GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Fakultät für Physik

Master's Thesis

Untersuchung von schnellen Hadronenschauersimulationsmethoden mit dem CALICE AHCAL Prototyp

Investigation of Fast Hadron Shower Simulation Methods with the CALICE AHCAL Prototype

prepared by

André Wilhahn

from Eschwege

at the II. Physikalischen Institut

Thesis number:	II.Physik-UniGö-MSc-2022/03
Thesis period:	15th October 2021 until 10th October 2022
First referee:	Prof. Dr. Stan Lai
Second referee:	Priv. Doz. Dr. Jörn Große-Knetter

Abstract

Diese Masterarbeit beschäftigt sich mit der Entwicklung und Untersuchung von schnellen Simulationsmethoden von hadronischen Teilchenschauern. Für diesen Zweck wurden Teststrahldaten der AHCAL-Gruppe der CALICE-Kollaboration verwendet. Diese wurden in drei Phasen im Jahr 2018 an der Europäischen Organisation für Kernforschung (CERN) mit dem technologischen Detektorprototyp der AHCAL-Gruppe aufgenommen. Während dieser Datenaufnahme wurde der Detektorprototyp Elektronen-, Myonen- und negativ geladenen Pionenstrahlen verschiedenster Energien ausgesetzt. Für diese Masterarbeit wurde der gesamte Pionendatensatz verwendet.

Die in dieser Arbeit präsentierte Simulation ist datenbasiert. Es wurden zunächst longitudinale Energiedifferenzen zwischen einzelnen Pionenschauern und einer longitudinalen Parametrisierung durchschnittlicher Pionenschauer berechnet. Danach wurden zwei verschiedene Methoden für die Simulation besagter Energiedifferenzen analysiert: Einerseits wurden Schauer mittels einer Hauptkomponentenanalyse und andererseits mittels Kerndichteschätzern simuliert. Des Weiteren wurden, basierend auf den Ergebnissen der Kerndichteschätzung, Interpolationen für simulierte longitudinale Energieverteilungen zwischen verschiedenen Strahlenergien durchgeführt.

Abstract

This thesis presents the development and investigation of fast hadron shower simulation methods. For this purpose, a test beam dataset of the AHCAL group of the CALICE Collaboration has been used. This dataset was recorded in 2018 at CERN, where the AHCAL Technological Detector Prototype was exposed to electron, muon, and negatively charged pion beams of various initial energies. For this thesis, only the pion dataset has been used.

The fast simulation presented in this thesis is a data-driven simulation. For the simulation of hadronic showers, differences in longitudinal energy distributions between single pion showers and an average pion shower parameterisation were calculated. Then, two different shower simulation methods were analysed: On the one hand, a principal component analysis was conducted, and on the other hand, kernel density estimators were applied to data. Furthermore, based on the results of the kernel density estimation, interpolations of simulated longitudinal energy distributions were implemented between different initial energies.

Contents

Nomenclature

Terminologies

Abbreviation	Meaning
COGZ	Centre of Gravity in the z-Direction
DCR	Dark Count Rate
HBU	HCAL Base Unit
KDE	Kernel Density Estimator
MIP	Minimal Ionising Particle
PCA	Principal Component Analysis
PCT	Principal Component Transformation
PDF	Probability Density Function
QED	Quantum Electrodynamics
SAPD	Single-Photon Avalanche Photodiode
SiPM	Silicon Photomultiplier
SM	Standard Model of Particle Physics

1. Introduction

With the beginning of the 20th century, an epoch of great advances and revolutionary ideas in the field of physics began. The newly discovered theory of quantum mechanics and its mathematical framework, for the first time described by physicists such as Max Planck, Niels Bohr, and many more, as well as Albert Einstein's General Theory of Relativity [1] shaped our understanding of nature like never before. Today, General Relativity is one of two theories that lay the foundation upon which modern science, and in particular physics, rests, describing gravity on large, macroscopic scales.

While the pure geometric theory of General Relativity has not been changed since its first publication in 1915, quantum mechanics has been continuously expanded upon and improved over the course of the last century, until it became what is today known as the Standard Model of particle physics (SM). The SM is a quantum field theory whose laws govern the behaviour of matter particles (fermions) and force carrier particles (bosons) on microscopic scales. Three of the four fundamental forces are incorporated into the SM, which includes the strong nuclear force described by quantum chromodynamics [2, 3] as well as the weak nuclear force and electromagnetism, which have been unified into the theory of electroweak interactions [4–6]. Together with the Brout-Englert-Higgs mechanism [7–10], which generates the masses of all massive SM constituents, many predictions of the SM have been proven to be correct, the last major one being the existence of the Higgs boson, which was discovered by the ATLAS and CMS Collaborations in 2012 at CERN [11, 12].

In order to detect particles such as the Higgs boson, and to claim a discovery, large and highly granular particle detectors are crucial for high energy physics experiments. However, not all SM particles can be directly detected. The majority of all SM constituents is unstable, and these particles decay before reaching the detector. Hence, physicists have to infer properties of unstable particles through their decay products, which is usually done in two ways: On the one hand, one can monitor trajectories of electrically charged particles with tracking chambers. On the other hand, energies of electromagnetically and strongly interacting particles can be measured with electromagnetic and hadronic calorimeters, respectively, which allows to deduce from which parent particle detected

particles might have originated.

The principle of calorimeters is always the same: An incoming particle interacts with the detector material, producing secondary particles that scatter within the calorimeter as well. These secondary particles are copiously produced and therefore create a cascade within the detector, called a particle shower, whose energy is then measured. Measuring the energy of a shower is a disruptive measurement because the shower deposits most or all of the initial particle's energy within the calorimeter. After deposition, this energy is first transformed into electric signals and after that into data, which contains valuable information about the nature of particle interactions. By evaluating it, physicists are able to put theoretical predictions of the SM to the test.

Predictions about the behaviour of particles are obtained by analysing the equations of motion of the SM numerically. Such computations allow scientists to predict, for instance, the distribution of energy deposition of a shower within the detector. However, simulating particle showers is often a very time- and computing-power-consuming task, since all particle interactions at the relevant energy scale must be considered. For higher energies, simulations therefore become more and more complex. A possibility to circumvent this problem is to build so called fast simulations. The aim of fast simulations is to reduce the amount of required computing resources significantly, while preserving as much information about the shower as possible. Fast simulations become continuously more important as physicists are currently exploring particle interactions at energies higher than ever before. This Master's thesis describes the fast simulation of hadronic showers, in particular pion showers, and compares its results to test beam data. The dataset that has been used for this comparison was recorded by the AHCAL group of the CALICE Collaboration during a test beam run in 2018 at CERN.

This thesis is structured as follows. Chapter 2 begins with a theoretical overview of electromagnetic and hadronic showers. This includes their longitudinal and radial shapes, the development of electromagnetic subshowers within hadronic showers, and the calorimeter responses to each type of shower. Following this, Chapter 3 introduces the CALICE Collaboration and the AHCAL detector prototype. In addition, a motivation underlining the importance of fast simulations, is given. This Chapter also discusses two large test beam campaigns which took place in 2018 and 2022, respectively, at CERN in Geneva. In Chapter 4, simulations obtained by applying a principal component analysis to data are presented. Chapter 5, on the other hand, presents simulation results of applying kernel density estimators to data. After this, results of interpolating simulated longitudinal energy distributions between different initial pion energies are shown in Chapter 6. In the end, a conclusion and an outlook are given in Chapter 7.

2. Theoretical Background

Particle showers exhibit distinct shapes and develop differently within a particle detector, depending on whether they originate from electromagnetic or strong (hadronic) processes. This Chapter therefore introduces the theory behind particle showers, first for electromagnetic showers in Section 2.1, and then for those emerging from hadronic interactions in Section 2.2. For both shower types, their developments as well as their characteristic longitudinal and radial shapes are covered. Furthermore, Section 2.2.2 also discusses the production of electromagnetic subshowers within hadronic showers and how differently calorimeters respond to each type of shower.

2.1. Electromagnetic Showers

2.1.1. Development of Electromagnetic Showers

The development of electromagnetic showers is governed by the laws of quantum electrodynamics (QED). QED describes the behaviour and interactions between electrically charged SM constituents and photons, the electrically neutral force carrier of QED. The charges of electrically charged particles are usually given in units of the elementary charge, e, and only differ by sign between particles and their corresponding antiparticles. Though there are many elementary particles that can interact electromagnetically, not all of them can also be part of an electromagnetic shower. In almost all cases, such a shower only comprises electrons, positrons, or photons [13], since only these particles deposit enough energy within an electromagnetic calorimeter to be detected. All remaining electromagnetically interacting particles are usually not detected because they decay before entering the detector (tau leptons), are mainly detected by hadronic calorimeters (quarks/hadrons), or deposit too little energy within electromagnetic calorimeters (muons).

At high energies, electrons and positrons lose energy via emission of bremsstrahlung, due to their small masses. Photons, on the other hand, undergo electron-positron pair production and thereby split their energy between the decay products. These two processes, however, are only dominant above an energy threshold called the critical energy

(per particle), E_c. Below the critical energy, ionisation becomes dominant for charged particles. Non-linear processes are involved in electromagnetic shower development too, for example multiple scattering, but they will not be discussed in this Chapter.

The size of an electromagnetic shower is commonly quantified by the mean free path of a particle within the detector material, that is to say, the distance after which an electron or a positron emits bremsstrahlung. This quantity is called the radiation length and is denoted as X_0. It is a material-dependent constant, and generally one finds that $X_0 \sim \frac{1}{Z^2}$ [14], where Z is the atomic number of the detector material. This proportionality implies that electromagnetic showers are on average shorter in denser detector materials.

The length of an electromagnetic shower also depends on the initial energy, E_0, of the incoming particle that initiates the shower. This can be shown using a simplified model of electromagnetic shower development. Suppose an incoming particle, for example an electron, enters an electromagnetic calorimeter. After one radiation length, on average, the electron will emit a photon, after two radiation lengths another, etc. Likewise, all emitted photons will eventually decay into another electron and a positron. In reality, this happens after one photon absorption length, λ_γ, the mean free path of a photon, which differs from the radiation length and can be approximated via [13, 15]:

$$\lambda_\gamma \approx \frac{9}{7} X_0 \,. \tag{2.1}$$

For this simplified model, though, $\lambda_\gamma = X_0$ is assumed. This results in an electromagnetic shower as schematically shown in Figure 2.1, which comprises on average 2^n particles after n radiation lengths.

When the critical energy is reached after n_{tot} radiation lengths, one finds that $E_c = \frac{E_0}{2^{n_{\text{tot}}}}$. The shower's total length is hence given by:

$$s_{\text{tot}} = n_{\text{tot}} X_0 = \frac{\log\left(\frac{E_0}{E_c}\right)}{\log(2)} X_0 \,. \tag{2.2}$$

Thus, one finds that an electromagnetic shower's length grows logarithmically as a function of the initial energy. Due to this, sizes of electromagnetic calorimeters grow logarithmically too, which makes calorimetry particularly attractive for high energy physics experiments, since a calorimeter with fixed size is able to detect particle showers with energies ranging between several orders of magnitude.

This simplified model of electromagnetic shower development already describes most general properties of electromagnetic showers but not all of them. The assumptions upon which the simplified model is based (energy loss via ionisation is energy-independent,

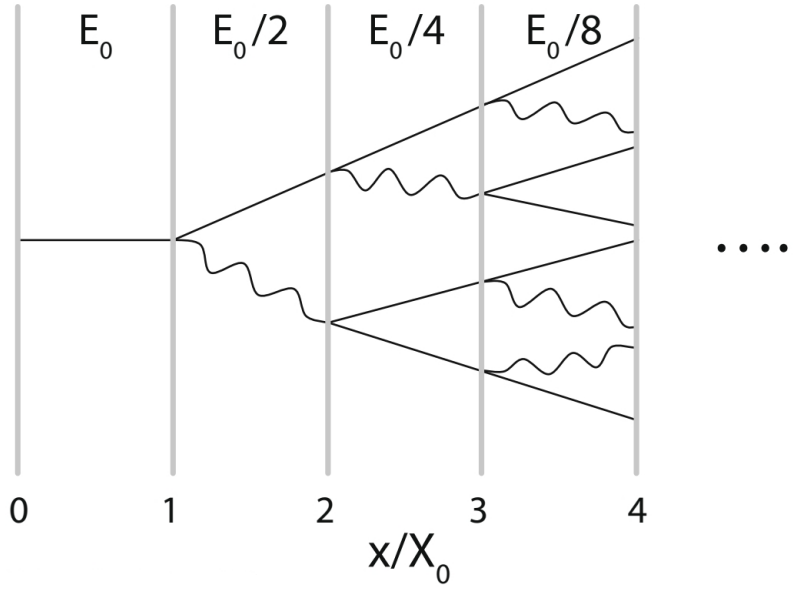


Figure 2.1.: Schematic depiction of an electromagnetic shower with radiation length X_0 [13]. After each radiation length the number of particles doubles and the energy per particle is halved. Electrons and positrons are indicated as straight lines and photons as waves.

neglected multiple scattering, one-dimensional shower development, etc.) no longer hold when dealing with real electromagnetic showers. For this reason, the following Section introduces an empirical model of electromagnetic shower development that aims at encapsulating all relevant processes and dependencies.

2.1.2. Longitudinal and Radial Shapes of Electromagnetic Showers

The previously presented simplified model is useful for understanding electromagnetic shower development conceptually. However, in order to predict energy distributions (both longitudinal as well as radial) of electromagnetic showers within an electromagnetic calorimeter, empirical models are necessary, particularly because the mean free path of a photon differs from the radiation length of an electron or a positron.

For the longitudinal development of electromagnetic showers, the following formula is commonly used for parameterising longitudinal energy distributions [16]:

$$\frac{\mathrm{d}E}{\mathrm{d}n} = E_0 \frac{b^a}{\Gamma(a)} n^{a-1} e^{-bn} \,. \tag{2.3}$$

Here, the left-hand side is the change in energy, E, over the number of radiation lengths,

n. On the right-hand side, E_0 is the initial energy of the shower, and Γ is the gamma function defined as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t \,. \tag{2.4}$$

Moreover, a and b are two empirical parameters that determine the position of the shower maximum in numbers of radiation lengths:

$$n_{\max} = \frac{a-1}{b} \,. \tag{2.5}$$

Longitudinal energy distributions of electromagnetic showers described by Equation (2.3) look like those shown in Figure 2.2. This Figure shows longitudinal energy distributions, initiated by electrons, for four different initial energies within a block of copper, where the energy deposition per centimetre is given as a function of the shower depth. All distributions exhibit a steep increase in deposited energy, followed by the shower maximum and a slow exponential decrease. Note that the position of the shower maximum shifts only slowly to larger depths, even though the initial energy is increased by orders of magnitude.

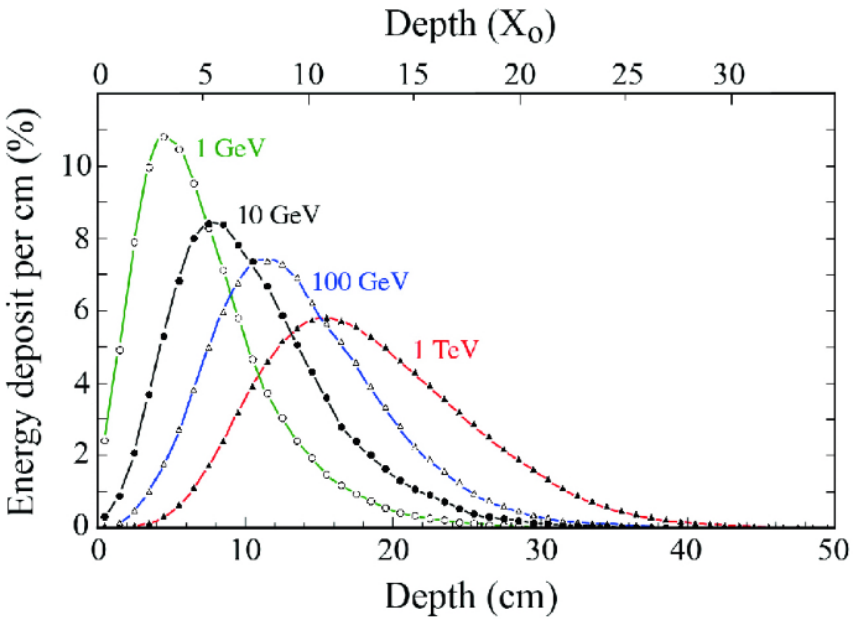


Figure 2.2.: Longitudinal energy distributions of electron showers in copper for four different initial energies [17]. The upper x-axis represents the depth given in radiation lengths, the lower one in centimetres. The y-axis shows the relative fraction of initial energy deposited within one centimetre of copper.

Unlike the longitudinal development of electromagnetic showers, their lateral development is neither dominated by emission of bremsstrahlung nor by electron-positron pair

production. The scattering angles of both processes are proportional to $\frac{1}{\gamma}$ [13], where γ is the Lorentz factor of the scattered particles. For highly boosted particles, the scattering angles are thus very small. It is multiple scattering between low-energy electrons and positrons, as well as Compton scattering of low-energy photons, that causes the shower to expand laterally. However, even though the lateral expansion is driven by different processes, the shower's width can still be very accurately quantified by a single parameter called the Molière radius, ρ_M. The reason why the Molière radius is so useful is because on average already 90 % of the initial energy are contained within a cylinder of radius ρ_M around the shower axis. It is defined as

$$\rho_M = \frac{E_s}{E_c} X_0 \,, \tag{2.6}$$

where X_0 and E_c are the already well-known radiation length and critical energy, respectively. E_s, on the other hand, is a new empirical parameter which is approximately equal to 21 MeV [15].

