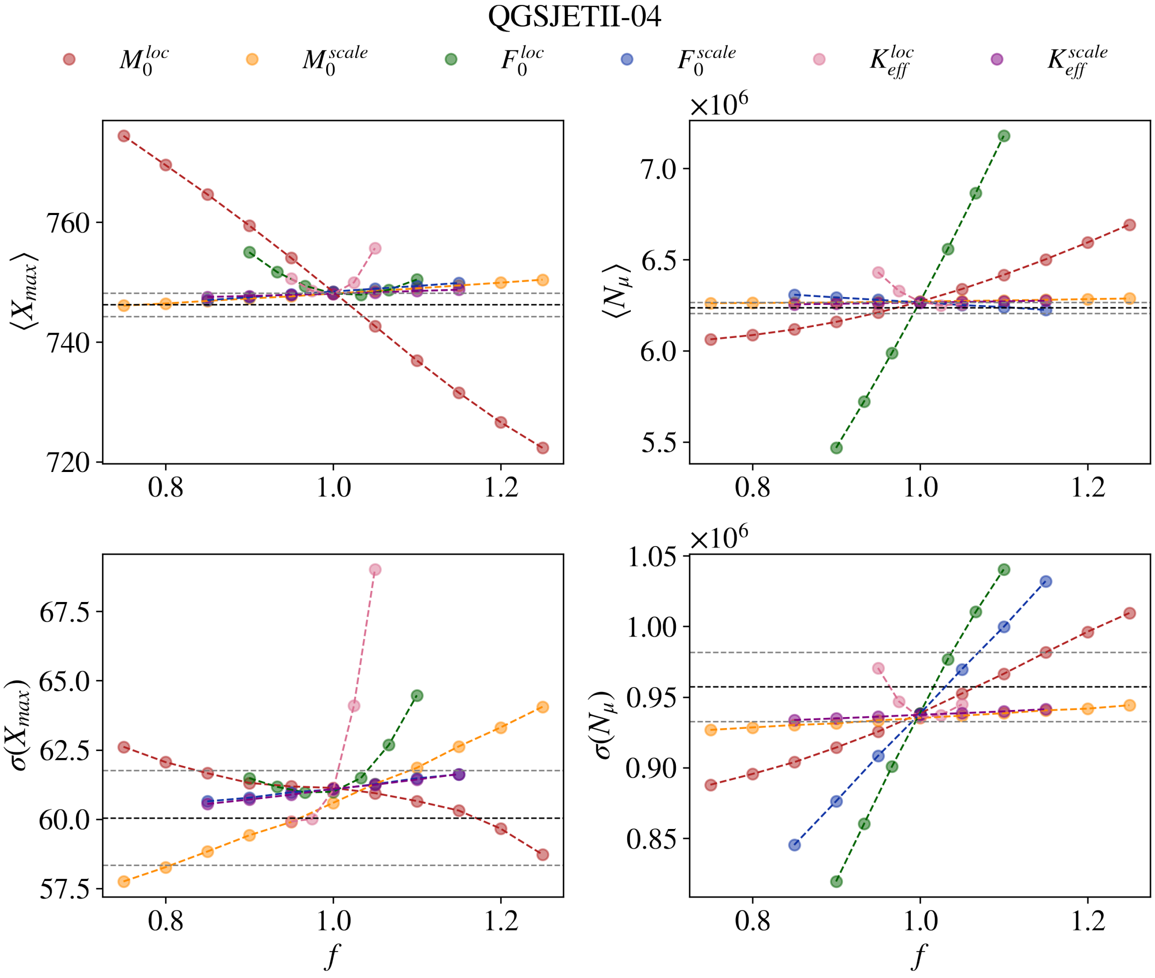

Figure 5.18: Impact of changing the basic parameters on the means and standard deviations of X_{max} and N_μ for QGSJETII-04.

The location parameter of the charged multiplicity M_0 (red circles) has the strongest impact on $\langle X_{\max} \rangle$. 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. At the same time, 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.

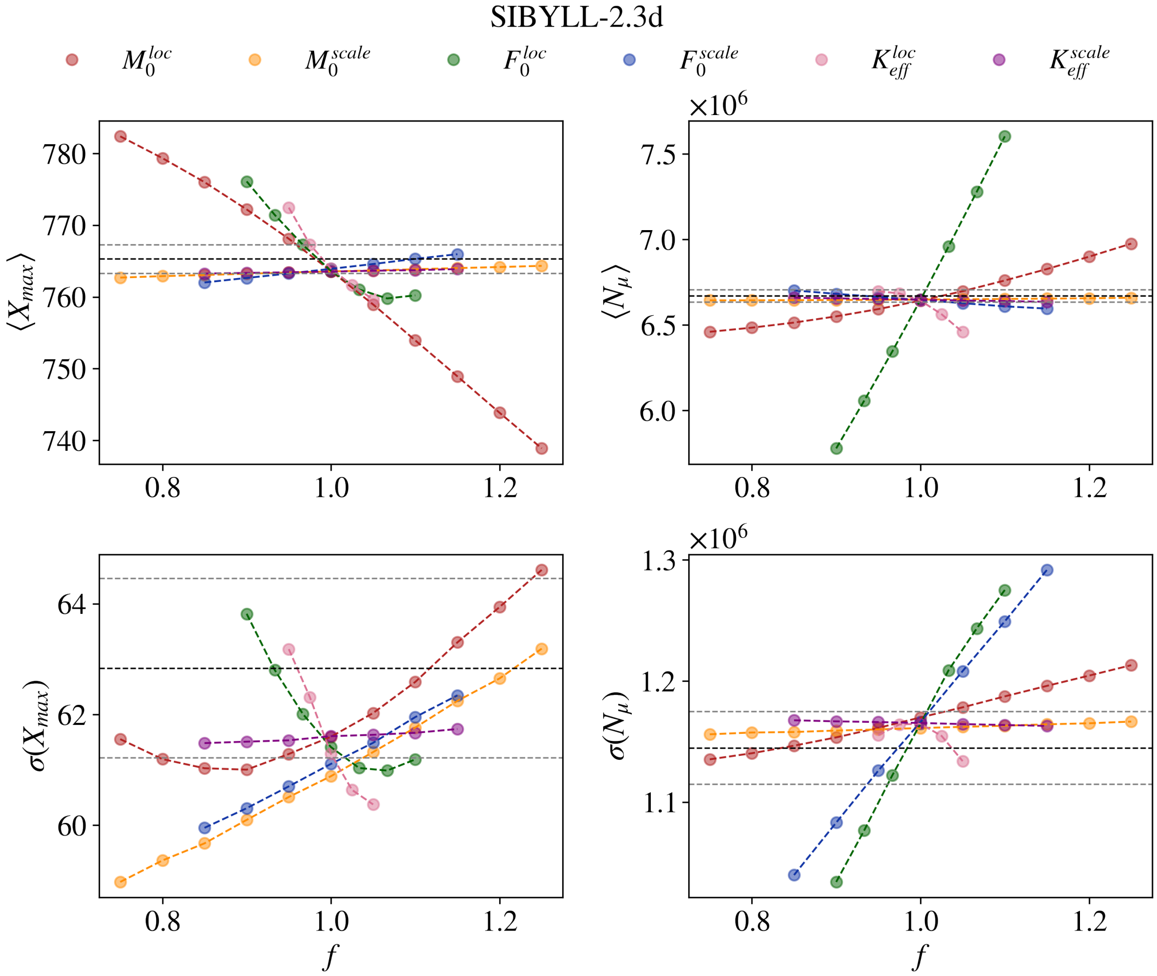

Figure 5.19: Impact of changing the basic parameters on the means and standard deviations of $X_{\max}$ and N_μ for SIBYLL-2.3d.