Figure 2.3 shows simulated radial energy distributions of electron-initiated electromagnetic showers in different detector materials (aluminium, copper, and lead), where the energy deposition within $0.1\rho_M$ is given as a function of the distance to the shower axis. The distributions fall off quickly and less than 0.1 % of the initial energy is deposited beyond approximately $3.5\rho_M$ [15].

The ratio of Molière radius to radiation length scales linearly with the atomic number: $\frac{\rho_M}{X_0} \sim Z$ [13]. Consequently, electromagnetic showers are slimmer, in relation to their length, in lighter detector materials than in denser ones. In terms of absolute values, though, a shower is wider in aluminium than in lead, as shown in Table 2.1. Small radiation lengths as well as Molière radii are the reason why modern calorimeters are usually made of heavy elements, such as tungsten or lead, because those materials keep calorimeter sizes small as well.

2.2. Hadronic Showers

2.2.1. Spallation and Evaporation

In contrast to electromagnetic showers, hadronic showers are not primarily created via electron-positron pair production or bremsstrahlung. Instead, inelastic scattering between highly energetic hadrons and the atomic nuclei of the detector material cause an exponential production of secondary particles, which in turn scatter with other atomic nuclei as well. Such a secondary particle can be another hadron or an electrically charged lepton.

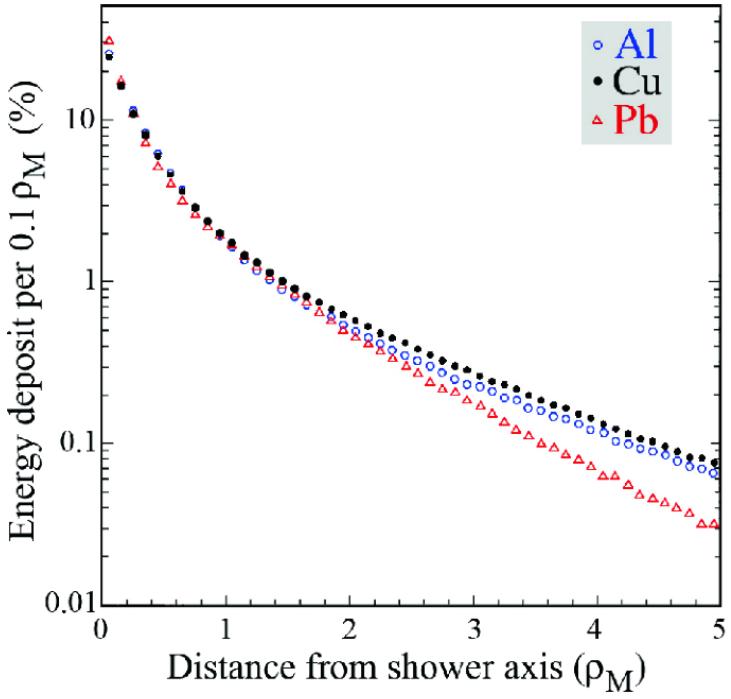


Figure 2.3.: Simulations of lateral energy distributions of electron-initiated electromagnetic showers within aluminium (blue), copper (black), and lead (red) [17]. The x-axis represents the distance from the shower axis in units of Molière radii. The y-axis shows the relative percentage of initial energy deposited within 0.1 Molière radii.

Table 2.1.: Values of different electromagnetic shower parameters for various detector materials [15]. The atomic number Z, the critical energy E_c, the radiation length X_0, the Molière radius ρ_M, and the ratio of Molière radius to radiation length $\frac{\rho_M}{X_0}$ are shown.

Material	Z	E_c for e^- [MeV]	E_c for e^+ [MeV]	X_0 [mm]	ρ_M [mm]	ρ_M/X_0
Al	13	42.7	41.5	89	44	0.49
Fe	26	21.7	21.0	18	17	0.94
W	74	8.0	7.7	3.5	9.3	2.66
Pb	82	7.4	7.1	5.6	16	2.86
U	92	6.7	6.4	3.2	10	3.13

The latter possibility is the reason why hadronic showers also comprise electromagnetic subshowers. Furthermore, neutrinos can be produced in hadronic interactions too. However, since they barely interact with matter, neutrinos are not detected, but only carry (significant) amounts of energy away, which makes the energy reconstruction of hadronic showers particularly difficult.

Hadronic shower development happens in two phases. The first one is called spallation

and takes place approximately 10^{-22} seconds [13] after the initial collision. The incoming hadron scatters inelastically with a nucleon (either proton or neutron), which produces large amounts of secondary particles. Very light hadrons, such as kaons or pions, are the most frequent produced in such collisions. These secondary particles scatter within the atomic nucleus as well, creating an intra-nuclear cascade. Eventually, the secondary particles escape the atom, leaving the nucleus highly excited. By emitting core fragments (protons or neutrons), α-particles, or strong X- or gamma radiation, all of them with energies between about 100 MeV and 1 GeV [13], the nucleus de-excites. Both core fragment emissions as well as secondary particles escaping the nucleus make up the first phase of hadronic shower development. Figure 2.4 shows a schematic representation of spallation.

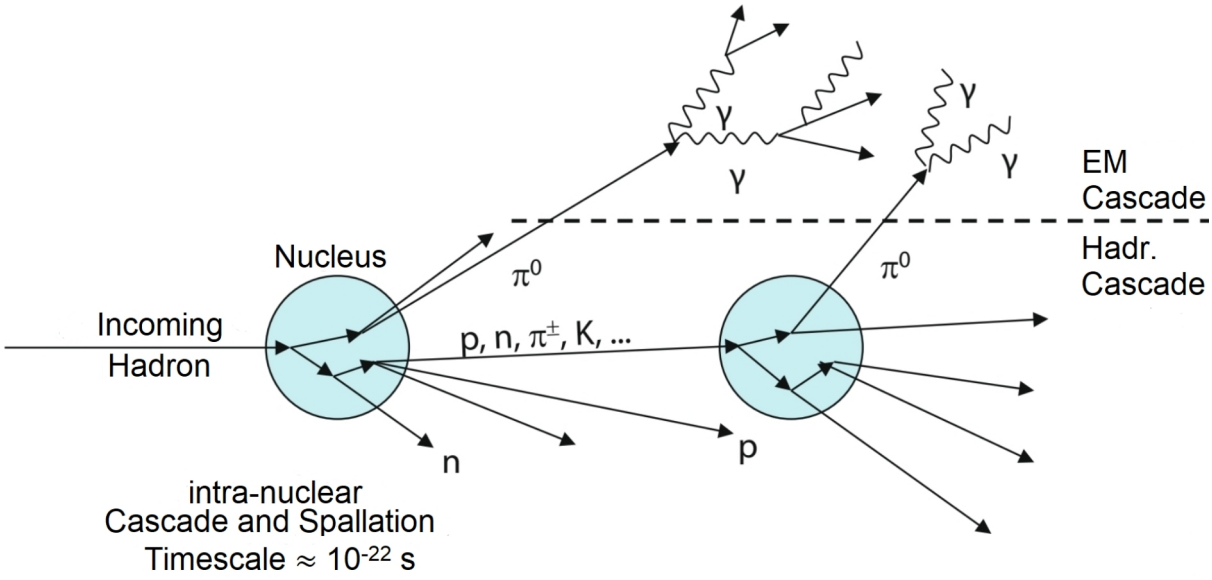


Figure 2.4.: Schematic depiction of the first phase of hadronic shower development (spallation) [13]. The hadronic shower is initiated by an incoming hadron that hits a nucleus and scatters inelastically. Secondary particles are produced and escape the nucleus, hitting other atoms of the detector material. Electromagnetic subshowers from neutral pion decays are also shown.

The second phase of hadronic shower development is called evaporation. Approximately 10^{-18} seconds [13] after the initial collision, atomic nuclei are still highly excited. In order to de-excite even further, a nucleus can undergo two processes: it either emits more core fragments, this time with energies of $\mathcal{O}(10\,\text{MeV})$ [13], or it splits into lighter atoms via nuclear fission. Both of these processes are shown in Figure 2.5. Of course, the daughter nuclei can evaporate or fissure into lighter elements as well if they are still energetic enough.

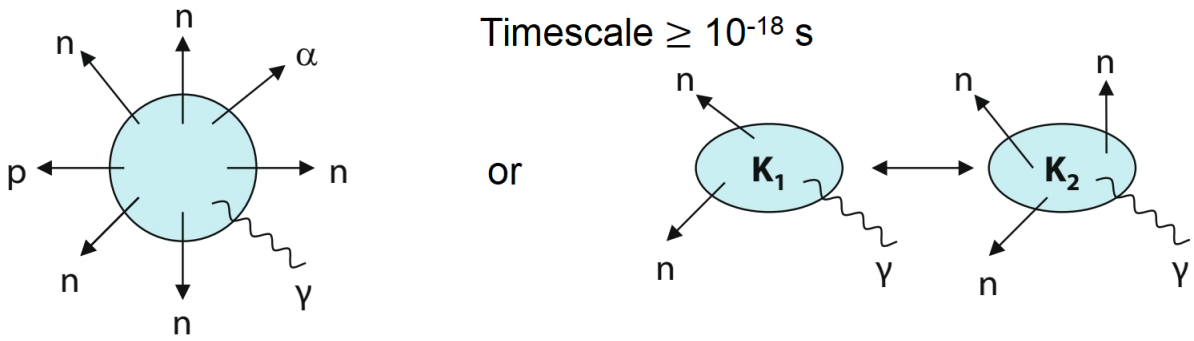


Figure 2.5.: Schematic depiction of low energy core fragment emission (left) and nuclear fission (right) [13]. Together, these processes form the second phase of hadronic shower development (evaporation).

Even though evaporation already starts 10^{-18} seconds after the initial hadron has struck the first nucleus, it can take up to microseconds before it ends. The reason for this are neutrons which barely interact with the electromagnetic field of a nucleus and therefore travel (mostly) unimpededly through the detector. Only when neutrons are caught by other nuclei, they emit photons and thereby deposit energy within the hadronic calorimeter. These photons are detected with a delay of multiple microseconds and can therefore, in addition to neutrinos, seriously deteriorate the energy reconstruction of the hadronic calorimeter. In the case of the LHC, for example, where proton bunches collide every 25 nanoseconds [18], 40 bunch crossings take place every single microsecond. It is thus not unlikely to detect photons in an event that belong to a previous one. One has to be aware of this issue when measuring energies of hadronic showers and react appropriately to ensure little to no data distortion.

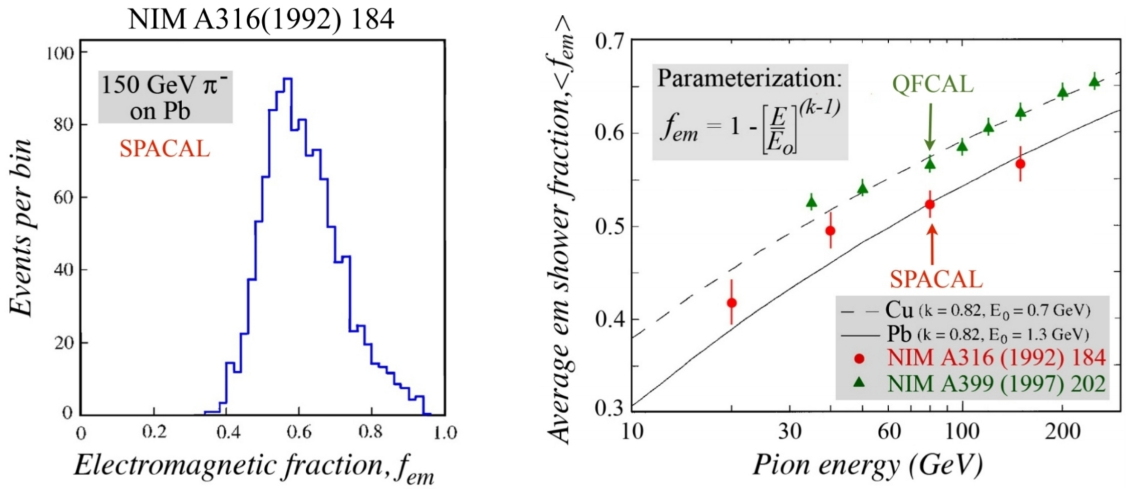
2.2.2. Electromagnetic Subshowers and Calorimeter Responses

As already mentioned and shown in Figure 2.4, hadronic showers are able to develop electromagnetic subshowers. These are caused by neutral pions which, in almost 99 % of all cases [15], decay into two photons. They are copiously produced in hadronic interactions, since they are one of the lightest hadrons, and therefore make up a large fraction of all secondary particles in a hadronic cascade. Hence, electromagnetic subshowers contribute a non-negligible fraction to the total deposited shower energy, E_{dep}, which can be mathematically described as

$$E_{\text{dep}} = (f_{\text{EM}} + f_{\text{had}})E_0\,. \tag{2.7}$$

Here, f_{EM} and f_{had} are the electromagnetic and hadronic fractions of the hadronic shower, respectively. The latter accounts for the energy that is deposited within the hadronic calorimeter by pure hadronic interactions, the former accounts for energy from electromagnetic subshowers and is therefore closely related to the total number of neutral pions produced in a hadronic shower.

Since the amount of neutral pions is in principle unbounded (as long as conservation laws are satisfied), the electromagnetic fraction can take on any value between 0 % and 100 %, which makes it prone to strong fluctuations. This is emphasised by Figure 2.6(a), where a distribution of the electromagnetic fraction of hadronic showers initiated by 150 GeV pions in a Pb-fiber calorimeter is shown. One can see that the distribution does not have a sharply defined peak, but is spread out from 0.4 to 1, even though neither the initial energy nor the initial particle were altered between events.



(a) (b)

Figure 2.6.: (a) Distribution of the electromagnetic fraction of hadronic showers initiated by 150 GeV pions, which were detected in a Pb-fiber calorimeter. The x-axis shows the relative electromagnetic fraction and the y-axis the number of events per bin. (b) The average electromagnetic fraction (y-axis) as a function of the initial pion energy (x-axis). The two black curves represent the theoretical prediction of Equation (2.8) for copper (dashed) and lead (continuous). Measurement results from calorimeters are also shown, one with copper absorber (triangles) and one with lead absorber (circles) [19].

Even though the electromagnetic fraction cannot be predicted very well for single events, its average value, $\langle f_{\mathrm{EM}} \rangle$, is well described by the following approximation [19]:

$$\langle f_{\mathrm{EM}} \rangle \approx 1 - \left(\frac{E}{E_0}\right)^{k-1} . \tag{2.8}$$

Here, E is the initial energy, E_0 is the average energy needed to produce a pion (material-dependent), and k is an empirical parameter approximately equal to 0.82 [13, 19]. Figure 2.6(b) shows the curves of Equation (2.8) for copper and lead as a function of the initial energy E. Furthermore, measurements of the average electromagnetic fraction are also shown for copper and lead.

Strong fluctuations of the electromagnetic fraction are the most important reason why one has to distinguish between electromagnetic and hadronic components, since hadronic calorimeter responses are very sensitive to changes in f_{EM}. Even tiny changes can have significant impacts on how well a hadronic calorimeter can detect electromagnetically and hadronically deposited energy. This becomes apparent when one describes calorimeter responses mathematically. The signal an electromagnetic shower induces within a hadronic calorimeter can be written as [13]

$$S(e) = f_{\mathrm{EM}} \epsilon_{\mathrm{EM}} E_e \,, \tag{2.9}$$

where E_e is the initial energy of an electromagnetically interacting particle, and ϵ_{EM} is the electromagnetic efficiency of the detector. A detector is never an ideal machine, and ϵ_{EM} therefore accounts for energy loss due to the imperfectness of the detector. Hence, $S(e)$ is just the total energy of an electromagnetic shower that a hadronic calorimeter is able to detect. Similarly, the hadronic signal can be written as [13]

$$S(\pi) = (f_{\mathrm{EM}} \epsilon_{\mathrm{EM}} + f_{\mathrm{had}} \epsilon_{\mathrm{had}}) E_\pi \,, \tag{2.10}$$

where ϵ_{had} plays the same role as ϵ_{EM}, but for hadronic interactions. E_π is the initial energy of a hadronically interacting particle. Combining Equations (2.9) and (2.10), and assuming equal initial energies ($E_e = E_\pi$), then yields:

$$\frac{S(e)}{S(\pi)} = \frac{\frac{\epsilon_{\mathrm{EM}}}{\epsilon_{\mathrm{had}}}}{1 - f_{\mathrm{EM}} \left(1 - \frac{\epsilon_{\mathrm{EM}}}{\epsilon_{\mathrm{had}}}\right)} \,. \tag{2.11}$$

Typically, one finds that $\frac{\epsilon_{\mathrm{EM}}}{\epsilon_{\mathrm{had}}} > 1$ [13] because hadronic showers have invisible components (neutrinos as well as binding energy needed to split up atomic nuclei), which implies that the right-hand side of Equation (2.11) is not equal to one but is a function of f_{EM}. Therefore, one can see that extreme fluctuations in the hadronic energy resolution are an unavoidable problem for hadronic calorimeters because the electromagnetic fraction has direct influence on the electromagnetic and hadronic signals and varies strongly between single events.

2.2.3. Longitudinal and Radial Shapes of Hadronic Showers

Hadronic showers distribute their energy similarly to electromagnetic showers, both in the longitudinal as well as the radial direction. Their lengths can be parameterised by a single quantity, just like for electromagnetic showers. This quantity is called the nuclear absorption length, λ. The nuclear absorption length can be approximated via [13]

$$\lambda \approx \lambda_0 \frac{\sqrt[3]{A}}{\rho}\,, \tag{2.12}$$

where A and ρ are the atomic weight and the density of the detector material, respectively. Moreover, λ_0 is a constant approximately equal to $35\,\mathrm{g\,cm^{-2}}$ [13]. The length of a hadronic shower grows logarithmically too, as it is the case for electromagnetic showers, and it scales with λ (or X_0) [13].

By comparing radiation lengths with nuclear absorption lengths, one can notice that the ratio of nuclear absorption length to radiation length scales approximately linearly with the atomic number Z as [13]

$$\frac{\lambda}{X_0} \approx aZ\,, \tag{2.13}$$

where $a \approx 0.37$ [13]. This relation implies that λ is much larger than X_0 in very dense detector materials, which also means that hadronic showers are on average much longer than electromagnetic showers. Hence, hadronic calorimeters have to be much larger than electromagnetic calorimeters. To emphasise this point, Table 2.2 shows values for the radiation length, the nuclear absorption length, their ratio, and the density of different detector materials. One can see very clearly that λ becomes much larger than X_0 in detectors made of heavy elements.

Table 2.2.: Values of shower parameters for different detector materials [15]. The atomic number Z, the density ρ, the radiation length X_0, the nuclear absorption length λ, and the ratio of nuclear absorption length to radiation length $\frac{\lambda}{X_0}$ are shown for different elements.

Material	Z	ρ [$\mathrm{g\,cm^{-3}}$]	X_0 [cm]	λ [cm]	λ/X_0
Al	13	2.7	8.9	39.7	4.5
Fe	26	7.9	1.8	16.8	9.3
W	74	19.3	0.35	9.9	28.3
Pb	82	11.4	0.56	17.6	31.4
U	92	19.0	0.32	11.0	34.4

To parameterise a hadronic shower longitudinally, Equation (2.3) can be generalised such that it now describes the energy deposition of a shower comprising a hadronic long component and an electromagnetically dominated short component. This parameterisation is of the form [20]

$$\frac{\mathrm{d}E}{\mathrm{d}z} = E_0 \cdot \left\{ \underbrace{\frac{f_{\mathrm{EM}}}{\Gamma(\alpha_s)} \cdot \left(\frac{z}{\beta_s}\right)^{\alpha_s - 1} \cdot \frac{e^{-\frac{z}{\beta_s}}}{\beta_s}}_{\text{short component}} + \underbrace{\frac{1 - f_{\mathrm{EM}}}{\Gamma(\alpha_l)} \cdot \left(\frac{z}{\beta_l}\right)^{\alpha_l - 1} \cdot \frac{e^{-\frac{z}{\beta_l}}}{\beta_l}}_{\text{long component}} \right\}, \tag{2.14}$$

where the left-hand side is the change in energy along the direction of propagation, z (parallel to the shower axis). Here, z is measured in units of the nuclear absorption length. Furthermore, f_{EM} is the electromagnetic fraction that now accounts for the fractional contribution of the short component to the whole energy deposition. The parameters α_i and β_i (both for the short and long component) are similar to a and b in Equation (2.3). In particular, α_i corresponds to a and β_i to $\frac{1}{b}$. They both determine the shape and the slope, respectively, of Equation (2.14).

Showers described by Equation (2.14) look like those depicted in Figure 2.7, which shows measurements of longitudinal energy distributions of 80 GeV pion- and proton-initiated hadronic showers. In addition to the data points, the short and long components of Equation (2.14) are shown too, each separately, as well as their sum. The x-axes represent the distance from the shower start in units of nuclear absorption lengths and the y-axes the energy deposited in one detector layer. Note that the energy is given in units of minimal ionising particles (MIPs) instead of GeV.

Similar to the Molière radius of electromagnetic showers, a radius, comprising 95 % of the shower's total energy, can be defined for hadronic showers too. It is approximately equal to the nuclear absorption length [13]:

$$R_{95\,\%} \approx \lambda\,. \tag{2.15}$$

Together with this radius, one can define, in a similar manner to Equation (2.14), a radial parameterisation of hadronic showers, encompassing an electromagnetically dominated core region surrounded by a hadronic halo [20]:

$$\frac{\Delta E}{\Delta S} = A_{\mathrm{core}} \exp\left(-\frac{r}{\beta_{\mathrm{core}}}\right) + A_{\mathrm{halo}} \exp\left(-\frac{r}{\beta_{\mathrm{halo}}}\right)\,. \tag{2.16}$$

Here, the left-hand side is the energy density as a function of the radius r measured from the shower axis. The energy density is defined as the energy, ΔE, divided by the area,

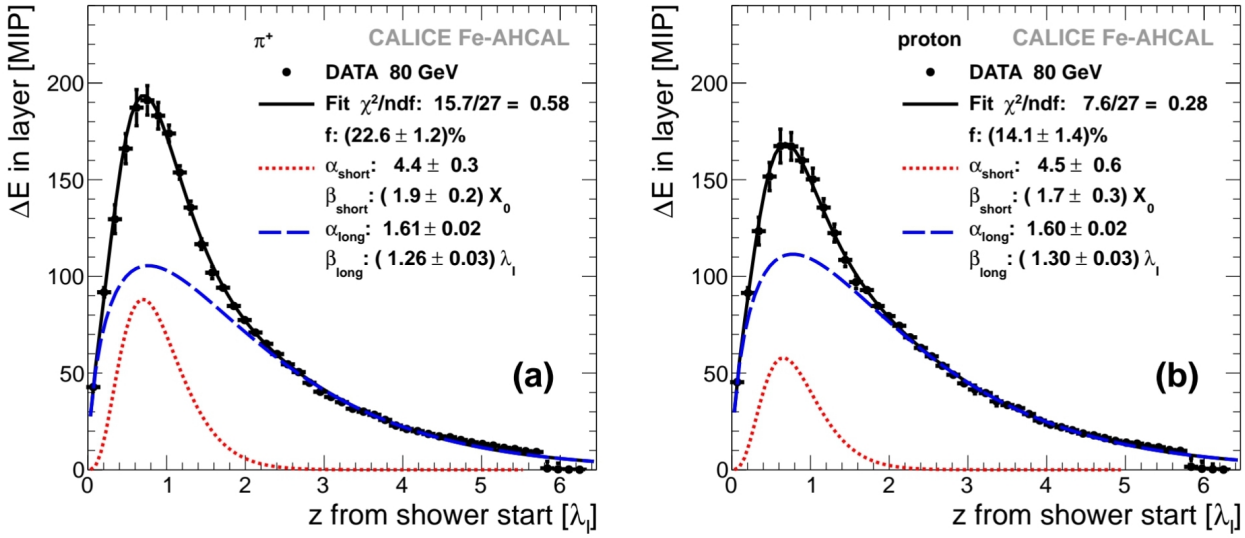


Figure 2.7.: Longitudinal energy distributions (in MIP) of 80 GeV pion- (left) and proton-initiated (right) hadronic showers as a function of the shower depth (in nuclear absorption lengths, here denoted as λ_I) [20]. Data points are shown as well as the fit of Equation (2.14) to the data (black curve). Furthermore, the short (red, dotted curve) and the long (blue, dashed curve) components of the fit are plotted separately too.

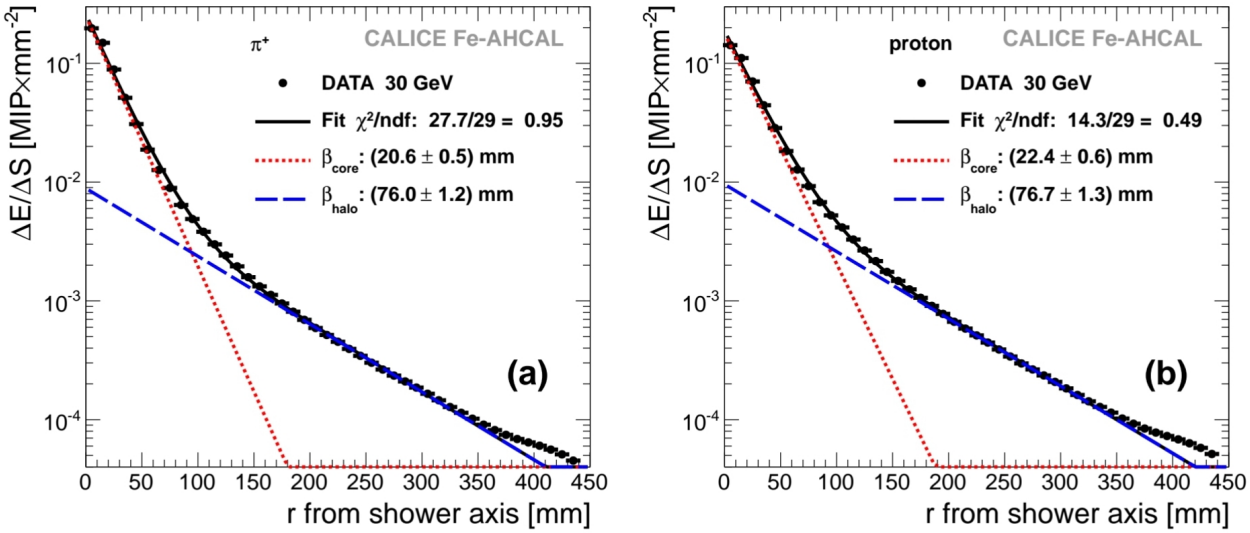


Figure 2.8.: Radial energy density distributions (in MIP mm^{-2}) of 30 GeV pion- (left) and proton-initiated (right) hadronic showers as a function of the distance to the shower axis (in millimetres) [20]. Data points are shown as well as the fit of Equation (2.16) to the data (black curve). Furthermore, the core (red, dotted curve) and the halo (blue, dashed curve) components of the fit are plotted separately too.

ΔS, of a ring with width Δr and radius r around the shower axis within which the energy is deposited. On the right-hand side, two scaling factors, A_{core} and A_{halo}, are included as well as two slope parameters, β_{core} and β_{halo}. Distributions of this shape look like those depicted in Figure 2.8, obtained from measurements of 30 GeV pion- and proton-initiated hadronic showers. The x-axes represent the radius from the shower axis in millimetres and the y-axes show $\frac{\Delta E}{\Delta S}$ in units of MIP mm^{-2}. The hadronic radial energy distributions fall off as quickly as those of electromagnetic showers.

3. The CALICE Collaboration and the AHCAL Prototype

Sophisticated particle detectors are an essential tool of modern high energy particle physics. In order to detect (elementary) particles and to study their properties, particle detectors comprise multiple components, each fulfilling a different purpose. Such detector components are carefully developed, built, and tested to ensure correct functionality, high performance, as well as fine resolution. A collaboration that dedicates itself to the research and development of highly granular calorimeters is the CALICE Collaboration, which is the subject of this chapter. In the following, a short introduction to CALICE and a motivation for fast hadron shower simulations, the topic of this thesis, are given, which are then followed by a description of the AHCAL detector prototype in Section 3.1 as well as two test beam campaigns in Sections 3.2 and 3.3.

CALICE stands for "**Ca**lorimeter for **Li**near **C**ollider **E**xperiment". It is an international collaboration of more than 300 physicists and engineers working on the research and development of high granularity and high performance detectors for a future International Linear e^+e^- Collider. CALICE is divided into different groups, with each group focusing on the construction and testing of one specific detector component (electromagnetic calorimeter, hadronic calorimeter, or tail catcher and muon tracker). For this thesis, work has been done in close cooperation with the AHCAL group ("**A**nalogue **H**adronic **Cal**orimeter") for which test beams were taken in 2018 and 2022.

Before test beam runs, such as those in 2018 and 2022, can take place, extensive simulations of test beam particles interacting with the detector material and producing showers have to be conducted. Based on simulated particle interactions with matter, one can enhance the detector design by evaluating which material is suited best for a certain part of a detector. However, simulations also allow for a sensible interpretation of data taken during a test beam run. Without simulations, one cannot tell whether a detector works properly and records data as expected or if there are too many disturbances caused by malfunctioning detector parts. Furthermore, testing a theoretical hypothesis does not make sense and is not possible if one does not know what to expect as an outcome of an

experiment. Particularly in high energy particle physics, detailed simulations are crucial in order to ensure correct functionality of detectors and to provide predictions which can be compared with data.

Currently, GEANT4 ("**Ge**ometry **an**d **t**racking") is used as a platform for the simulation of particle passages through matter in high energy physics experiments. Based on Monte-Carlo methods, GEANT4 is able to generate event kinematics, track single particles through simulated detectors, simulate interactions between particles and matter while taking into account all possible physics processes, and much more. While all these features yield highly accurate predictions about nature, they also require very large amounts of computational power and a lot of computing time. To counteract this issue, fast simulations can be used, for they are a useful tool to capture and provide the most important information about physical processes, for instance the energy distributions of a particle shower, without relying on large amounts of resources and computing time. Such simulations can be implemented as data-driven simulations, which, for instance, renders computations of highly complex equations of motion in full simulations unnecessary.

3.1. The AHCAL Prototype

As already mentioned previously, the CALICE Collaboration works on the development and construction of highly granular calorimeters. For this purpose, the AHCAL group has built its own prototype of a hadronic calorimeter called the AHCAL Technological Prototype (henceforth only referred to as "the AHCAL"). The AHCAL is a sampling calorimeter, using non-magnetic stainless steel as an absorber with a total of 38 active scintillator layers placed within the absorber structure. Figure 3.1 shows a picture of the fully assembled prototype. One absorber layer has a thickness of 17 mm, corresponding to about one radiation length or 0.1 nuclear absorption lengths, whereas a single active layer has a thickness of only 3 mm. In total, the thickness of the AHCAL amounts to 4.4 nuclear absorption lengths.

One active layer of the AHCAL is formed by four HCAL Base Units (HBUs), each with an area of 36×36 cm^2. Together, they are arranged quadratically, such that one active layer covers an area of 72×72 cm^2. Furthermore, one HBU is made of 144 (12×12) active scintillator tiles, each with a size of 3×3 cm^2, which means that one active layer encompasses a grid of 576 (24×24) scintillator tiles. Thus, the whole AHCAL comprises 21 888 channels that are read out individually via silicon photomultipliers (SiPMs). The SiPM model selected for the AHCAL is the Hamamatsu MPPC of type S13360-1325PE. In addition, every tile is individually wrapped in reflector foil in order to minimise optical

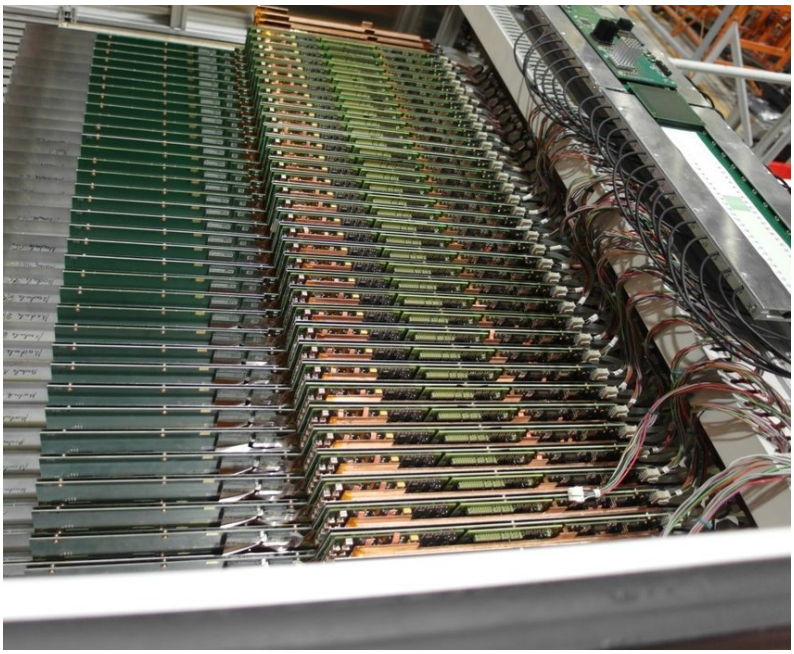
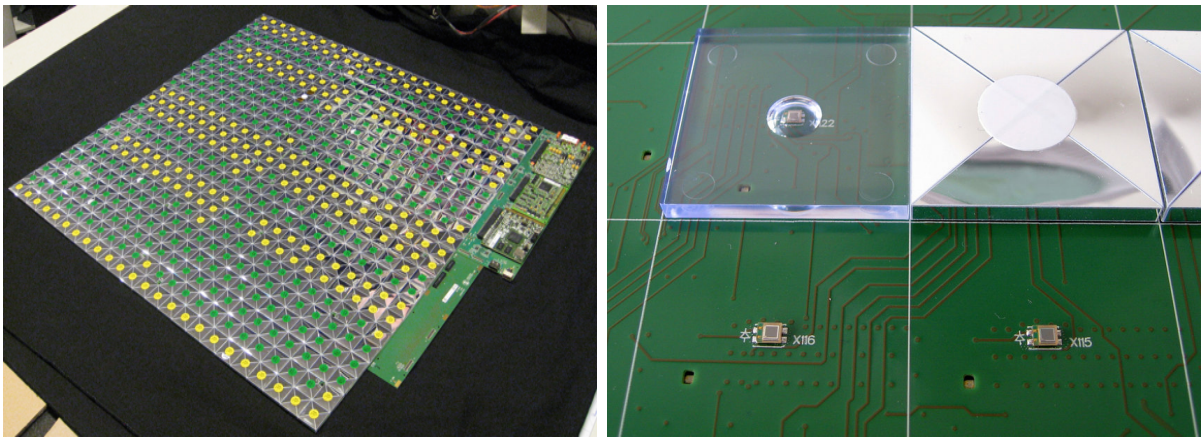


Figure 3.1.: Top view of the fully assembled 38-layer AHCAL stack [21].