An increasing value of the scale parameter of the charged multiplicity M_0 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 energy 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 behavior 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 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 behavior. 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 [112], 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 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}, $K_{\mathrm{eff}}^{\mathrm{loc}}$, F_0^{loc} and F_0^{scale}. We do this for each high-energy interaction model, in order to test

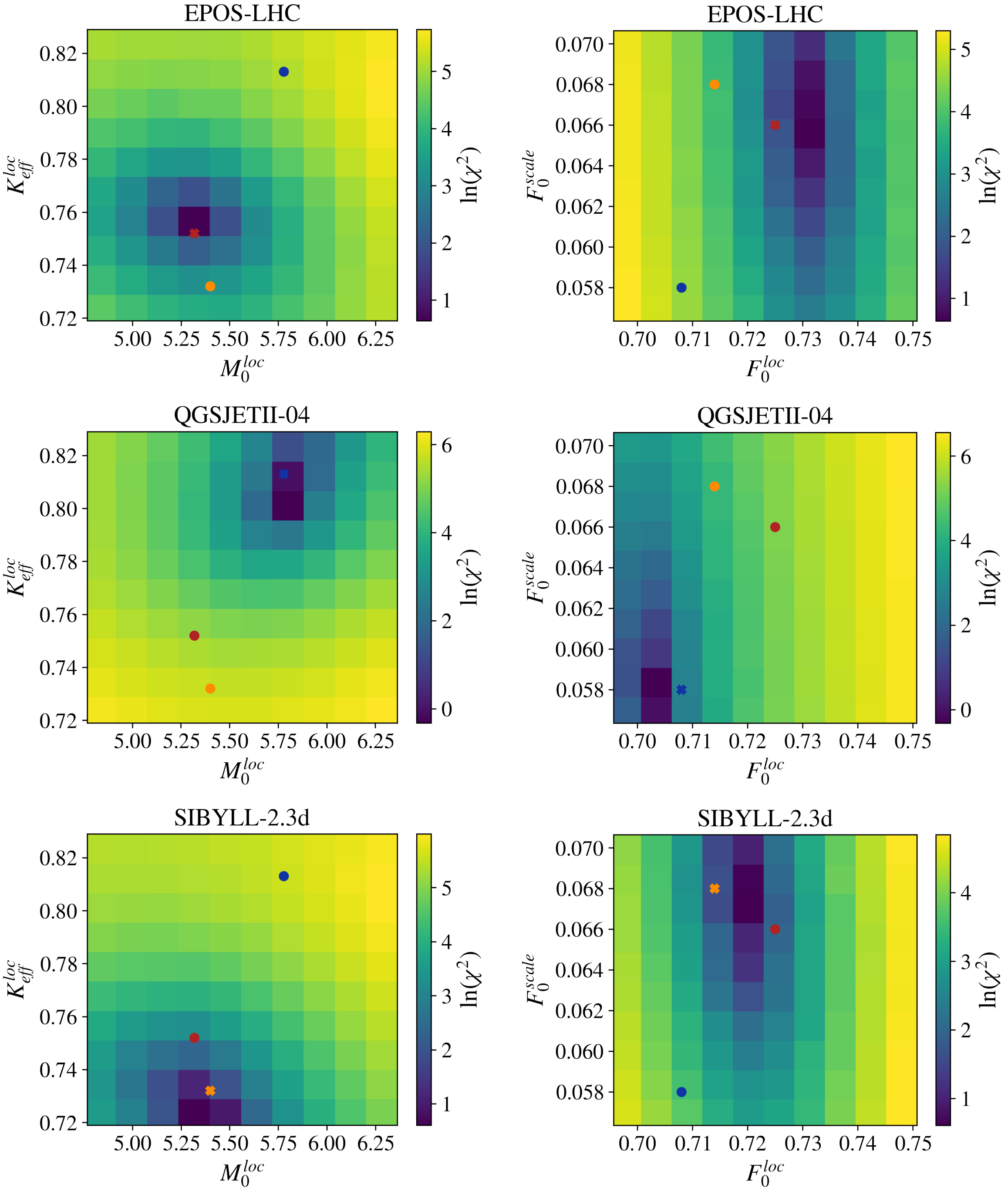

Figure 5.20: Logarithm of the 4-dimensional χ^2-function, calculated to get the unknown parameters M_0^{loc}, $K_{\mathrm{eff}}^{\mathrm{loc}}$, F_0^{loc} and F_0^{scale}. The real values for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d are shown as red, blue and yellow markers, respectively. In each figure, 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: $\langle X_{\mathrm{max}} \rangle$, $\langle N_\mu \rangle$, $\sigma(X_{\mathrm{max}})$ 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} to where the minimum is obtained and show the logarithm of the χ^2 as a function of M_0^{loc} and $K_{\mathrm{eff}}^{\mathrm{loc}}$ (left). Equivalently, the χ^2 as a function of F_0^{loc} and F_0^{scale} is shown on the right. In each case, the minimum is achieved very close to the position of the corresponding real value. The logarithm of the χ^2-function is used here only to make the position of the minimum more visible.

the performance of our model in predicting the unknown parameters and, in particular, to test how well our model can distinguish between the high-energy interaction models. More in detail, we define

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

where $\bar{\theta}$ is the vector of the parameter values that are tested for M_0^{loc}, $K_{\mathrm{eff}}^{\mathrm{loc}}$, F_0^{loc} and F_0^{scale}. $\hat{\bar{x}}$ 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 the contributions due to statistical and systematic errors:

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

as suggested in [143, 144]. Since their measurements are correlated, the covariance matrix $(V_{\mathrm{stat}})_{ij} = \mathrm{cov}[\hat{x}_i, \hat{x}_j]$ needs to be used here [145]. For the systematic errors, we use $V_{\mathrm{syst}} = \bar{\mathbf{s}}\bar{\mathbf{s}}^{\mathrm{T}}$, where $\mathbf{s}$ contains the difference between the real values of the observables and the values predicted when 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 $\hat{\bar{\theta}}$ 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

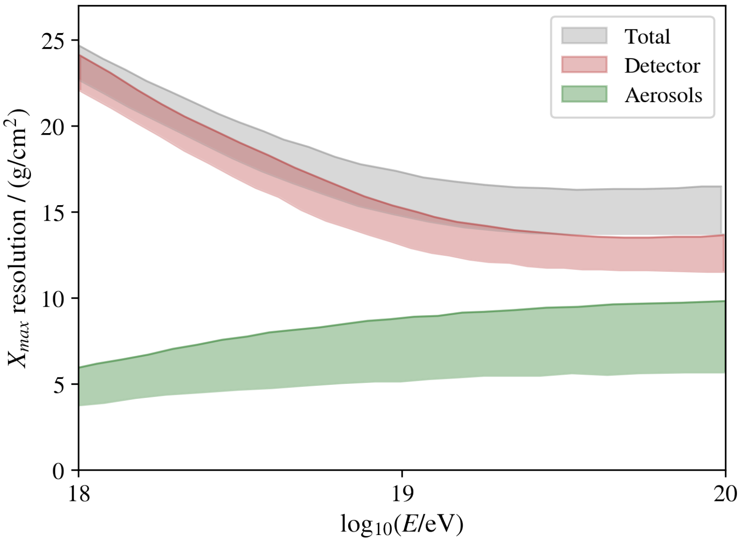

Figure 5.21: For our analysis, we need the $X_{\max}$ value that is observed by the fluorescence telescopes of the Pierre Auger Observatory. In this plot, the $X_{\max}$ resolution as a function of energy is shown (reproduced from [142]). Bands denote the estimated systematic uncertainties. Contributions 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.

obtained and show the logarithm of the χ^2 as a function of M_0^{loc} and $K_{\mathrm{eff}}^{\mathrm{loc}}$ on the left. Similarly, we show the χ^2 as a function of F_0^{loc} and F_0^{scale} on the right. In each case, the minimum is achieved very close to the position of the corresponding real value. 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 two-dimensional slice and searching for the minimum gives the predictions (stars in figure 5.22) for the four parameters M_0^{loc}, $K_{\mathrm{eff}}^{\mathrm{loc}}$, 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 $\chi^2_{\mathrm{min}} + 4.72$ and $\chi^2_{\mathrm{min}} + 9.7$, respectively. These are the values that need to be used in the case of 4 dimensions [146]. Comparing the real values (filled circles) with the predictions, one can see that M_0^{loc}, $K_{\mathrm{eff}}^{\mathrm{loc}}$ 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 three high-energy scenarios.

Before moving on to the dataset from the Pierre Auger Observatory in the final chapter, we can test the performance of our models including the effect of the resolution of the measurements. As is shown in figure 5.21 (reproduced from [142]), for the showers in the small energy range we used for our simulations, an absolute value of around $24\,\mathrm{g\,cm^{-2}}$ is present. For the number of muons, it is usual to assume a value of $20\,\%$, which is considered quite conservative. We include these effects by smearing the X_{max} values according to a normal distribution centered at the true value and with a standard deviation of $24\,\mathrm{g\,cm^{-2}}$. The N_μ values are also spread 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}, $K_{\mathrm{eff}}^{\mathrm{loc}}$, F_0^{loc} and F_0^{scale} under study, the values of X_{max} and N_μ that are obtained from our neural networks 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.