(a) (b)

Figure 3.2.: (a) A picture of one active layer of the AHCAL with scintillator tiles and integrated read-out electronics [22]. All scintillator tiles are wrapped in reflector foil. (b) A picture of two naked SiPMs (bottom left and right) and two SiPMs with scintillator tiles on top (top left unwrapped, top right wrapped in foil) [23].

crosstalk between the tiles. Figure 3.2 shows a picture of an active layer on the left and one of single SiPMs on the right.

SiPMs are a relatively new technology compared to other photomultiplier models, and they are used for the sensing, timing, and quantifying of light signals (even down to single

photons) via the creation and amplification of electric signals. They are silicon-based arrays of self-quenching single-photon avalanche photodiodes (SAPDs) with a p^+-i-p-n^+ doping profile as shown in Figure 3.3, meaning that a single SAPD is designed such that a heavily p-doped layer (p^+) is followed by a slightly p-, almost undoped intrinsic layer (i), which itself in turn is followed by a normally p-doped (p) and a heavily n-doped (n^+) layer. The latter two layers make up the multiplication zone of the SAPD. Due to an external reverse bias that operates at a few volts above the breakdown voltage, an electric field is created, and the p^+-, i-, and p-layer become negatively charged, whereas the n^+-layer becomes positively charged. The strength of the electric field grows only slowly from the p^+- to the p-layer, but experiences a sudden increase within the multiplication zone with its maximum located at the p-n^+ junction.

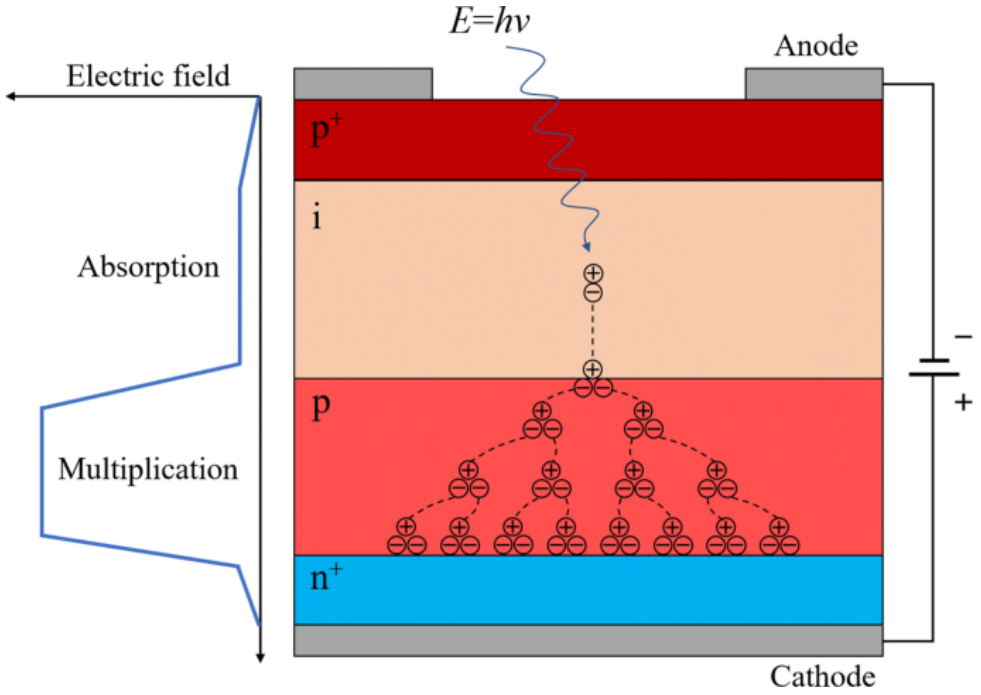


Figure 3.3.: Structure of an SAPD [24]. Its doping profile comprises a heavily p-doped, an intrinsic, a p-doped, and a heavily n-doped layer (from top to bottom). An external electric field charges the p^+-, i-, and p-doped layers negatively, whereas the n^+-layer is positively charged. The field strength increases slowly from the p^+- to the p-layer. From here, it rises quickly, reaches a maximum at the p-n^+ junction, and afterwards falls off quickly again.

A photon that enters an SAPD is completely absorbed by the intrinsic layer. During this process, an electron-hole pair is created. Due to the external electric field, the electron is drawn towards the multiplication zone, whereas the hole travels in the opposite direction. Within the multiplication zone, the electron is accelerated to high velocities by the strong

electric field, which enables it to create other electron-hole pairs via impact ionisation. These secondary charges are also accelerated and ionise the material even further, creating an "avalanche" of electrons, i.e. a measurable electric current. While the holes travel towards the negatively charged p^+-layer, the electrons are collected at the n^+-layer where they are read out as electric signals. Via this process, the initial light signal can be amplified by an amplification factor of the order of multiple millions [25], depending on the difference between breakdown voltage and reverse bias (called the overvoltage). Electric signals can, however, also be initiated by thermal electrons, the main source of noise within an SiPM [25]. The rate at which such signals are generated is called the dark count rate (DCR). Since the DCR is also overvoltage-dependent, a trade-off between increasing the amplification factor as well as the DCR is created. It is thus more advantageous to operate SiPMs at low temperatures in order to keep the DCR as small as possible [26].

3.2. Test Beam Run in 2018

The test beam run in 2018 was conducted in three single data-taking periods in the H2 test beam line at the CERN Super Proton Synchrotron beam test facility. The first one took place in May 2018, where the AHCAL was setup with all 38 active layer modules, integrated within the first 38 gaps of the absorber structure. Furthermore, the detector was placed on a movable platform that allowed to move the detector up and down or left and right in the x-y plane perpendicular to the beam axis. An event was then recorded by the detector if two external trigger scintillators, placed in the beam line, coincided with each other. The AHCAL was exposed to three different particle types during the run: electrons, muons, and negatively charged pions. For electrons, data was recorded within the energy range from 10 GeV and 100 GeV, whereas for pions, energies between 10 GeV and 160 GeV were measured. For muons, data with beam energies between 40 GeV and 120 GeV was acquired.

The second data-taking period took place in June 2018 for which the 38th active detector layer was exchanged by a module with scintillator tiles of size 6×6 cm^2 instead of 3×3 cm^2. The layer that had been previously integrated into the 38th gap was then installed within the 41st. Moreover, a single HBU was installed in front of the detector, acting as a pre-shower layer, and a tail catcher was setup in the rear as well. The tail catcher was made of twelve single-HBU active layers in total, alternated with 7.4 mm thick steel absorber layers. In this period, data was again recorded for muons and negative pions. For muons, only 40 GeV events were recorded, and for pions, energies between

10 GeV and 200 GeV were measured. In addition to that, positron test beams with energies ranging from 10 GeV and 100 GeV were also recorded.

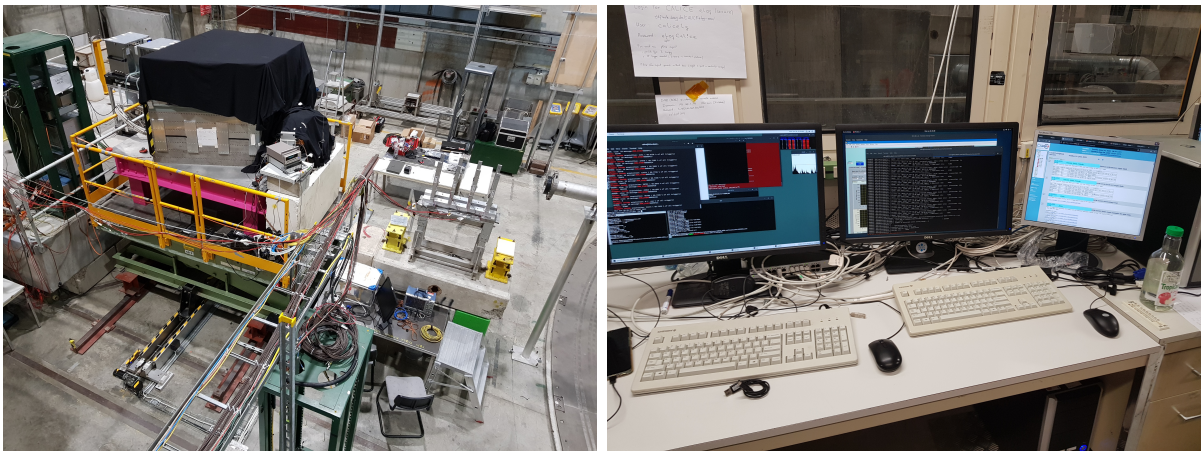
For the data-taking period that took place in October 2018, the AHCAL with 39 layers was tested together with the prototype of the silicon part of the CMS HGCAL ("**H**igh-**G**ranularity **Cal**orimeter Upgrade") [27, 28], which were now both mounted on a fixed platform. Since the AHCAL was placed behind the HGCAL prototype, which has a thickness of approximately five nuclear absorption lengths, only tails of hadronic showers and muons reached the AHCAL, which limited the data taking rate to approximately 50 events per second.

In total, about 93 million events were recorded during the test beam run in 2018 at CERN. For this thesis, only the data acquired for negatively charged pions has been used, limited to the following nine initial energies: $E_{\text{initial}} = \{10, 20, 30, 40, 60, 80, 120, 160, 200\}$ GeV.

3.3. Test Beam Run in 2022

The test beam run at CERN in June 2022 was similar to that four years prior. The only difference, though, was that the AHCAL was not tested solely this time. Instead, the 2022 test beam run was a combined run of the AHCAL and the SiW ECAL group ("**Si**licon-tungsten (**W**) **E**lectromagnetic **Cal**orimeter") [29] of CALICE. The SiW ECAL was mounted in front of the AHCAL, and both of them were again positioned on a movable platform, adjustable within the x-y plane perpendicular to the particle test beam. Figure 3.4 shows pictures of the detector setup, the beam pipe, and of the workplace from where the data recording was monitored.

During the data-taking period, electron beams with energies between 10 GeV and 150 GeV as well as positively charged pion beams with energies in the range of 20 GeV to 200 GeV were measured. Furthermore, 150 GeV muon data was acquired. In total, approximately ten million events were recorded in 2022. Significant contributions in terms of data-taking shifts were provided to the test beam campaign in the context of this thesis.



(a) (b)

(c)

Figure 3.4.: (a) Diagonal aerial perspective onto the whole detector setup. The detector is mounted on a movable platform (with orange-yellow railing) and covered with a black blanket. The beam pipe is visible, too, at the right edge of the picture. (b) A picture of a workplace from where the data taking was controlled. (c) View onto the front of the detector (black cube in the background) along the beam pipe (centre of the picture).

4. Longitudinal Simulation of Pion Showers using a Principal Component Analysis

The aim of this thesis is to develop a data-based fast simulation of pion showers. For this purpose, two simulation techniques have been investigated. One of these two methods is a principal component analysis (PCA) whose application to differences between longitudinal energy distributions of single pion showers and an average pion shower parameterisation is presented in this Chapter.

To begin with, Section 4.1 introduces the average pion shower parameterisation. Moreover, an explanation of how longitudinal energy differences per layer were calculated is given, and probability density functions (PDFs) of energy differences[1] are shown. Following this, Section 4.2 presents the results of the conducted PCA. First, the theoretical background behind principal component transformations is introduced. Then, the results of the PCA are presented, including variance plots, correlation factors, as well as PDFs and simulations of the principal components. In the end, results of converting simulated principal components back into simulated energy differences are shown in Section 4.3.

4.1. Average Longitudinal Pion Showers and Distributions of Individual Shower Energies

The following Section is based upon average shower shape studies conducted by Olin Pinto at DESY [30]. Average shower shapes are necessary for this research because they allow to transform absolute energies into energy differences by subtracting the energy of single pion showers from the average shower energy per layer. This simplifies the analysis considerably, since the mean of all energy differences is centred around zero, independent of the initial pion energy, which also makes it easier to compare the behaviour of the

[1]The terms "energy difference per layer" and "energy difference" are used equivalently in this thesis.

simulation between different initial energies.

The idea behind the application of a PCA to energy differences is schematically shown in Figure 4.1. After energy differences were computed, they are centred around zero and normalised to their standard deviations. After that, all energy differences are transformed into uncorrelated principal components. Vanishing correlation factors are important for the simulation because a single principal component can then be simulated without having to consider influence of other principal components.

The simulation occurs, first by simulating principal components according to their PDFs. The inverse transformation is then applied to them, resulting in centred and normalised simulated energy differences. Lastly, the respective standard deviations are multiplied and the respective means added to all centred and normalised simulated energy differences, which will yield the desired simulated energy differences. The energy in each calorimeter layer is thus simulated, in principle with correlations preserved.

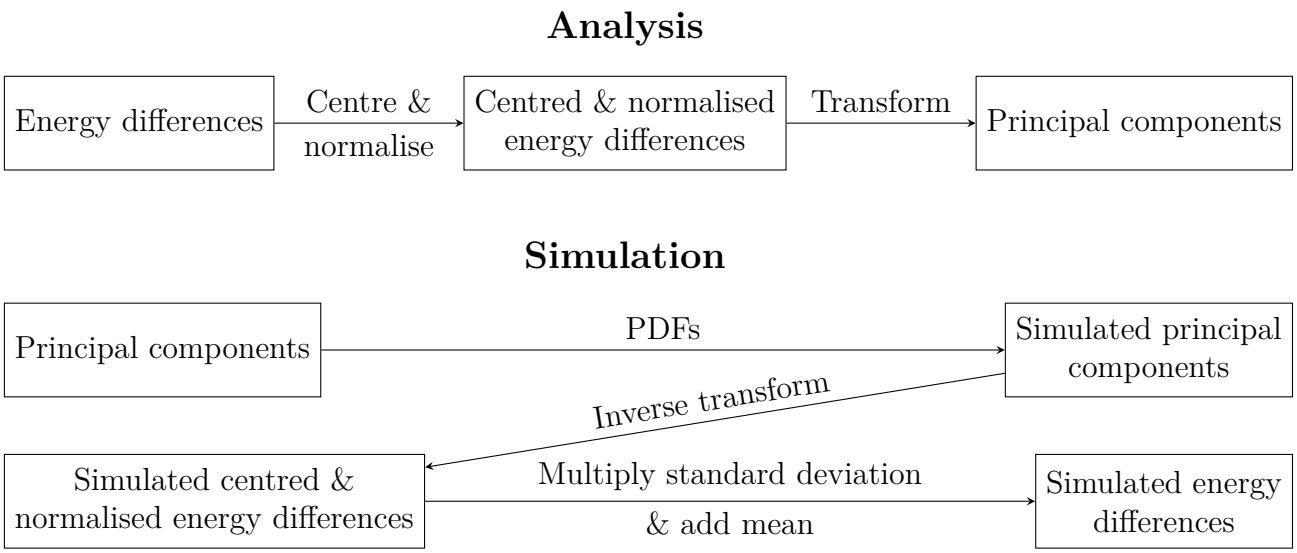


Figure 4.1.: Diagrams that schematically depict the principal component analysis and the simulation of longitudinal energy differences.

The dataset used for this thesis was recorded in 2018 at CERN by the AHCAL group of the CALICE Collaboration. It contains pion shower data of various initial energies, namely 10 GeV, 20 GeV, 30 GeV, 40 GeV, 60 GeV, 80 GeV, 120 GeV, 160 GeV, and 200 GeV, and it was also refined before being used during this and the following analyses. The criteria by which showers either passed or failed the selection are the following: First, a particle identification [31], based on boosted decision trees, was applied to the whole dataset in order to remove beam contamination. Furthermore, the first physical layer of the AHCAL was excluded due to uncertainties in the shower start finding algorithm [31]. Then, showers

with a shower start beyond the sixth physical detector layer were excluded to minimise leakage loss, and those that fulfilled the requirement had to have exactly one track and a matching hit within the first three layers of the detector. It is important to mention here that due to the shower start finding algorithm, pre-shower energies, i.e. energies deposited before the calculated shower start layer, are not considered in this thesis. Lastly, a gap rejection of two millimetres was applied to prevent that the impact point of a shower lies between two tiles within a single layer. This event selection is necessary for this thesis in order to be able to correctly use the average pion shower parameterisation that is presented next.

To obtain an average longitudinal energy distribution of pion showers, one has to average the shower energy layerwise for all events in a dataset. The resulting distribution shows how a pion shower on average distributes its energy longitudinally through all detector layers. The equation

$$E_A(z) = E_0 \cdot \left\{ \underbrace{\frac{f_{\mathrm{EM}}}{\Gamma(\alpha_s)} \cdot \left(\frac{z}{\beta_s}\right)^{\alpha_s - 1} \cdot \frac{e^{-\frac{z}{\beta_s}}}{\beta_s}}_{\text{short component}} + \underbrace{\frac{1 - f_{\mathrm{EM}}}{\Gamma(\alpha_l)} \cdot \left(\frac{z}{\beta_l}\right)^{\alpha_l - 1} \cdot \frac{e^{-\frac{z}{\beta_l}}}{\beta_l}}_{\text{long component}} \right\} \tag{4.1}$$

can then be fitted to all data points in order to describe the shower's behaviour quantitatively. Figure 4.2 shows an example of how this parameterisation appears for 60 GeV pions. It is important to note here that, from this point on, z is no longer expressed in units of nuclear absorption lengths (as in Figure 2.7) but in layers (of the AHCAL) measured from the shower start (thus, $z \in \mathbb{N}_0$). The conversion factor from GeV to MIPs used for this thesis was determined to be $f_{\mathrm{GeV}\to\mathrm{MIP}} = 37.3\,\mathrm{MIP\,GeV^{-1}}$ [32].