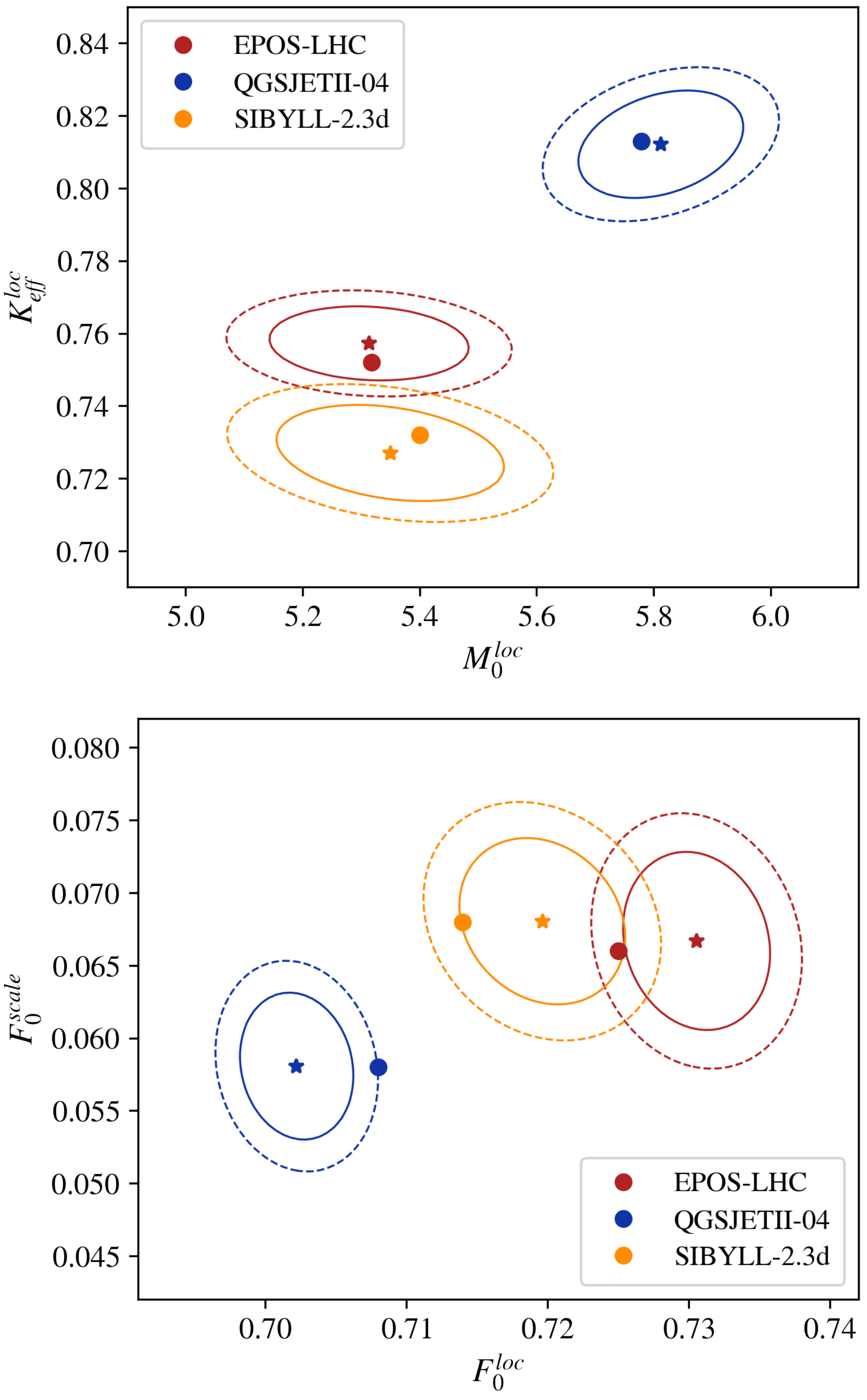

Figure 5.22: Minimum χ^2 fit to obtain the predictions (stars) of the four parameters M_0^{loc}, $K_{\text{eff}}^{\text{loc}}$, F_0^{loc} and F_0^{scale} for EPOS-LHC (red), QGSJETII-04 (blue) and SIBYLL-2.3d (yellow). The χ^2-function shown in figure 5.20 is used for the fit. The 1σ (full lines) and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

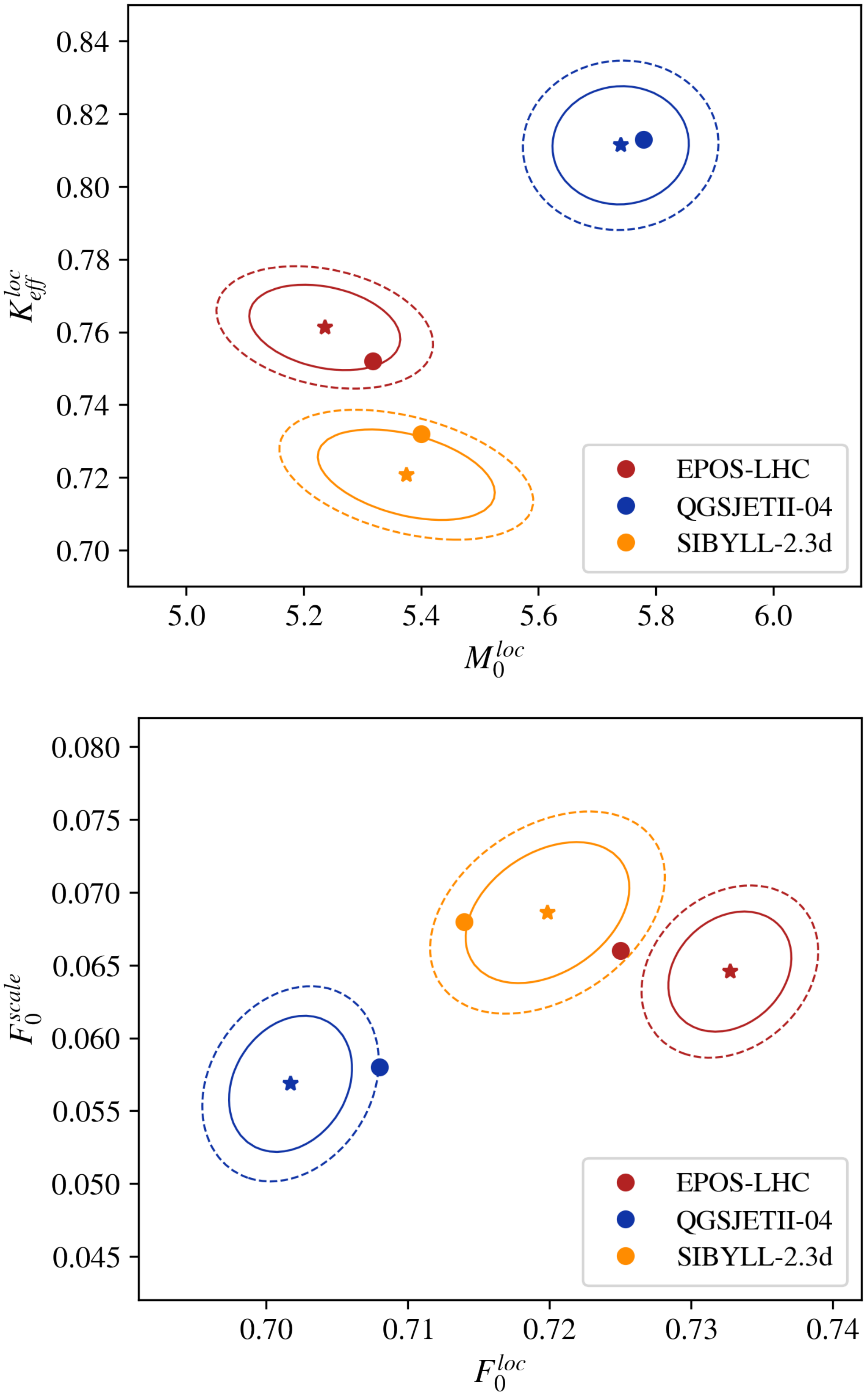

Figure 5.23: Minimum χ^2 fit to obtain the predictions (stars) of the four parameters $M_0^{\rm loc}$, $K_{\rm eff}^{\rm loc}$, $F_0^{\rm loc}$ and $F_0^{\rm scale}$ for EPOS-LHC (red), QGSJETII-04 (blue) and SIBYLL-2.3d (yellow), as calculated for figure 5.22. The difference is that for this analysis we performed modifications on the $X_{\rm max}$-N_μ 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 the Pierre Auger Observatory. For this purpose, the mixed composition in the cosmic ray flux is taken into account [40]. 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 that 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 that contains unbiased values of the depth of maximum development X_{max} and the total number of muons at the ground N_μ for vertical showers ($\theta < 60°$). In [147], 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. In [147], it is stated that an unbiased value of N_μ can be obtained, if one corrects for the fact that the muon number for 10^{19} eV showers is 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 considered here are from the time period between 2005 and 2012, which gives a total of 7145 events. The estimated primary energy E_0, angle of incidence θ and value of X_{max}, together with the estimator used to infer N_μ, are at our disposal. This calibration method has not been 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 signal at the ground:

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

S_μ is a fixed mean muon reference signal for proton showers of 10^{19} eV simulated using QGSJETII-03. Consequently, $N_\mu^{\mathrm{n}}(E_0)$ is the relative muon normalization needed to obtain the signal for a primary energy of E_0. Here, E_0 and $N_\mu^{\mathrm{n}}(E_0)$ are the only unknowns and are calculated in [147].