The parameterisation represented by Equation (4.1) has two components: a short and a long component, each with its own shape and slope parameters. The short component accounts for energy deposited by electromagnetic subshowers close to the beam axis, whereas the long component represents energy deposited by hadronic interactions which usually have a wider range than electromagnetic interactions.

For each initial energy, energy differences, ΔE, with the average shower deposition were computed via

$$\Delta E = E_A - E_S\,, \tag{4.2}$$

where E_A is the average energy in layer z from Equation (4.1). E_S, on the other hand, is the energy of a single event in the same layer[2]. The above computation was performed layerwise for the first 32 layers of the AHCAL (layers 0 to 31), where layer 0 represents the

[2]The term "event" henceforth means "shower".

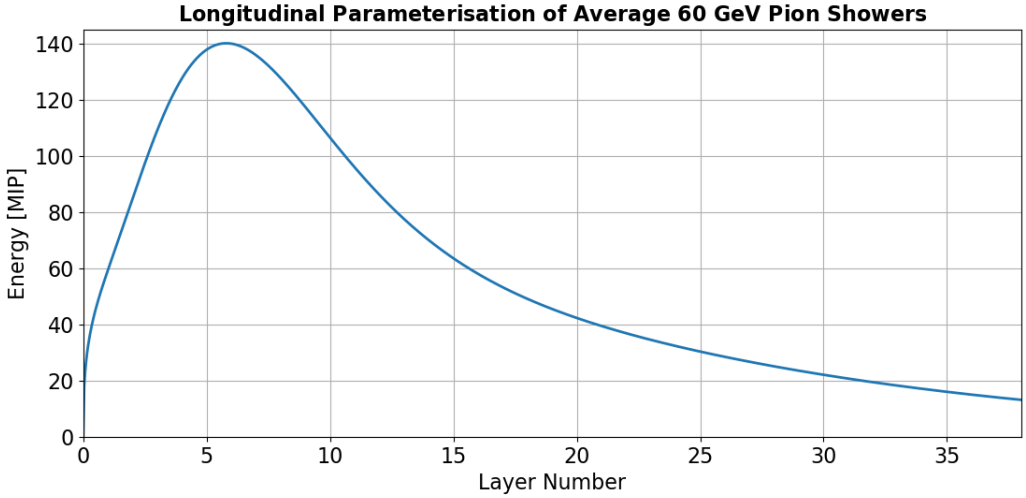


Figure 4.2.: A depiction of the parameterisation Equation (4.1) for 60 GeV pion showers. The energy deposited by the hadronic shower within each detector layer is shown in units of MIPs as a function of the calorimeter layer.

shower start layer. For the remaining seven layers, all energy differences were summed up and combined into a single variable, representing the energy differences in layers 32 to 38, since only a small fraction of the initial energy is deposited at the very end of the detector.

Equation (4.2) implies two things: First of all, energy differences are either positive or negative (in the former case, they are bounded from above by E_A because $E_S \geq 0$), depending on whether E_S is greater than average or less. Secondly, when averaging over many events, the mean of all energy differences is close to zero, independent of the initial energy (the reason why energy differences were chosen over absolute energies). This becomes clear when calculating and plotting average longitudinal energy distributions of pion showers, obtained from data, together with Equation (4.1) in one coordinate system, as shown in Figure 4.3. This Figure shows Equation (4.1) (orange curve) and average longitudinal energy distributions obtained from pion shower data (blue curve). For each initial energy, differences between both curves are small, which emphasises the validity of Equation (4.1).

The PDFs of energy difference distributions are well described by the product of a Gaussian and a Landau distribution:

$$f(x) = A \cdot \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)\right\}}_{\text{Gaussian}} \cdot \underbrace{\frac{1}{\pi c}\int_0^\infty e^{-t}\cos\left\{t\left(\frac{x-\mu}{c}\right)+\frac{2t}{\pi}\log\left(\frac{t}{c}\right)\right\}\mathrm{d}t}_{\text{Landau}} \,. \quad (4.3)$$

The parameters of this function include the mean, $\bar{x}$, and standard deviation, σ, of the

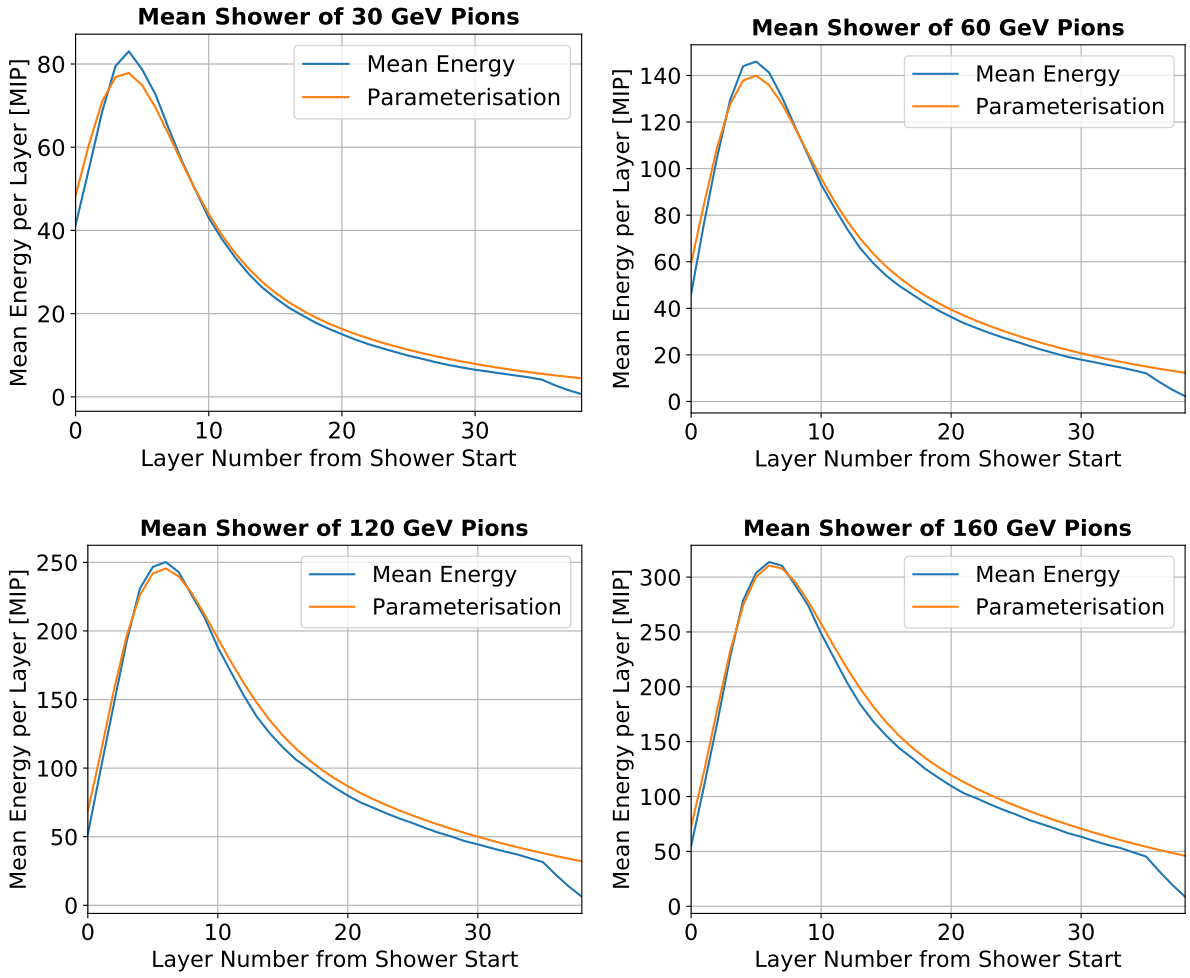


Figure 4.3.: Distributions of average longitudinal energy depositions of 30 GeV, 60 GeV, 120 GeV, and 160 GeV pion showers, calculated from data (blue), as a function of the detector layer. In addition, Equation (4.1) (orange) is also plotted in order to allow for a direct comparison between data and theory.

Gaussian distribution, a location parameter, μ, and a scale parameter, c, for the Landau distribution, and a multiplicative factor, A. Figures 4.4 and 4.5 show examples of the PDFs of energy differences and their corresponding fit functions for different detector layers for 60 GeV pions. All Figures are given on linear as well as logarithmic scale and exhibit very good agreement between the histograms and their corresponding fit curves. Deviations are visible at the tails of the distributions, but these do not exceed a relative fraction of 0.1 % or more. Furthermore, Figure 4.6 shows distributions of energy differences within the same detector layer, but for different initial pion energies. Here, the PDFs exhibit similar shapes and behaviour and clearly show that Equation (4.3) describes energy differences of various initial pion energies.

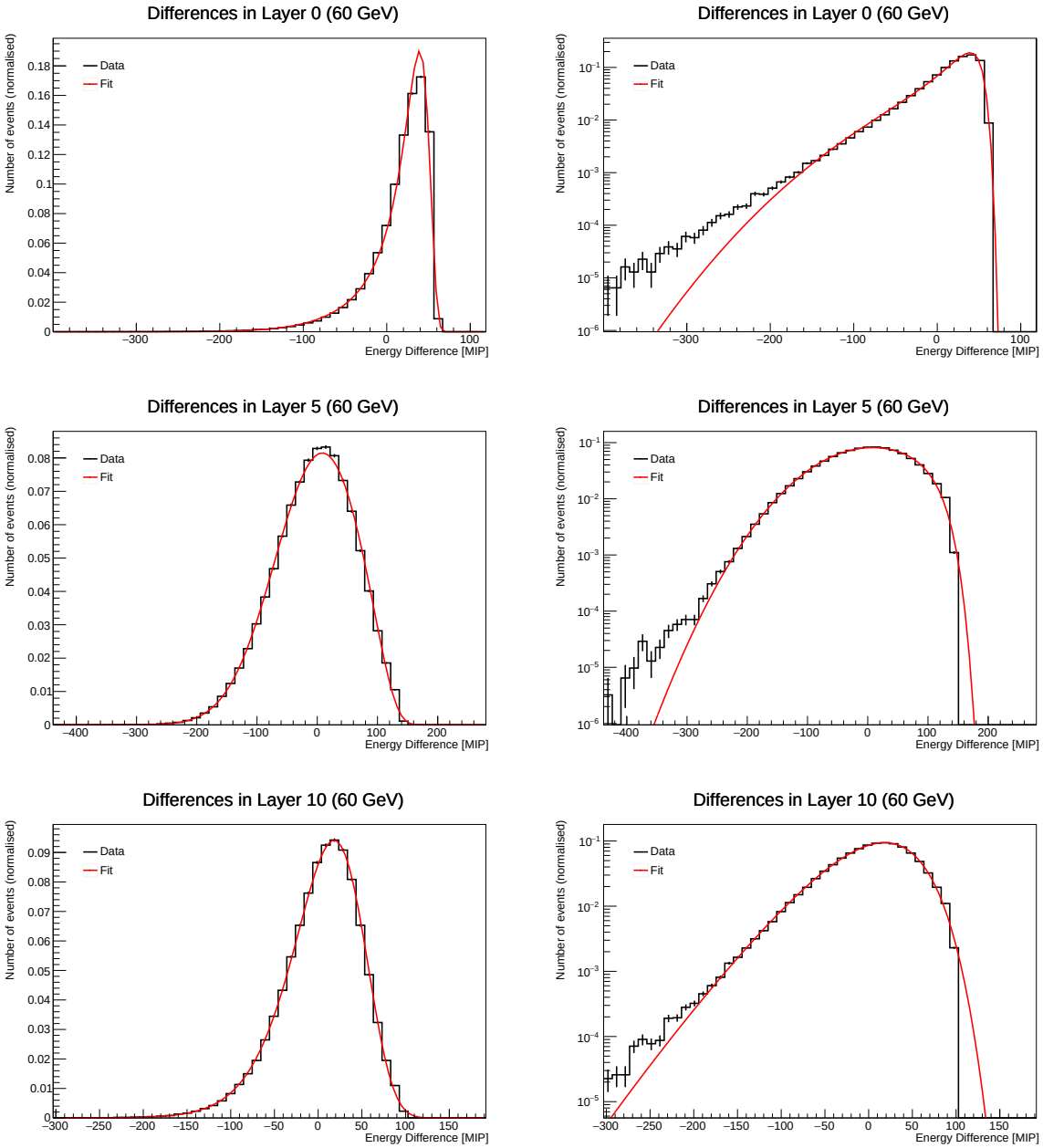


Figure 4.4.: PDFs of energy differences for 60 GeV pions. The left column shows distributions and their corresponding fit functions for layers 0, 5, and 10 (measured from the shower start) on linear scale. The right column shows the same distributions but on logarithmic scale.

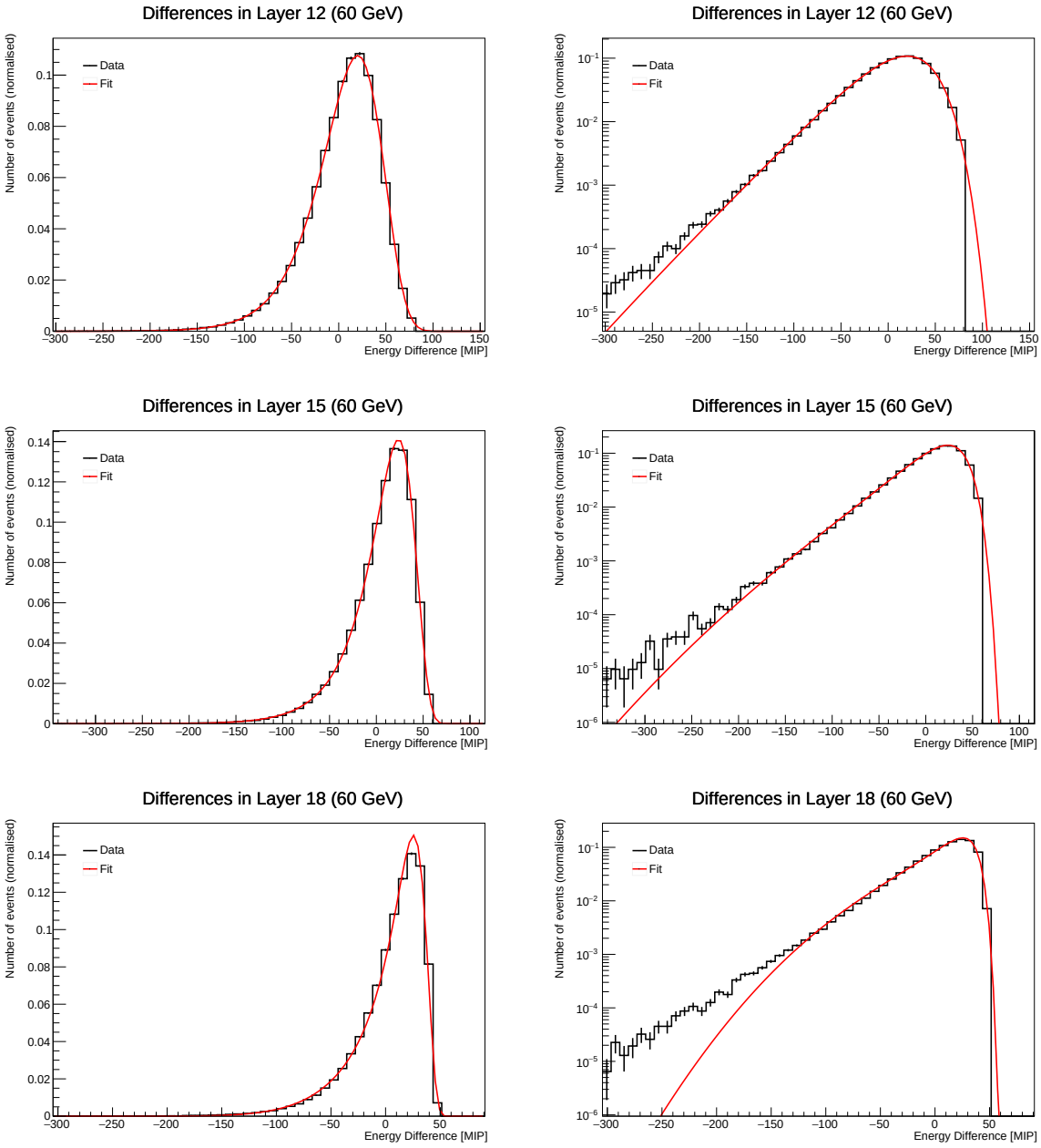


Figure 4.5.: PDFs of energy differences for 60 GeV pions. The left column shows distributions and their corresponding fit functions for layers 12, 15, and 18 (measured from the shower start) on linear scale. The right column shows the same distributions but on logarithmic scale.