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

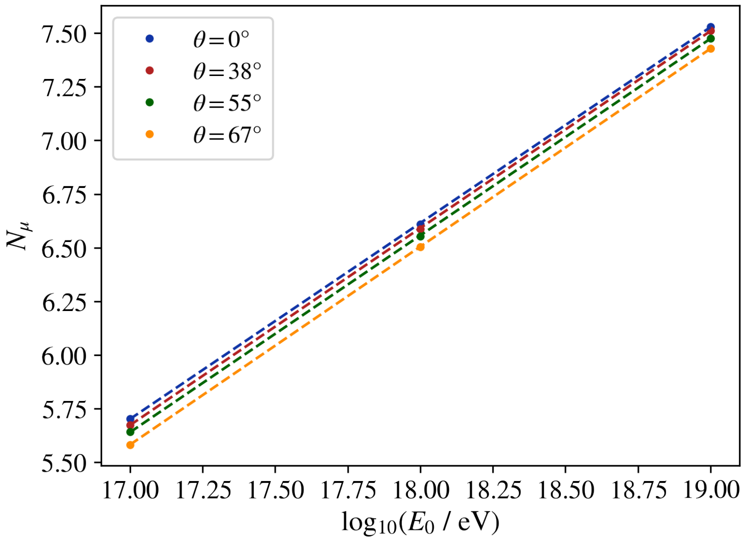

Figure 6.1: Mean number of muons at the ground $\log_{10}(N_\mu)$ as a function of primary energy for different zenith angles, obtained from simulations performed using QGSJETII-03. Knowing the primary energy and angle of incidence of an event, we can extract the corresponding number of muons from the relations presented here. The corresponding value times the normalization factor $N_\mu^{\mathrm{n}}(E_0)$ (given for the dataset) gives the experimental number of muons we need.

between $\log_{10}(E_0/\text{eV}) = 18.098$ and 18.198, the energy range covered by the simulations used to train the final model. With respect to the zenith angle, our simulations are done for $\theta = 38°$. In order to not change the conditions significantly, we keep those showers with zenith angle between $\theta = 34°$ and $40°$. Discarding in addition events for which the calibration did not converge properly, we end up with a dataset comprising 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 $\theta = 38°$, 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 protons is minimized here, as is visible in figure 1.6 [40, 148]. We chose a small interval 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 a mixed composition into our analysis, we repeat the calculations done in section 5.4. This time, we create X_{max}-N_μ distributions for simulations respecting the fractions of primaries valid at the energy range under study. In addition, we apply the smearing, as done in section 5.4. The resulting distributions are shown 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.

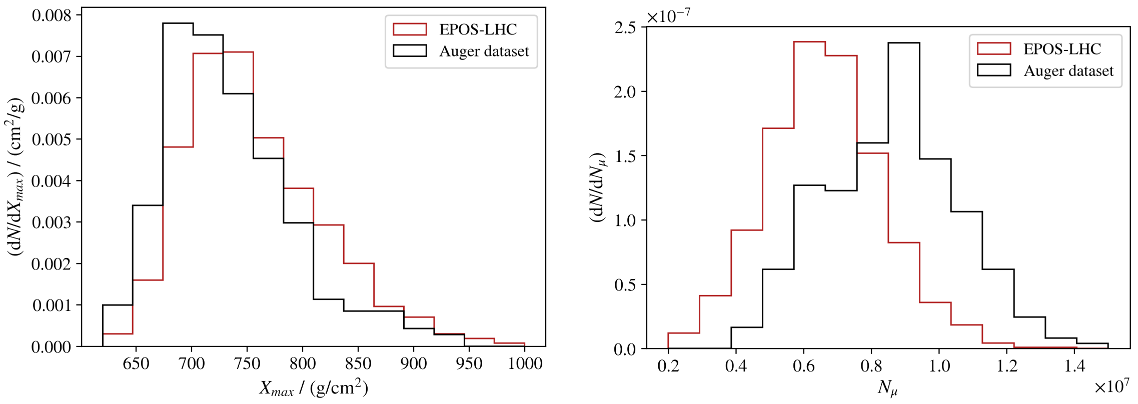

Figure 6.2: X_{max} and N_μ distributions for the Auger dataset. The corresponding distributions for EPOS-LHC simulations of proton primaries with $\theta = 38°$ and $\log_{10}(E_0/\text{eV})$ between 18.098 and 18.198 are added for comparison.

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 those coming from 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 in the first interaction very differently from proton initiated ones. 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 the

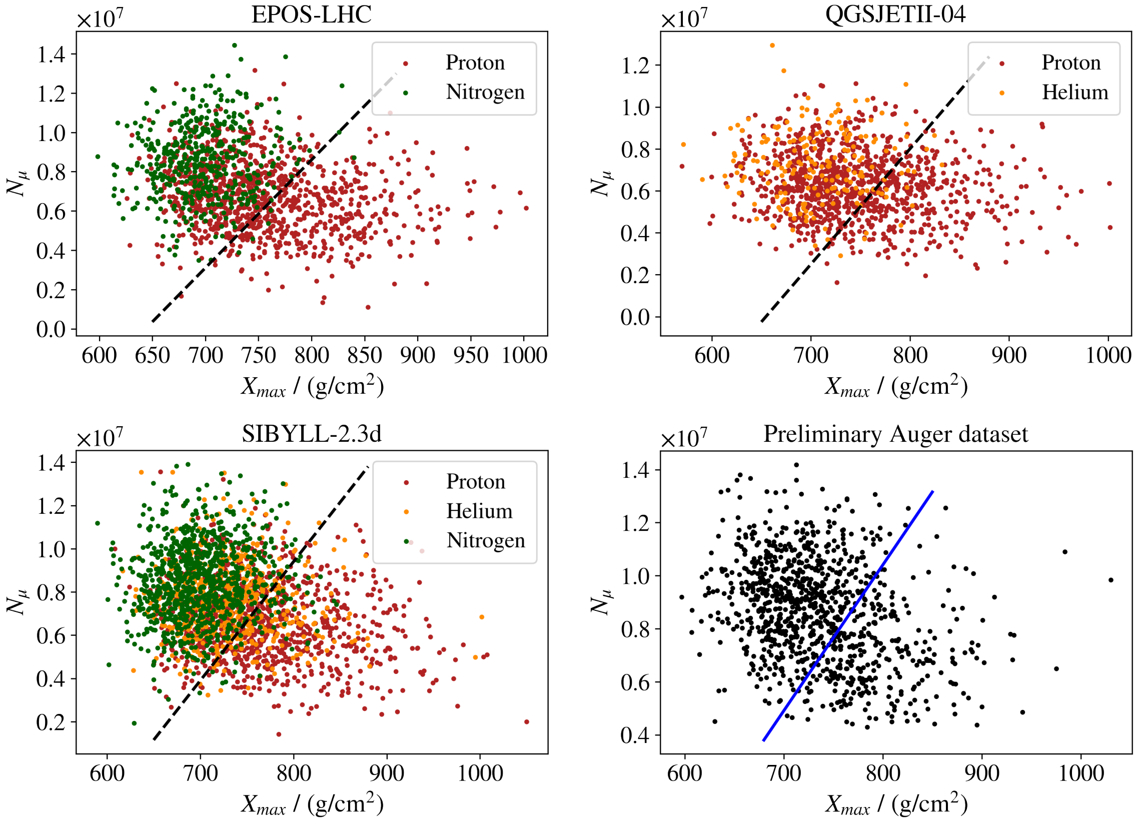

Figure 6.3: $X_{\max}$-N_μ distributions for EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d, after applying an artificial smear in order to mimic the measurement resolution, as described in section 5.4, together with the Auger dataset. The mixed composition is also taken into account for the simulated distributions. For each high-energy interaction model, the fractions of proton (red dots), helium (yellow), nitrogen (green) and iron from [40] 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.

most "contaminated" areas in the $X_{\max}$-N_μ plane. We choose 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, a cut with respect to

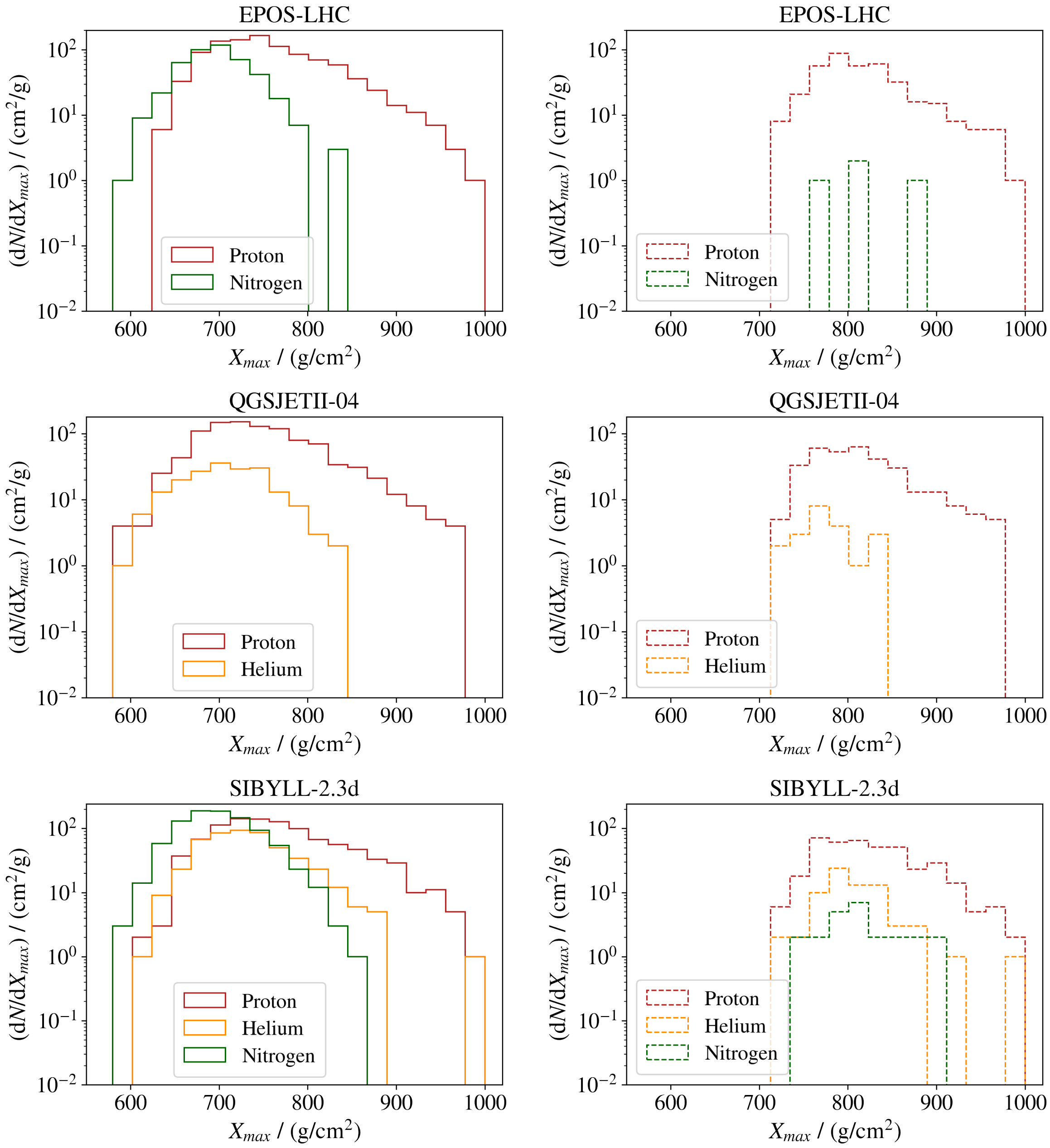

Figure 6.4: $X_{\max}$ distribution for different primaries present in the cosmic ray flux, using the fractions calculated in [40] for the different high-energy interaction models, before (left) and after (right) performing the cut defined in figure 6.3. The axes are fixed at equal values in this comparison, in order to see which portion of the events is selected out when performing the cut.

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 basic parameters (the distributions for all three high-energy interaction models and the Auger dataset) and on the distributions generated by each combination of parameters θ, with which they will be compared in the χ^2-analysis.

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

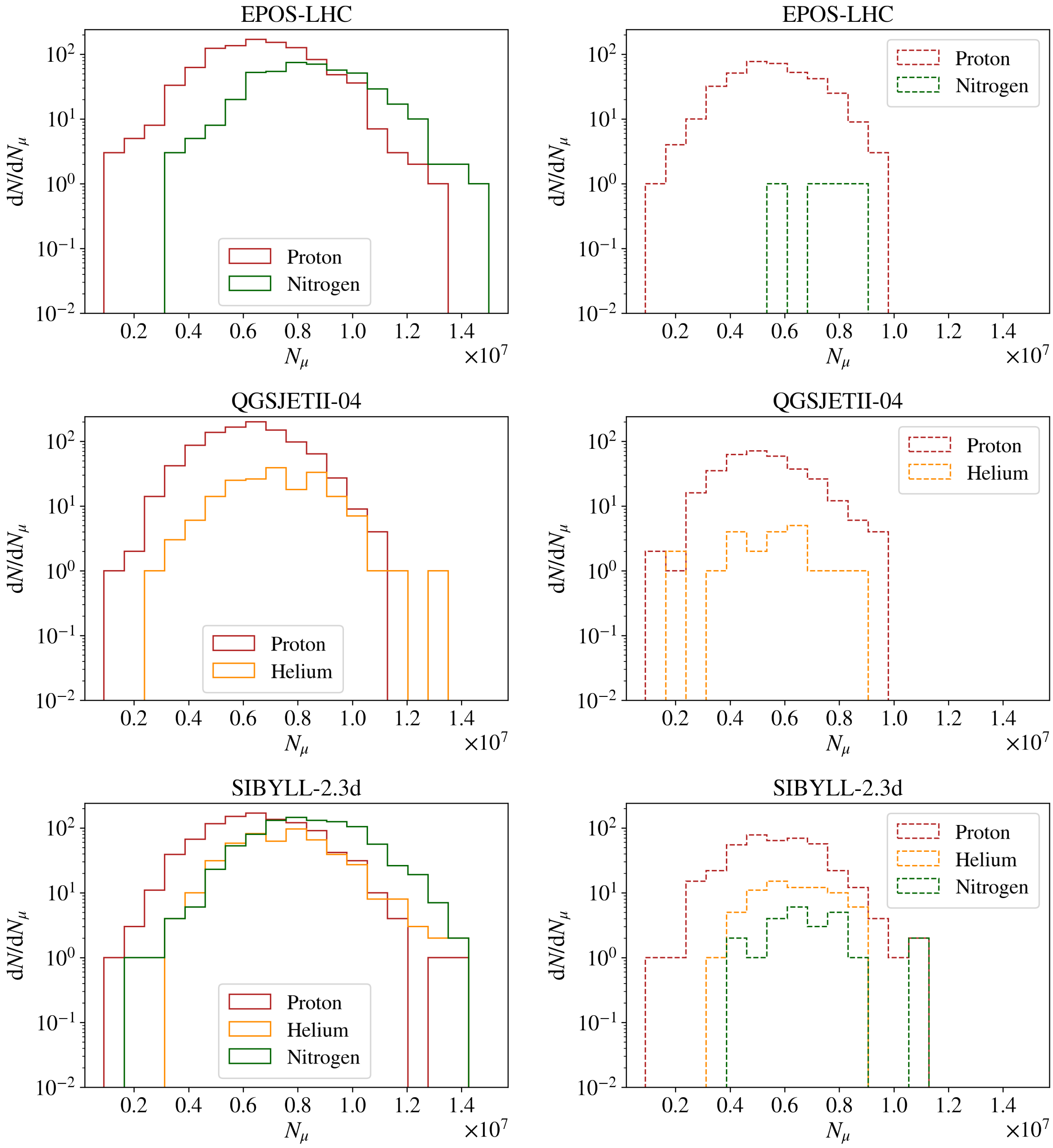

Figure 6.5: N_μ distribution for different primaries present in the cosmic ray flux, using the fractions calculated in [40] for the different high-energy interaction models, before (left) and after (right) performing the cut defined in figure 6.3. The axes are fixed at equal values in this comparison, in order to see which portion of the events is selected out when performing the cut.

simulated distributions have a stronger bias than for the cut corresponding to 30 %. In figures 6.4 and 6.5, we show on the left the complete distributions of $X_{\max}$ and N_μ, respectively. Each color corresponds to a different primary type present for the high-energy interaction model considered (only protons, helium, nitrogen and iron are displayed). The fraction of each particle type is obtained from [40]. 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 using the architecture 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 having a mixed composition and subjected to the cut corresponding to 30 %. We thus allow for the neural network to understand, at least partially, this