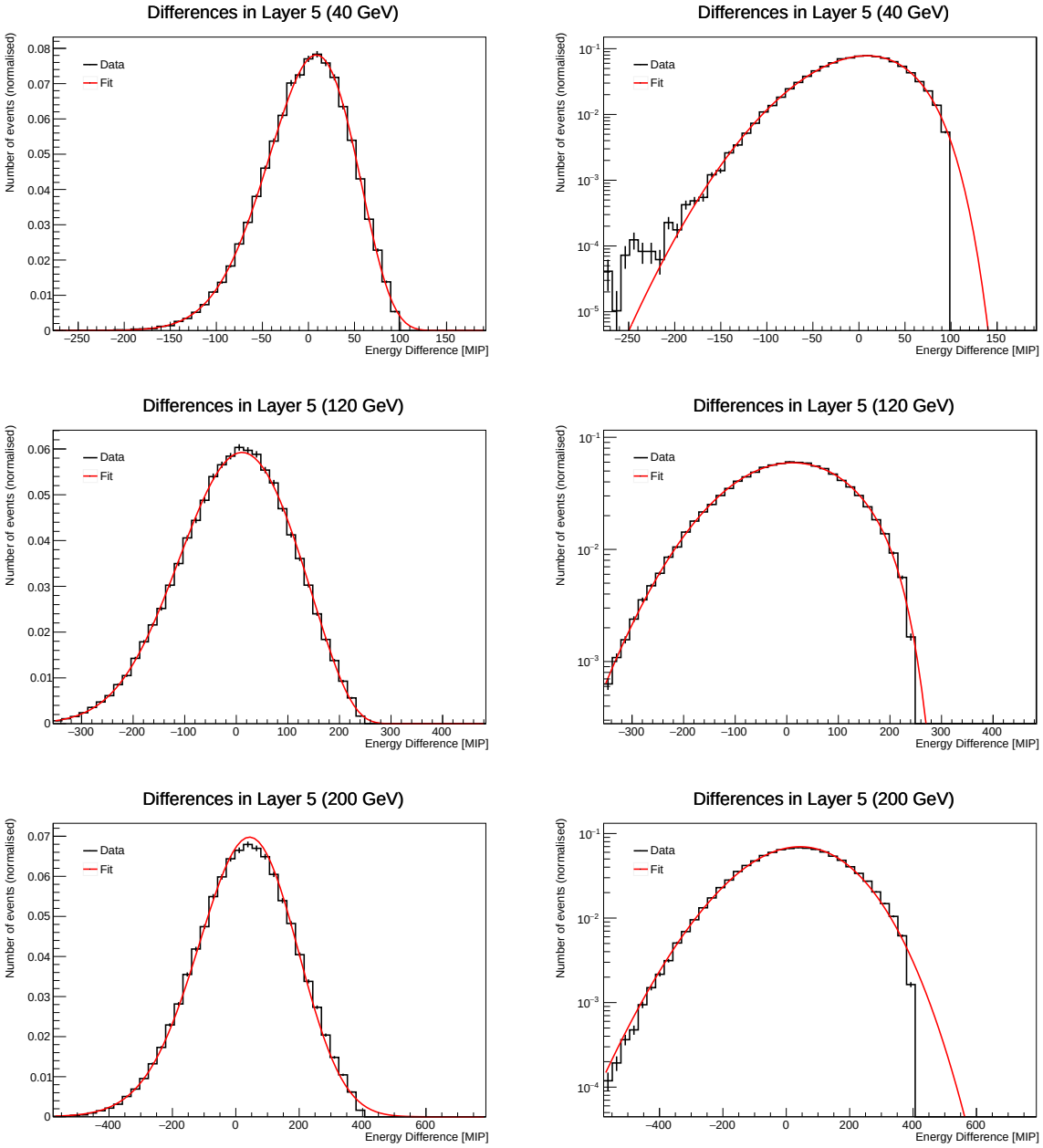


Figure 4.6.: PDFs of energy differences for 40 GeV (upper row), 120 GeV (middle row), and 200 GeV (bottom row) pions. The left column shows distributions and their corresponding fit functions for layer 5 (measured from the shower start) on linear scale. The right column shows the same distributions but on logarithmic scale.

4.2. Principal Component Analysis

4.2.1. Principal Component Transformation

A principal component transformation (PCT) is a linear transformation that transforms a set of correlated variables into a set of uncorrelated principal components. This transformation is done in the following way. Consider an $n \times p$ data matrix, $\mathbf{X}$, whose rows represent n events (or repetitions of an experiment) and its columns p variables, sometimes also called "features". For each column, one can find a mean, $\bar{x}_j$, and a standard deviation, σ_j ($j = 1, ..., p$), which are used to centre and normalise all matrix elements via

$$x_{\mathrm{norm},\,ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}\,. \tag{4.4}$$

The resulting matrix, $\mathbf{X}_{\mathrm{norm}}$, contains data of p variables, f_i, with entries that have a mean of zero column-wise. From this, a covariance matrix, $\mathbf{C}$, is then constructed. This $p \times p$ matrix is of the form

$$\mathbf{C} = \begin{pmatrix} \mathrm{Cov}(f_1, f_1) & \mathrm{Cov}(f_1, f_2) & \dots & \mathrm{Cov}(f_1, f_p) \\ \mathrm{Cov}(f_2, f_1) & \mathrm{Cov}(f_2, f_2) & \dots & \mathrm{Cov}(f_2, f_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(f_p, f_1) & \mathrm{Cov}(f_p, f_2) & \dots & \mathrm{Cov}(f_p, f_p) \end{pmatrix}. \tag{4.5}$$

Since $\mathrm{Cov}(x, x) = 1$ and $\mathrm{Cov}(x, y) = \mathrm{Cov}(y, x)$, $\mathbf{C}$ is symmetric.

With the covariance matrix $\mathbf{C}$, the transformation matrix, $\mathbf{V}$, can now be determined. For this, all eigenvalues and normalised eigenvectors of $\mathbf{C}$ have to be found. There are p pairs of eigenvalues, λ_i, and eigenvectors, $\mathbf{v}_i$, in total because $\mathbf{C}$ has dimensions $p \times p$. Each eigenvalue corresponds to a principal component, and its value is proportional to the amount of information about the initial variables that is carried by the respective principal component. Hence, by sorting all eigenvalues in descending order, one can easily see which principal components are the most and the least significant. Depending on this order, all eigenvectors are arranged similarly in a matrix whose jth column is the eigenvector $\mathbf{v}_j$, which forms the transformation matrix $\mathbf{V}$ that is required for the PCT.

The matrix of principal components, $\mathbf{Y}$, is now computed via

$$\mathbf{Y} = \mathbf{X}_{\mathrm{norm}}\mathbf{V}\,. \tag{4.6}$$

Due to the laws of matrix multiplication, one can notice two things: firstly, $\mathbf{Y}$ is an $n \times p$ matrix; secondly, all principal components are superpositions of the initial variables. Since

the initial variables might not all have the same physical dimensions or meanings, it is difficult to assign sensible physical interpretations to the principal components.

In the case discussed above, the dimensionality of the analysis remains unchanged. If, however, one decides to reduce the dimensionality and simplify the problem, only a subset of size k of eigenvectors may be used to construct $\mathbf{V}$. Usually, eigenvectors with small corresponding eigenvalues are discarded in this case because this minimises the inevitable information loss. The shape of $\mathbf{V}$ is then reduced to a $p \times k$ matrix, and $\mathbf{Y}$ will therefore only be an $n \times k$ matrix.

The inverse PCT is done in a similar manner to Equation (4.6):

$$\tilde{\mathbf{X}}_{\text{norm}} = \mathbf{Y}\mathbf{V}^{\text{T}} = \mathbf{X}_{\text{norm}}\mathbf{V}\mathbf{V}^{\text{T}} \,. \tag{4.7}$$

Here, $\mathbf{V}^{\text{T}}$ is the transposed of the matrix $\mathbf{V}$. Note that $\tilde{\mathbf{X}}_{\text{norm}} = \mathbf{X}_{\text{norm}}$ is only true if the dimensionality is not decreased. Otherwise, $\tilde{\mathbf{X}}_{\text{norm}}$ differs from $\mathbf{X}_{\text{norm}}$ due to the aforementioned information loss. Lastly, in order to obtain $\tilde{\mathbf{X}}$, one can just invert Equation (4.4) and exchange all x with $\tilde{x}$:

$$\tilde{x}_{ij} = \sigma_j \tilde{x}_{\text{norm},\,ij} + \bar{x}_j \,. \tag{4.8}$$

4.2.2. Determination and Simulation of Principal Components

For each initial energy, all energy differences, introduced in Section 4.1, were transformed into uncorrelated principal components. Since 33 initial variables were transformed, 33 principal components were obtained from the PCT. The variances of the principal components are shown in descending order in Figure 4.7.

Figure 4.7 shows that the first eight principal components already possess a significant amount of information about the initial variables, as their corresponding bars are much larger compared to those of higher principal components. Therefore, only the first eight principal components were considered during the following analysis. Those that remained were neglected, and as a consequence thereof, information loss was accepted. The relevant principal components were then displayed as PDFs and a double crystal ball function [33, 34] was fitted to the distributions in order to describe them analytically. In terms of four parameters, the crystal ball function is defined as

$$f(\mu, \sigma, \alpha, n; x) = N \cdot \begin{cases} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} & \text{if } \frac{x-\mu}{\sigma} > -\alpha \\ A \cdot \left(B - \frac{x-\mu}{\sigma}\right)^{-n} & \text{if } \frac{x-\mu}{\sigma} \leq -\alpha \end{cases} \,, \tag{4.9}$$

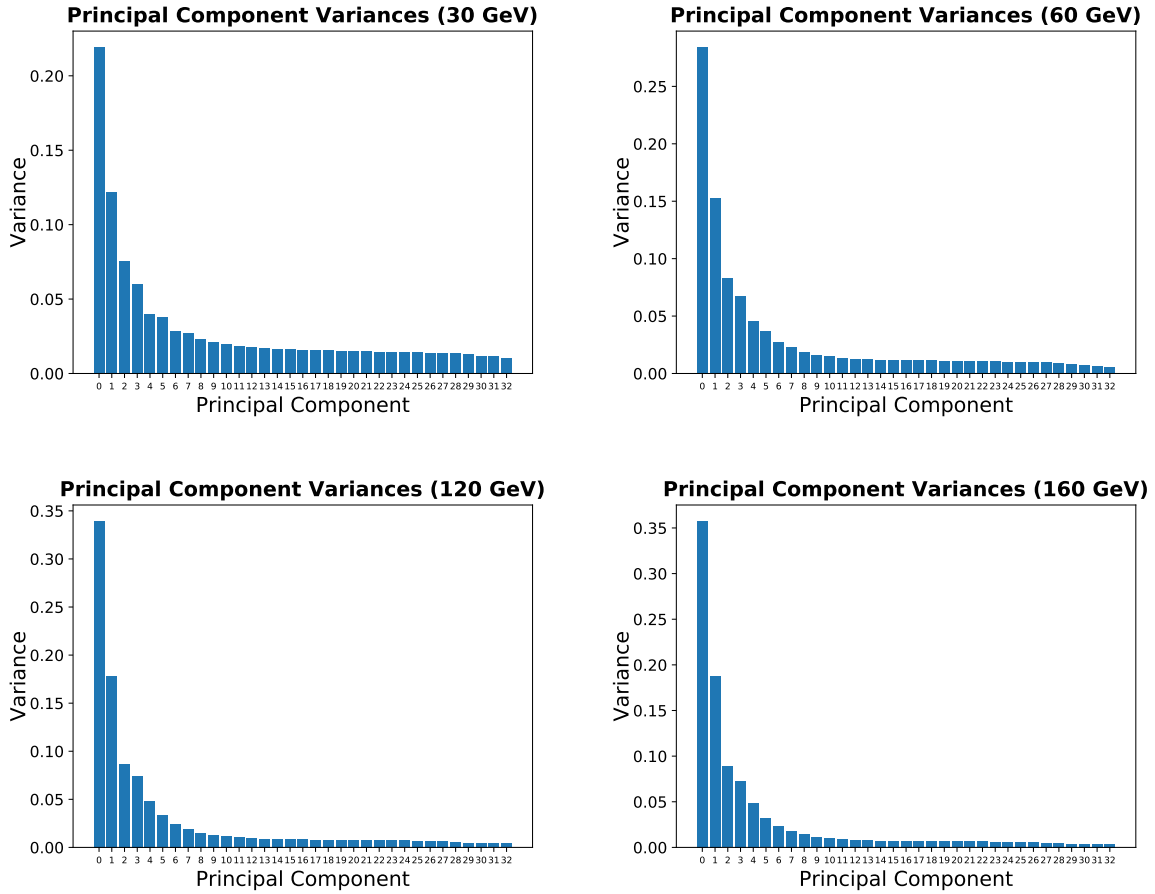


Figure 4.7.: Variances of principal components for 30 GeV, 60 GeV, 120 GeV, and 160 GeV shown in descending order. The variances fall quickly off for the first eight principal components and remain almost constant afterwards. Moreover, the higher the initial energy is, the larger is the contribution of the first eight principal components to the total sum of variances.

with A being defined as

$$A = \left(\frac{n}{|\alpha|}\right)^n \cdot \exp\left(-\frac{|\alpha|^2}{2}\right) \tag{4.10}$$

and B as

$$B = \frac{n}{|\alpha|} - |\alpha| \,. \tag{4.11}$$

N is a function of σ and other constants C and D:

$$N = \frac{1}{\sigma(C+D)} \,. \tag{4.12}$$

The constants C and D are of the form

$$C = \frac{n}{|\alpha|} \cdot \frac{1}{n-1} \cdot \exp\left(-\frac{|\alpha|^2}{2}\right) \tag{4.13}$$

and

$$D = \sqrt{\frac{\pi}{2}} \cdot \left[1 + \mathrm{erf}\left(\frac{|\alpha|}{\sqrt{2}}\right)\right] , \tag{4.14}$$

where "erf" is the error function:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t \,. \tag{4.15}$$

The resulting histograms and fit curves are displayed in Figures 4.8 and 4.9 for 60 GeV and 120 GeV, respectively. One can see good agreement between the data histograms and the corresponding fit curves. The shapes around the maxima are well described by Equation (4.9) and minor deviations are only noticeable at the tails of the distributions. Furthermore, the vanishing correlation factors of these principal components can be seen in Figure 4.10 (also for 60 GeV and 120 GeV pions). As expected, all correlations are close to zero, except those on the diagonal which are equal to one.

The principal component distributions were used as input PDFs for a random number generator. With this, 100 000 events were randomly generated, each event containing eight values that correspond to eight simulated principal components. These events were then transformed back via Equations (4.7) and (4.8) into simulated energy differences.

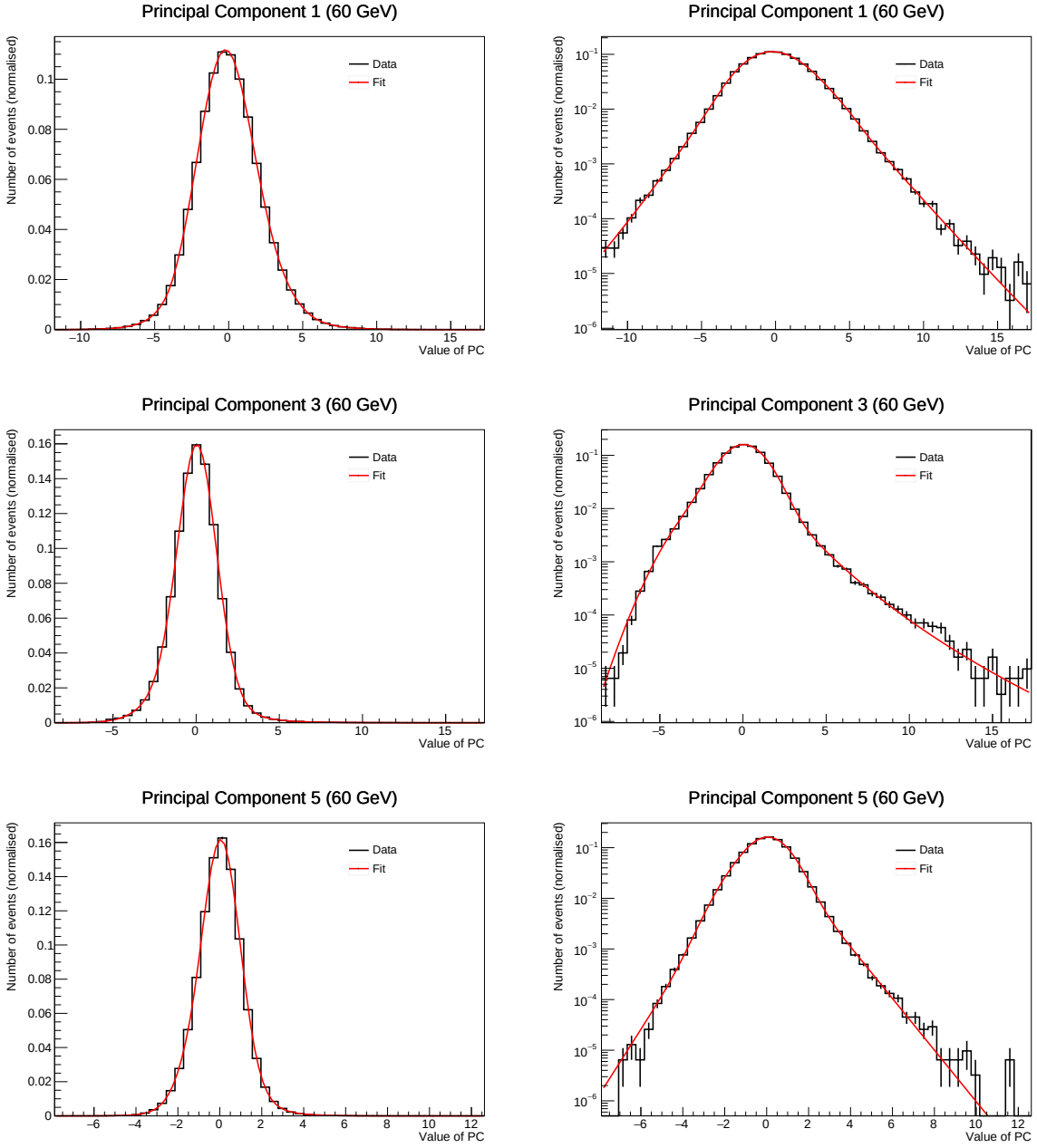


Figure 4.8.: PDFs of principal components for 60 GeV pions. The left column shows distributions and their corresponding fit functions for principal components 1, 3, and 5 on linear scale. The right column shows the same distributions but on logarithmic scale.