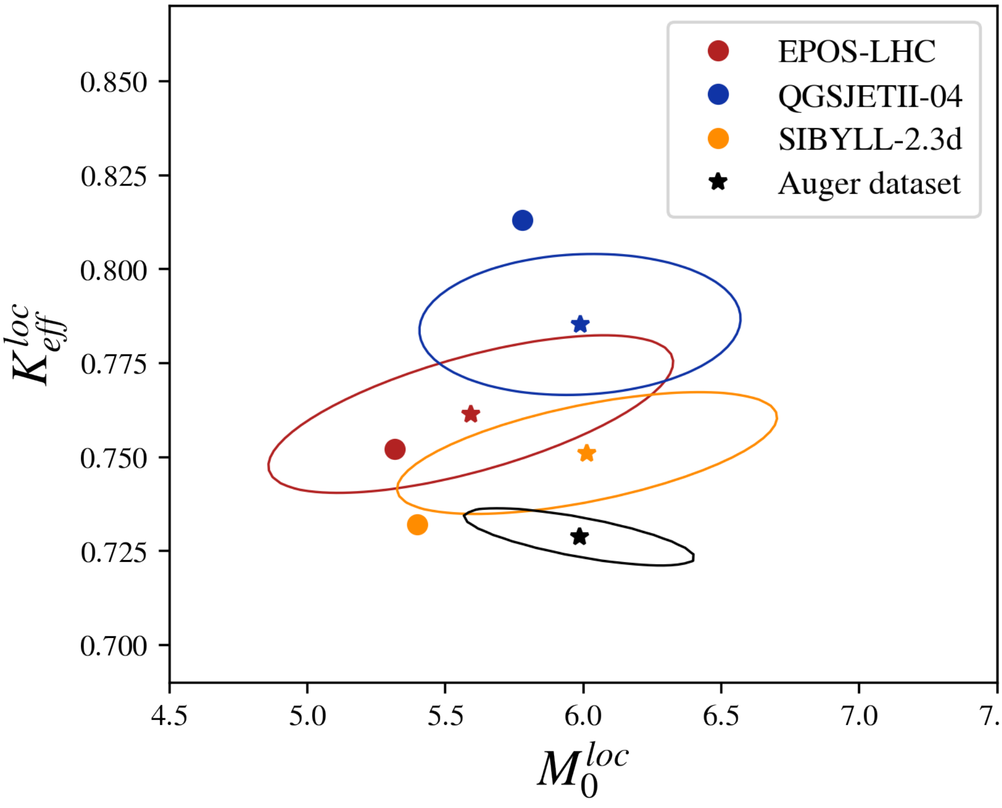

Figure 6.6: Minimum χ^2 fit to obtain the estimated values (stars) of the parameters M_0^{loc} and $K_{\text{eff}}^{\text{loc}}$ for EPOS-LHC (red), QGSJETII-04 (blue) and SIBYLL-2.3d (yellow), as calculated for figure 5.22. In this analysis, we include the effect of having a mixed composition in the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

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 defined 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 a 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 corresponding to the estimation of the parameters for 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 $M_0^{\text{loc}} = 5.99$, $K_{\text{eff}}^{\text{loc}} = 0.73$, $F_0^{\text{loc}} = 0.77$ and $F_0^{\text{scale}} = 0.058$. The distributions of M_0, F_0, K_{eff} and $N_{\text{ch}}^{\text{FI}}$ obtained

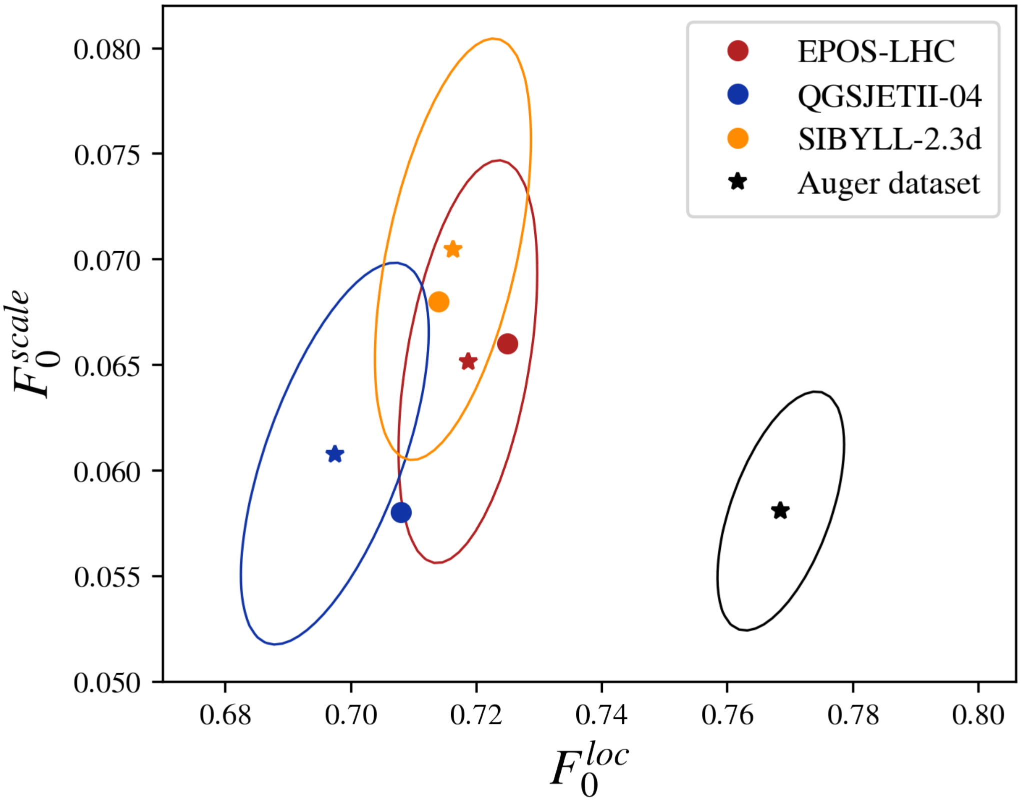

Figure 6.7: Minimum χ^2 fit to obtain the estimated values (stars) of the parameters F_0^{loc} and F_0^{scale} for EPOS-LHC (red), QGSJETII-04 (blue) and SIBYLL-2.3d (yellow), as calculated for figure 5.22. In this analysis, we include the effect of having a mixed composition in the cosmic ray flux.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

from these values are presented in figure 6.8, together with those corresponding to EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d. The distributions of $N_{\rm ch}^{\rm FI}$ for the different cases 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 high-energy interaction models used in the simulations. This makes sense because having a higher hadronic multiplicity at the first interaction implies having more subshowers of less energy. Thus, the overall 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 energy fraction of the first interaction F_0 is higher, as happens for the Auger dataset. Having a higher value of $F_0^{\rm loc}$ than for any high-energy interaction scenario shifts its distribution to higher values. Having a small value of $F_0^{\rm scale}$ concentrates the peak around these higher values. The $K_{\rm eff}$ distribution for the dataset is quite close to those corresponding to SIBYLL-2.3d and EPOS-LHC.

These final results reveal which conditions need to be met in order to solve the Muon

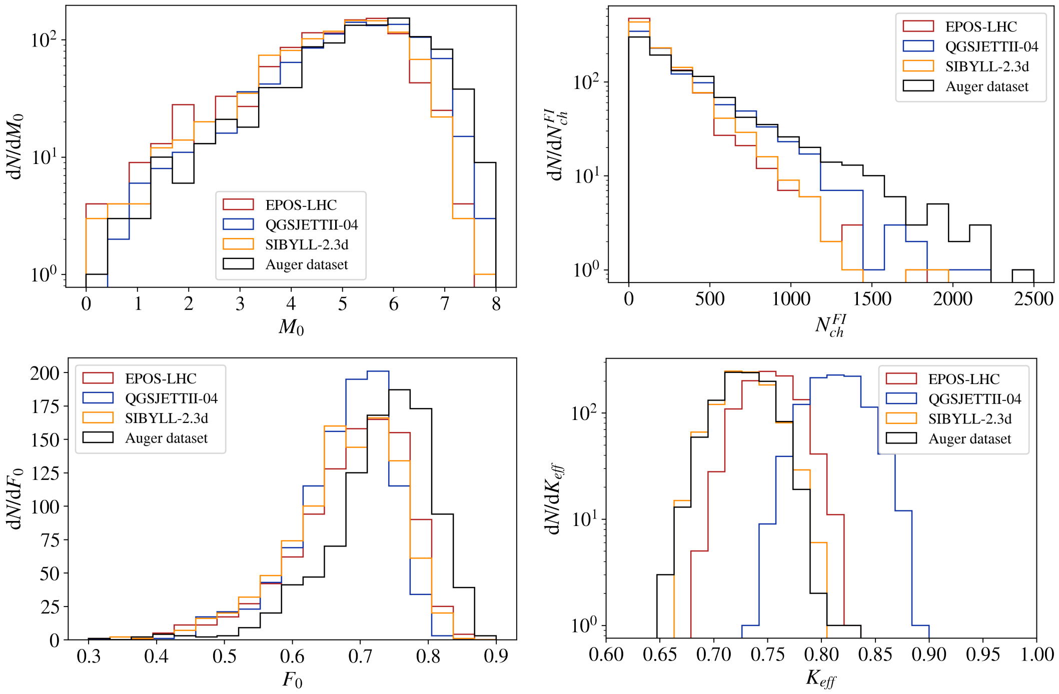

Figure 6.8: Predicted distributions of the unknown features M_0, $N_{\rm ch}^{\rm FI}$, F_0 and $K_{\rm eff}$ for the Auger dataset, obtained using the predicted basic parameters (see figures 6.6 and 6.7). The hadronic multiplicity M_0 tends to be higher than for any of the models used in the simulations.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

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 implementing the *core-corona model* [149]. 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 [150, 151].

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 study 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 describing the inelasticity of the first interaction and a parameter F_0 describing the fraction of energy that is taken by the hadronically interacting particles of the first interaction, together with three effective versions of these parameters representative for the rest of the shower, M_{eff}, K_{eff} and F_{eff}.

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. In order to make sure that this behavior is no coincidence, the distributions of all used parameters are 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 behavior. The three effective parameters 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 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 describing 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 predict $X_{\max}$ and N_μ, separately. A joint mixed ensemble of simulations performed with CONEX using the high-energy interaction models EPOS-LHC, QGSJETII-04 and SIBYLL-2.3d is used for training.