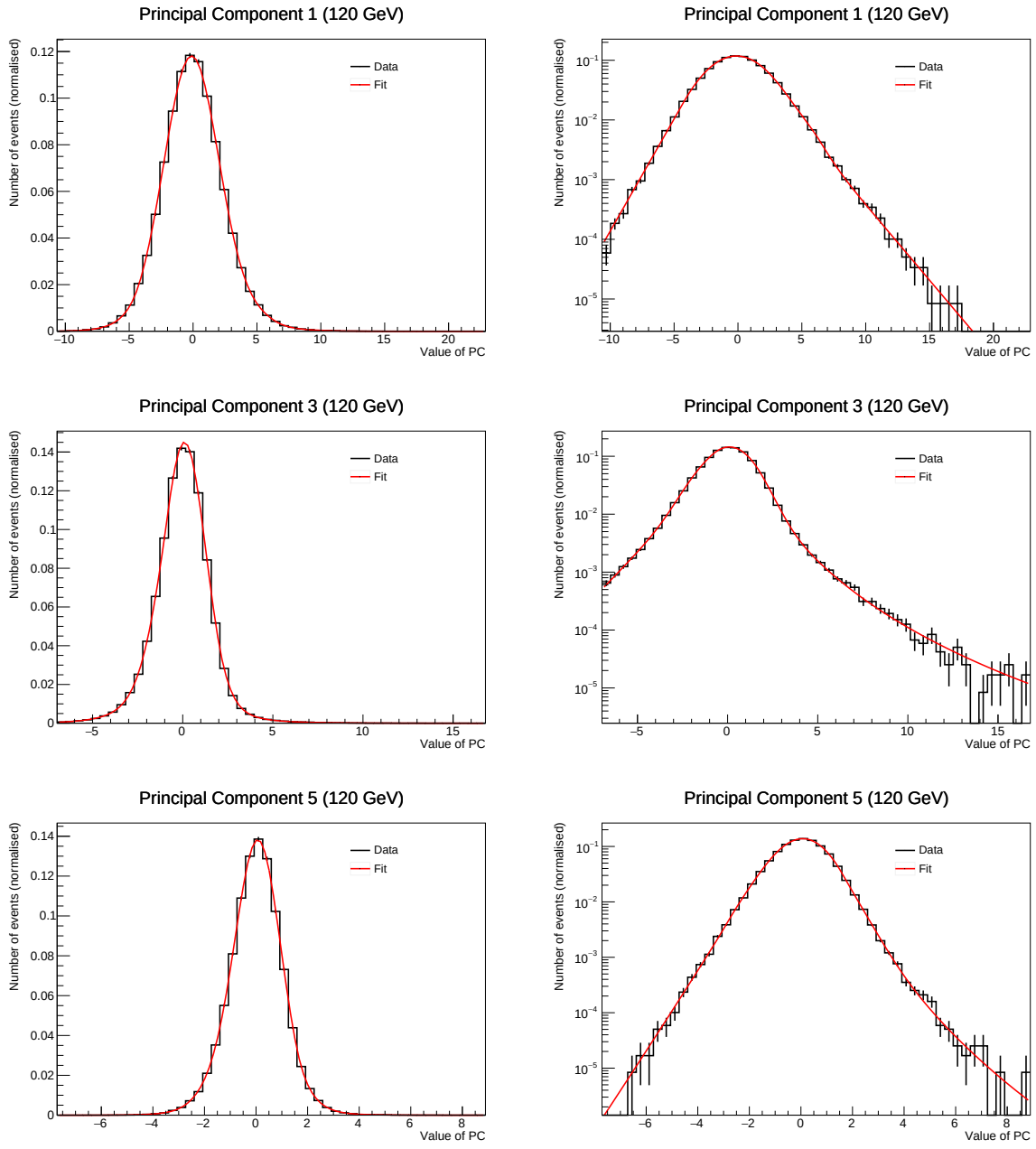


Figure 4.9.: PDFs of principal components for 120 GeV pions. The left column shows distributions and their corresponding fit functions for principal components 1, 3, and 5 on linear scale. The right column shows the same distributions but on logarithmic scale.

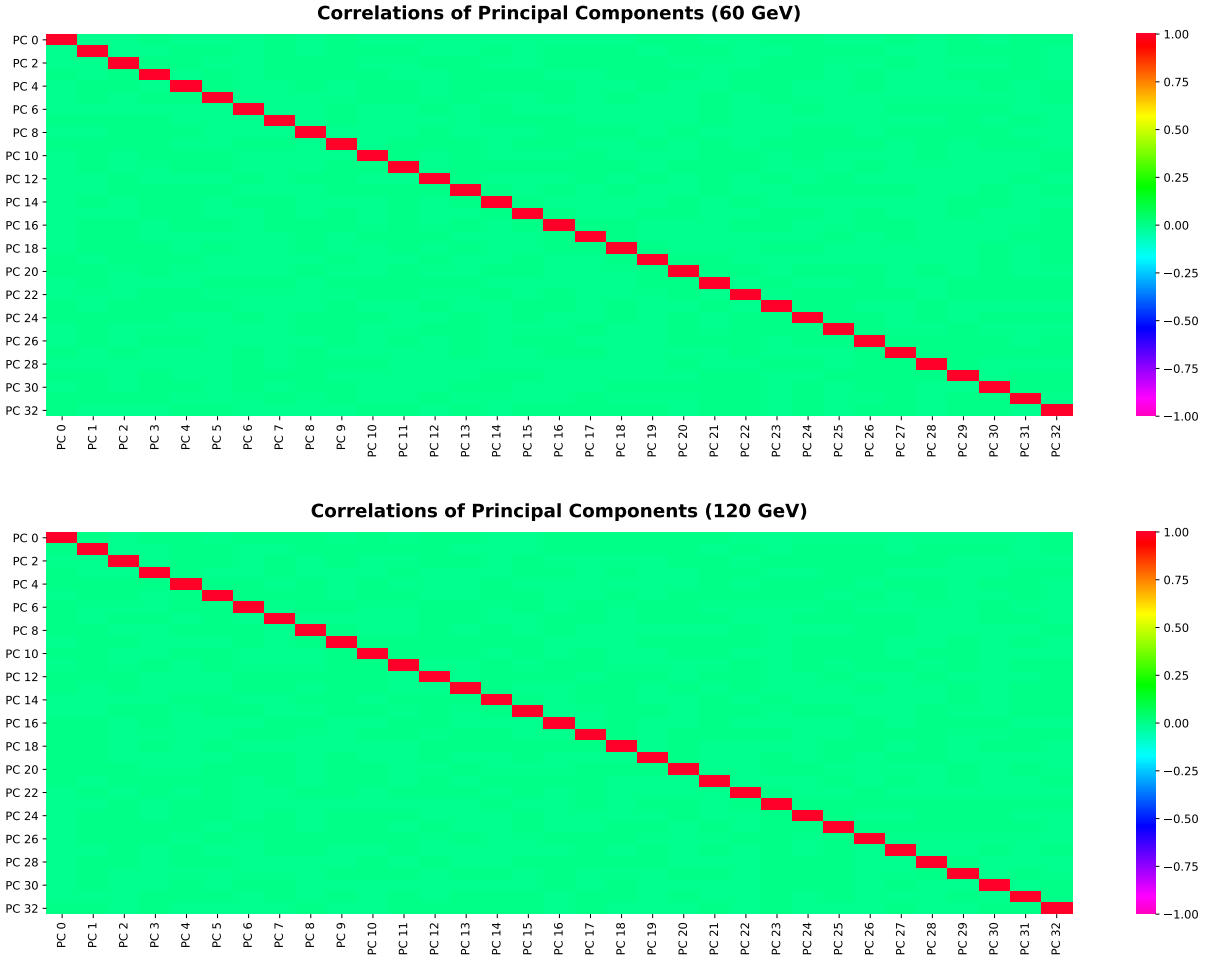


Figure 4.10.: Correlation factors between principal components for 60 GeV (upper plot) and 120 GeV (lower plot) pions. Only those correlation factors that lie on the diagonal are, as expected, equal to one; all others are close to zero.

4.3. Simulation of Individual Shower Energies

From the previously conducted PCA, 100 000 events (each event containing 33 simulated energy differences) were obtained. Based on these events, simulated energy differences were compared with data layerwise. For each detector layer, histograms of data and simulation were first normalised and then plotted together. The results of this procedure are shown in Figure 4.11 for 60 GeV and in Figure 4.12 for 120 GeV pions, both on linear and logarithmic scales. In both cases, one can see that the simulation does not agree very well with data. In particular, the simulated PDFs have shapes that are too broad around their respective maxima compared to those from data. Furthermore, the simulation is not able to recreate the correct fraction of negative energy differences, which causes its distributions to fall off too quickly for negative values. Lastly, all simulated PDFs include energy differences larger than the upper bounds of their respective data histograms. Such energy differences correspond to negative energies, which are unphysical values, and are therefore undesired.

One might argue now that the information loss due to neglecting 25 principal components in Section 4.2.2 might be the reason for the disagreement between data and simulation, or that the functional fits should have been used instead of the principal component PDFs in order to generate simulated principal components. To investigate this, the whole PCA has been conducted again twice, once without neglecting any principal components and once with functional fits as input for a random number generator. Energy differences were then simulated in the same manner as has been described previously, and the combined results are shown in Figure 4.13 for 60 GeV pions. By comparing all distributions in Figure 4.13, one can notice that the functional fits do not seem to have a significant impact on the goodness of the simulation, nor do they improve it in any other way. The distributions stemming from the analysis with all principal components, on the other hand, yield improvement, though it still does not match with expectations. One reason is that information about moments of second order or higher of the data energy difference distributions is lost during the PCA.

To check whether the simulation agrees on average with data, simulated average longitudinal energy distributions were obtained. The resulting distributions are depicted in Figure 4.14. It is clearly visible that the simulation is, at least on average, able to reproduce the data exactly because both the simulation and the data curve are congruent to each other. Similar results were also achieved for distributions of the centre of gravity of pion showers along the z-axis of the detector (the beam axis), which are shown in Figure 4.15. Here, good agreement between data and simulation is visible, although the simulation is not exactly able to reproduce the expected distributions.

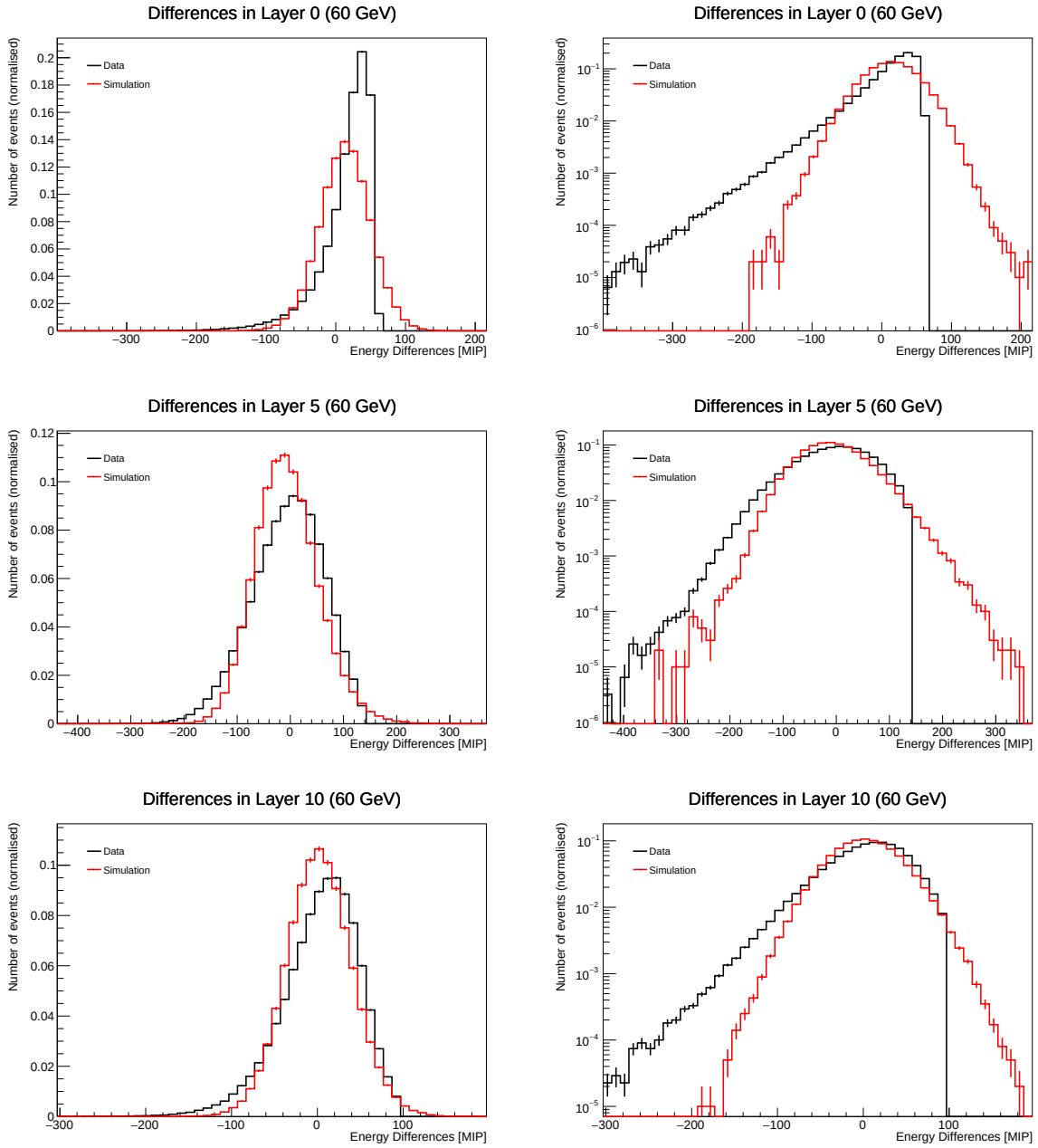


Figure 4.11.: Comparison of 60 GeV energy difference PDFs between data (black) and simulation (red). The left column shows energy difference distributions in layers 0, 5, and 10 on linear scale, whereas the right column shows the same plots on logarithmic scale. These histograms were obtained from a PCA where only eight principal components were considered.

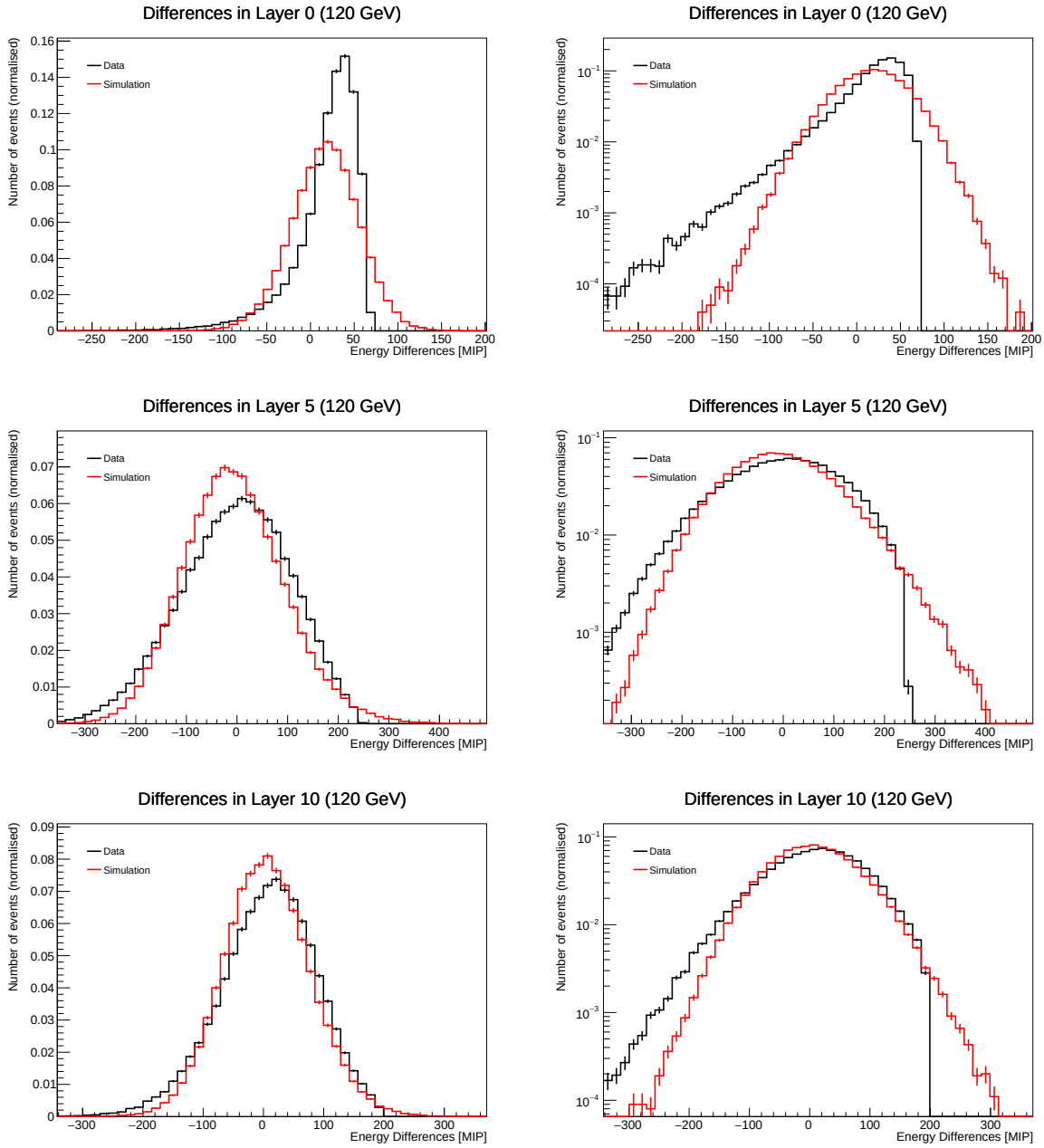


Figure 4.12.: Comparison of 120 GeV energy difference PDFs between data (black) and simulation (red). The left column shows energy difference distributions in layers 0, 5, and 10 on linear scale, whereas the right column shows the same plots on logarithmic scale. These histograms were obtained from a PCA where only eight principal components were considered.