A particular characteristic that is 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 we kept, while the behavior that is common to all three models is included in the final neural network. Furthermore, the errors in the predictions of $X_{\max}$ and N_μ are for all high-energy interaction scenarios below $36.6\,\text{g}\,\text{cm}^{-2}$ and $7.7\,\%$, respectively.

The performance of this model is tested emulating the process that is applied at a final stage to a dataset from Auger. The distributions of the features used for training are described through suitable distributions. Four parameters M_0^{loc}, $K_{\text{eff}}^{\text{loc}}$, F_0^{loc} and F_0^{scale} are sufficient to describe the unknown distributions. A χ^2-analysis is carried out where $\langle X_{\max}\rangle$, $\langle N_\mu\rangle$, $\sigma(X_{\max})$ and $\sigma(N_\mu)$ from the true distributions from simulations are compared to those generated by our model for a grid of possible values of the parameters M_0^{loc}, $K_{\text{eff}}^{\text{loc}}$, F_0^{loc} and F_0^{scale}. Remarkably good results are obtained for the predicted values of the parameters. The predictions of M_0^{loc}, $K_{\text{eff}}^{\text{loc}}$ 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 is applied to a hybrid dataset from Auger in which the number of muons at the ground is provided making use of the concept of universality. Since in a realistic scenario, the cosmic ray flux presents a 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 is solved performing a cut on the distributions, in order to obtain samples with a high proton fraction. Furthermore, the model is retrained on the sets of simulations that respect the estimated fractions of proton, helium, nitrogen and iron. This way, it is possible to obtain predictions of the parameters M_0^{loc}, $K_{\text{eff}}^{\text{loc}}$, F_0^{loc} and F_0^{scale} for the Auger dataset and 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 in the near future 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, the hadronic multiplicity of the first interaction and the effective inelasticity are inferred for a preliminary dataset from the Pierre Auger Observatory. The first two distributions represent highly interesting properties related to high-energy hadronic interactions that cannot be studied at accelerators. It is the first time that such distributions are inferred that show hadronic properties at energies 10 times higher than the highest ones 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 energy 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, 43

[2] Kampert, K.-H., Unger, M. Measurements of the cosmic ray composition with air shower experiments. *Astroparticle Physics*, **35** (10), 660–678, 2012. URL `http://dx.doi.org/10.1016/j.astropartphys.2012.02.004`. vii, 28, 38, 39, 40, 43, 45

[3] Hess, V. On the observations of the penetrating radiation during seven balloon flights. *arXiv preprint arXiv:1808.02927*, 2018. 1

[4] Anderson, C. D. The positive electron. *Phys. Rev.*, **43**, 491–494, 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**, 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, 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. *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, 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, 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, 1949. URL https://link.aps.org/doi/10.1103/PhysRev.75.581. 3

[13] Fretter, W. B. Penetrating showers. *Phys. Rev.*, **76**, 511–517, 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, 1950. URL https://link.aps.org/doi/10.1103/PhysRev.78.290. 3

[17] Rossi, B. Misure sulla distribuzione angolare di intensita 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, 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, 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, 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, 1964. URL https://doi.org/10.1143/PTPS.32.82. 5

[25] Aab, A., Abreu, P., Aglietta, M. Pierre Auger Collaboration. *Phys. Rev. D*, **91**, 2015. 5, 13

[26] Shellard, R. C. First results from the Pierre Auger Observatory. *Brazilian journal of physics*, **36** (4A), 2006. 5, 14

[27] The Pierre Auger Cosmic Ray Observatory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **798**, 172–213, 2015. URL https://www.sciencedirect.com/science/article/pii/S0168900215008086. 5, 15, 16, 17

[28] Mollerach, S., Roulet, E. Progress in high-energy cosmic ray physics. *Progress in Particle and Nuclear Physics*, **98**, 85–118, 2018. URL https://www.sciencedirect.com/science/article/pii/S0146641017300881. 6

[29] Kampert, K.-H., Collaboration, K.-G. Cosmic rays in the knee-region - recent results from KASCADE. *arXiv preprint astro-ph/0405608*, 2004. 6

[30] 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), 2011. URL http://dx.doi.org/10.1103/PhysRevLett.107.171104. 7

[31] Hillas, A. M. The origin of ultra-high-energy cosmic rays. *Annual review of astronomy and astrophysics*, **22** (1), 1984. 7, 8, 15

[32] Pierre Auger Collaboration, *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

[33] 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**, 2019. URL http://dx.doi.org/10.3389/fspas.2019.00023. 7, 8

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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), 2017. URL http://dx.doi.org/10.1088/1475-7516/2017/04/038. 8

[39] 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), 2015. URL http://dx.doi.org/10.1088/0004-637X/804/1/15. 9

[40] Bellido, J., Pierre Auger Collaboration, *et al.* Depth of maximum of air-shower profiles at the Pierre Auger Observatory: Measurements above $10^{17.2}$ eV and composition implications. *35th International Cosmic Ray Conference*, **301**, 2018. 10, 11, 103, 105, 106, 107, 108, 109

[41] 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. 12

[42] 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

[43] 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

[44] 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), 2008. URL http://dx.doi.org/10.1016/j.nima.2007.12.016. 13, 17

[45] 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), 2010. URL http://dx.doi.org/10.1016/j.nima.2010.04.023. 14

[46] 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

[47] Greisen, K. The extensive air showers. *Progress in cosmic ray physics*, **3** (1), 1956. 15

[48] Newton, D., Knapp, J., Watson, A. The optimum distance at which to determine the size of a giant air shower. *Astroparticle Physics*, **26** (6), 2007. URL http://dx.doi.org/10.1016/j.astropartphys.2006.08.003. 15

[49] Schulz, A. The measurement of the cosmic ray spectrum above 3×10^{17} eV with the Pierre Auger Observatory. *International Cosmic Ray Conference*, **33**, 2013. 15

[50] 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

[51] Meurer, C., Scharf, N. on behalf of the Pierre Auger Collaboration. HEAT – a low energy enhancement of the Pierre Auger Observatory. *Astrophysics and Space Sciences Transactions*, **7** (2), 183–186, 2011. URL http://dx.doi.org/10.5194/astra-7-183-2011. 17

[52] 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

[53] 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

[54] 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

[55] Stasielak, J. AugerPrime–the upgrade of the Pierre Auger Observatory. *arXiv preprint arXiv:2110.09487*, 2021. 18, 20

[56] 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

[57] Gora, D., Pierre Auger Collaboration, *et al.* The Pierre Auger Observatory: Review of latest results and perspectives. *Universe*, **4** (11), 2018. 19

[58] Suarez, F. The AMIGA muon detectors of the Pierre Auger Observatory: Overview and status. *Proceedings, 33rd International Cosmic Ray Conference (ICRC2013)*, **712**, 2013. 19

[59] Müller, S. N. Measurement of the cosmic ray composition with air showers detected by the AMIGA extension at the Pierre Auger Observatory, 2019. 19

[60] 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

[61] Pierog, T., Alekseeva, M., Bergmann, T., Chernatkin, V., Engel, R., Heck, D., *et al.* First results of fast one-dimensional hybrid simulation of extensive air showers using CONEX. *arXiv preprint astro-ph/0411260*, 2004. 21, 25

[62] 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

[63] Heitler, W. The quantum theory of radiation, Second edition. International Series of Monographs on Physics. Oxford University Press, 1936. 21, 28, 29

[64] Matthews, J. Energy flow in extensive air showers. *Proceedings of ICRC*, **2001** (261), 2001. 21, 28, 30, 34, 40

[65] Matthews, J. A Heitler model of extensive air showers. *Astropart. Phys.*, **22**, 387–397, 2005. 21, 28, 30, 40

[66] Cazon, L. Working group report on the combined analysis of muon density measurements from eight air shower experiments. *arXiv preprint arXiv:2001.07508*, 2020. 21, 48, 49

[67] Cazon, L. Probing high-energy hadronic interactions with extensive air showers. *arXiv preprint arXiv:1909.02962*, 2019. 21, 32, 33, 54, 58, 86

[68] 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. *EPJ Web of Conferences*, **210**, 02006, 2019. URL http://dx.doi.org/10.1051/epjconf/201921002006. 21, 54, 58

[69] Boezio, M., Mocchiutti, E. Chemical composition of galactic cosmic rays with space experiments. *Astroparticle Physics*, **39-40**, 95–108, 2012. URL http://dx.doi.org/10.1016/j.astropartphys.2012.05.007. 21

[70] Kampert, K.-H., Unger, M. Measurements of the cosmic ray composition with air shower experiments. *Astroparticle Physics*, **35**, 660–678, 2012. 21

[71] Zyla, P., *et al.* Review of Particle Physics. *PTEP*, **2020** (8), 083C01, 2020. 22, 33, 34

[72] Grieder, P. Extensive air showers. High energy phenomena and astrophysical aspects. A tutorial, reference manual and data book. vol. I and II. *EAS Publications Series*, 2010. 23

[73] Kobal, M., *et al.* A thinning method using weight limitation for air-shower simulations. *Astroparticle Physics*, **15** (3), 2001. 26

[74] Risse, M., Heck, D., Knapp, J. EAS simulations at Auger energies with CORSIKA. *International Cosmic Ray Conference*, **2**, 2001. 26

[75] 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

[76] Pierog, T., Engel, R., Heck, D., Poghosyan, G. Future of Monte Carlo simulations of atmospheric showers. *EPJ Web of Conferences*, **89**, 2015. 26

[77] Pierog, T., Engel, R., Heck, D., Ulrich, R. 3D hybrid air shower simulation in CORSIKA. *Proc. of the International Cosmic Ray Conference*, 2011. 26

[78] Pierog, T., Engel, R., Heck, D. 3D air shower simulations using CONEX in CORSIKA. *Contributions to the 31st International Cosmic Ray Conference, Poland, 2009*, 2009. 26

[79] 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. *29th International Cosmic Ray Conference (ICRC29), Volume 7*, **7**, 2005. 26

[80] Kalmykov, N., Ostapchenko, S., Werner, K. Results of EAS characteristics calculations in the framework of the universal hadronic interaction model NEXUS. *International Cosmic Ray Conference*, **2**, 2001. 26

[81] 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

[82] 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

[83] 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

[84] 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

[85] Pierog, T., Werner, K. EPOS model and ultra high energy cosmic rays. *Nuclear Physics B-Proceedings Supplements*, **196**, 2009. 27

[86] 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

[87] Ostapchenko, S. Total and diffractive cross sections in enhanced Pomeron scheme. *Physical Review D*, **81** (11), 2010. 27

[88] Ostapchenko, S. On the re-summation of enhanced Pomeron diagrams. *Physics Letters B*, **636** (1), 2006. 27

[89] Ostapchenko, S. Nonlinear screening effects in high energy hadronic interactions. *Physical Review D*, **74** (1), 2006. 27

[90] 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

[91] 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

[92] 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

[93] Thakuria, C. C., Boruah, K. Comparison of EPOS and QGSJET-II in EAS simulations using CORSIKA. *arXiv preprint arXiv:1202.3661*, 2012. 27

[94] Linsley, J. Structure of large air showers at depth 834 g/cm^2 III - applications. *International Cosmic Ray Conference*, **12**, 89–96, 1977. 33

[95] Pierog, T. Hadronic interactions and air showers: Where do we stand? *EPJ Web Conf.*, **208**, 2019. 34, 36, 37, 46, 47

[96] 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

[97] Adam, J., Adamova, D., Aggarwal, M. M., Rinella, G. A., Agnello, M., Agrawal, N., *et al.* Measurement of pion, kaon and proton production in proton–proton collisions at $\sqrt{s} = 7$ TeV. *The European Physical Journal C*, **75** (5), 1–23, 2015. 36

[98] Gaisser, T. K., Engel, R., Resconi, E. Cosmic Rays and Particle Physics. Cambridge University Press, 2016. 37, 48

[99] Veberič, D. Lambert W function for applications in physics. *Computer Physics Communications*, **183** (12), 2622–2628, 2012. URL http://dx.doi.org/10.1016/j.cpc.2012.07.008. 39

[100] Abu-Zayyad, T., Belov, K., Bird, D., Boyer, J., Cao, Z., Catanese, M., *et al.* Evidence for changing of cosmic ray composition between 10^{17} and 10^{18} eV from multicomponent measurements. *Physical Review Letters*, **84**, 4276–4279, 2000. 47

[101] 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), 032003, 2015. 47, 52

[102] Dembinski, H. Computing mean logarithmic mass from muon counts in air shower experiments. *Astroparticle Physics*, **102**, 89–94, 2018. URL http://dx.doi.org/10.1016/j.astropartphys.2018.05.008. 48

[103] 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 10^{11} GeV. *arXiv preprint arXiv:1711.11432*, 2017. 48

[104] 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, 2006. URL http://dx.doi.org/10.1007/s10582-006-0156-9. 50

[105] 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

[106] Drescher, H., Hladik, M., Ostapchenko, S., Pierog, T., Werner, K. Parton-based Gribov–Regge theory. *Physics Reports*, **350** (2-4), 93–289, 2001. URL http://dx.doi.org/10.1016/S0370-1573(00)00122-8. 50

[107] 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, 1994. URL https://link.aps.org/doi/10.1103/PhysRevD.50.5710. 50

[108] Engel, J., Gaisser, T. K., Lipari, P., Stanev, T. Nucleus-nucleus collisions and interpretation of cosmic-ray cascades. *Phys. Rev. D*, **46**, 5013–5025, 1992. URL https://link.aps.org/doi/10.1103/PhysRevD.46.5013. 50

[109] Glauber, R. J., Matthiae, G. High-energy scattering of protons by nuclei. *Nuclear Physics*, **21**, 135–157, 1970. 50

[110] 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 $\sqrt{s} = 13$ TeV. *Physics Letters B*, **751**, 143–163, 2015. 50

[111] 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. *Astrophysics and Space Science*, **367** (3), 27, 2022. 51, 52

[112] Ulrich, R., Engel, R., Unger, M. Hadronic multiparticle production at ultrahigh energies and extensive air showers. *Physical Review D*, **83** (5), 2011. URL http://dx.doi.org/10.1103/PhysRevD.83.054026. 51, 93, 97

[113] 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. 52, 53

[114] 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. 53

[115] ALICE Collaboration. Enhanced production of multi-strange hadrons in high-multiplicity proton–proton collisions. *Nature Physics*, **13** (6), 535–539, 2017. 53

[116] Ostapchenko, S. QGSJET: Physics, recent improvements, and results for air showers. *EPJ Web of Conferences*, **52**, 2013. 53

[117] Pierog, T., Werner, K. Muon production in extended air shower simulations. *Phys. Rev. Lett.*, **101**, 171101, 2008. URL `https://link.aps.org/doi/10.1103/PhysRevLett.101.171101`. 53

[118] 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. *Physics Letters B*, **784**, 68–76, 2018. URL `http://dx.doi.org/10.1016/j.physletb.2018.07.026`. 54, 58

[119] Cazon, L., Conceição, R., Martins, M., Riehn, F. Probing the high energy spectrum of neutral pions in ultra-high energy proton-air interactions. *arXiv preprint arXiv:1908.09668*, 2019. 54, 58

[120] Cazon, L., Conceição, R., Martins, M. A., Riehn, F. Constraining the energy spectrum of neutral pions in ultra-high-energy proton-air interactions. *Physical Review D*, **103** (2), 2021. URL `http://dx.doi.org/10.1103/PhysRevD.103.022001`. 54, 58

[121] 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

[122] Jones, E., Oliphant, T., Peterson, P., *et al.* Scipy: Open source scientific tools for Python, 2001. 71

[123] 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

[124] Chollet, F., *et al.* Keras, 2015. 71

[125] 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

[126] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., *et al.* Tensorflow: A system for large-scale machine learning. *12th symposium on operating systems design and implementation 16*, 2016. 71

[127] Van Rossum, G., *et al.* Python programming language. *USENIX annual technical conference*, **41**, 2007. 71

[128] 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), 2019. URL `http://dx.doi.org/10.1103/RevModPhys.91.045002`. 72

[129] A living review of machin learning for paticle physics. 72

[130] Banko, M., Brill, E. Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, págs. 26–33, 2001. 73

[131] Smith, L. N. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018. 76

[132] Ge, R., Huang, F., Jin, C., Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. *Conference on learning theory*, 2015. 76

[133] Glorot, X., Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010. 77

[134] 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

[135] Teodorescu, L. Artificial neural networks in high-energy physics, 2008. 77

[136] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*, **61**, 2015. URL `http://dx.doi.org/10.1016/j.neunet.2014.09.003`. 77

[137] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. Classification and regression trees. Belmont, CA: Wadsworth International Group, 1984. 85

[138] Breiman, L. Bagging predictors. *Machine learning*, **24** (2), 123–140, 1996. 85

[139] Breiman, L. Machine learning, volume 45, number 1 - springerlink. *Machine Learning*, **45**, 5–32, 2001. 85

[140] Bergstra, J., Bengio, Y. Random search for hyper-parameter optimization. *Journal of machine learning research*, **13** (2), 2012. 86

[141] Davison, A. C., Hinkley, D. V. Bootstrap methods and their application. Cambridge university press, 1997. 94

[142] 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

[143] 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

[144] 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

[145] Lista, L. Statistical methods for data analysis in particle physics. Springer, 2017. 99

[146] 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

[147] 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

[148] 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

[149] 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

[150] Pierog, T., Guiot, B., Karpenko, I., Sophys, G., Stefaniak, M., Werner, K. EPOS 3 and air showers. *EPJ Web of Conferences*, **210**, 2019. 112

[151] 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 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 Belén Andrada, Michael Unger, Matias Perlin and Ana Botti for all the great discussions. I would also like to thank Sabine Bucher and Marie-Christine Mundt, for helping me with the bureaucracy.

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, Kathrin, Martin, Juli, Mariza, Edu, 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!