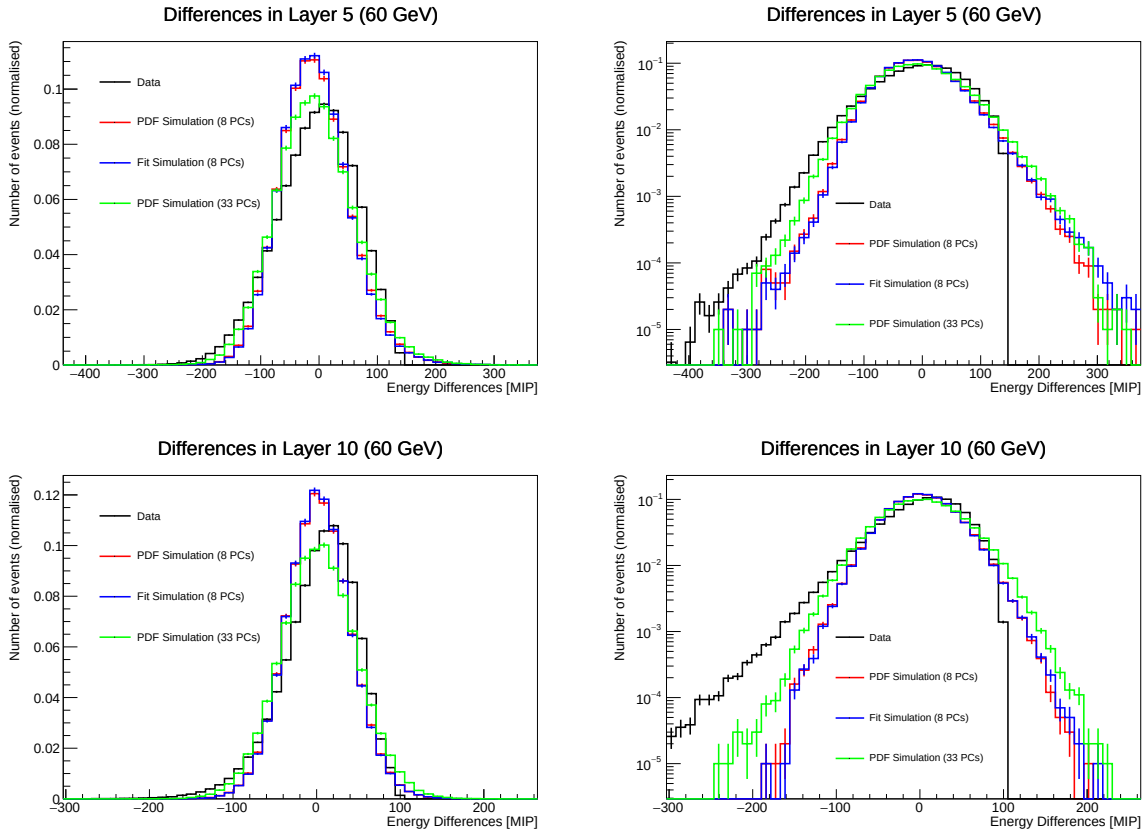


Figure 4.13.: Comparison of 60 GeV energy difference PDFs between data (black), simulations obtained from principal component PDFs with eight (red) and all principal components (green), and simulations with eight principal components obtained from functional fits (blue). The left column shows energy difference distributions in layers 5 and 10 on linear scale, whereas the right column shows the same plots on logarithmic scale. The red curve is almost not visible because it is congruent to the blue one.

Next to average longitudinal energy distributions, simulated correlation factors were compared with data, which can be seen in Figure 4.16 for 60 GeV and in Figure 4.17 for 120 GeV pions. One can already see here that the simulated correlation factors do not agree well with data. In particular, positive simulated correlation factors are too large, whereas negative ones are too small, in comparison with the expectation from data. This becomes even clearer by subtracting the data correlation matrix from that obtained from simulations. The result is another matrix whose entries are correlation differences, ΔC, of energy differences in layers x and y:

$$\Delta C = \mathrm{Corr}_{\mathrm{sim}}(x, y) - \mathrm{Corr}_{\mathrm{data}}(x, y)\,. \tag{4.16}$$

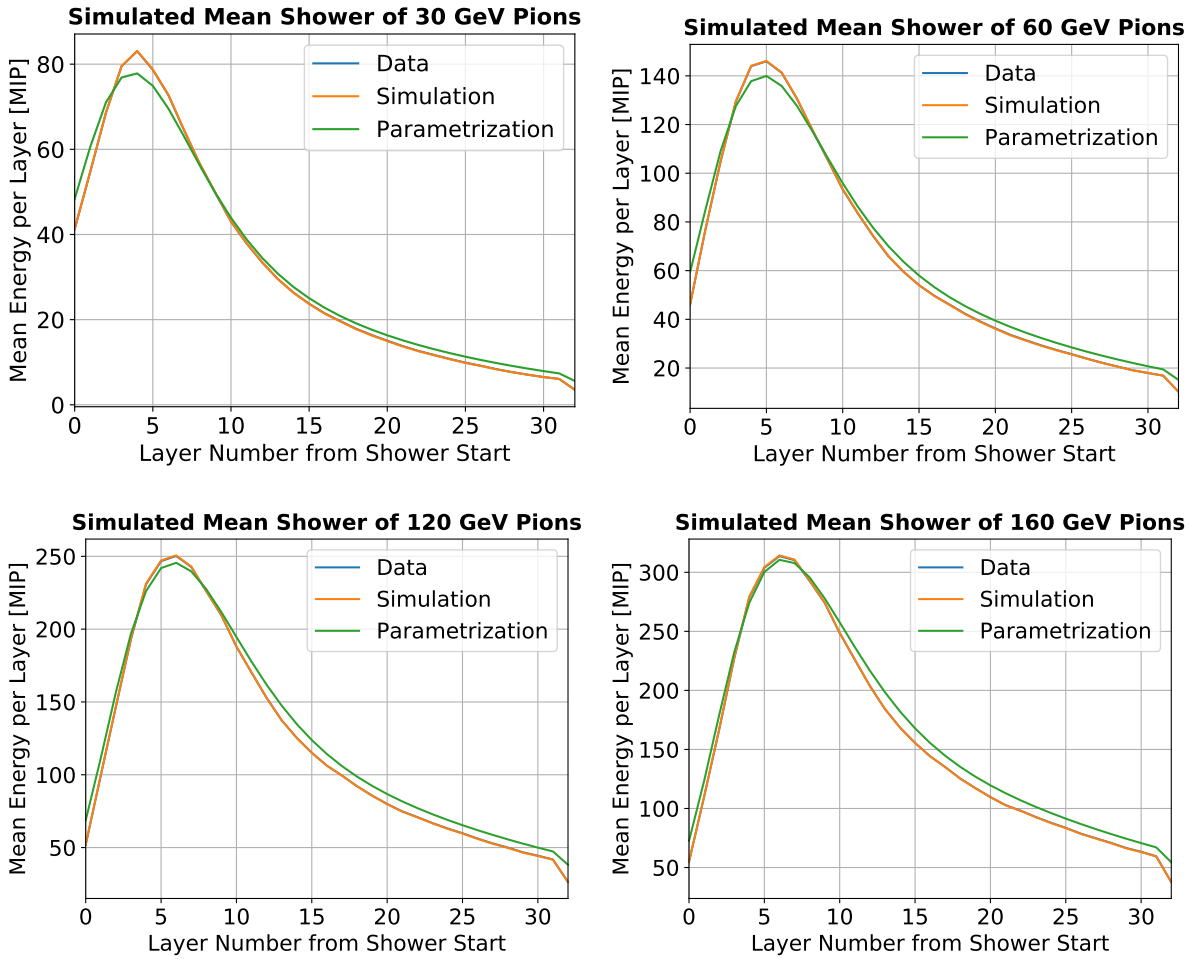


Figure 4.14.: These plots show average longitudinal energy depositions of pion showers, obtained from data (simulation), in blue (orange) as a function of the detector layer. In addition, Equation (4.1) is plotted too (green) in order to compare it with data and simulation. The data curve is not visible because it lies exactly beneath the simulation curve.

These matrix elements were visually displayed of which those from 60 GeV and 120 GeV pions are, as examples, shown in Figure 4.18. One can notice that for low initial pion energies differences between simulation and data are more pronounced, and the simulation preserves correlation factors better for higher initial energies than for lower.

To summarise, the PCA was able to simulate average longitudinal energy distributions of pion showers, but did not give accurate simulations at the single-shower level. Both the distributions of simulated energy differences as well as the simulated correlation matrices did not yield good agreement with data and deviated too much from their corresponding expectations that were obtained from the dataset. Therefore, other methods for pion shower simulation have to be investigated.

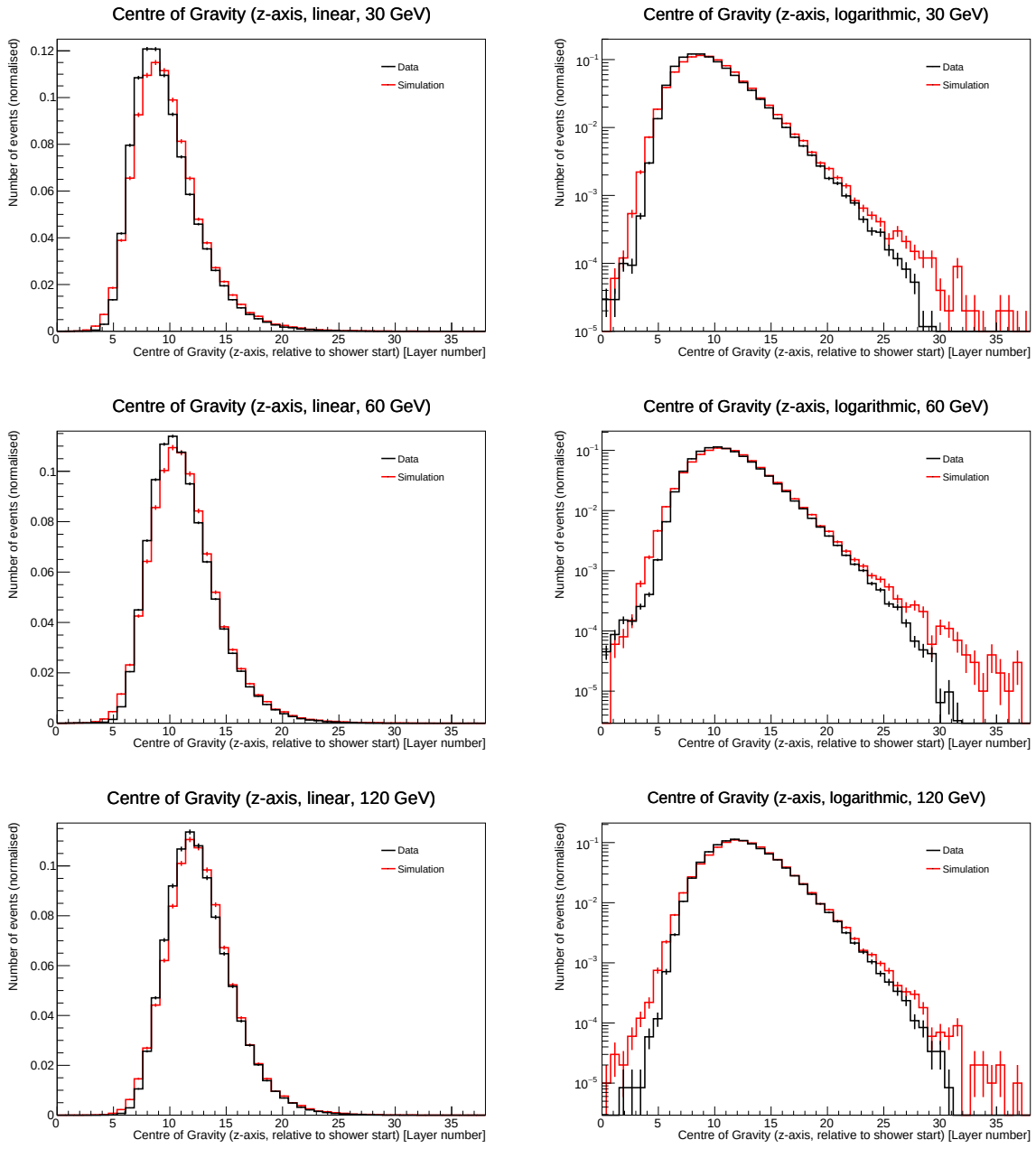


Figure 4.15.: Comparisons of centre of gravity (z-axis) distributions between data (black) and simulation (red) for 30 GeV, 60 GeV, and 120 GeV pions. Apart from smaller deviations, the distributions are in good agreement with each other.

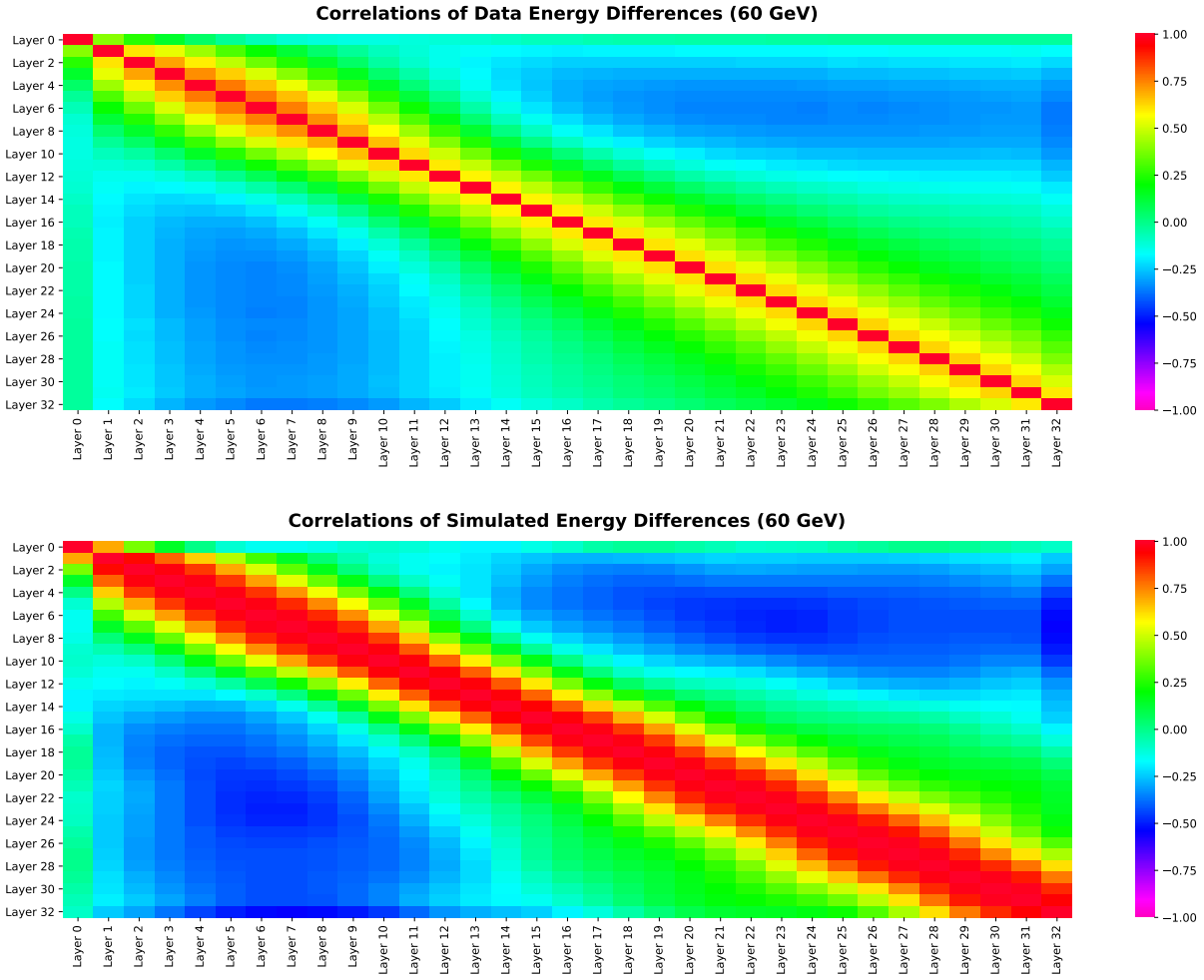


Figure 4.16.: Comparison of correlation factors between data (upper plot) and simulation (lower plot) for 60 GeV pions. The simulated (anti-)correlations are much stronger than their expectations from data.

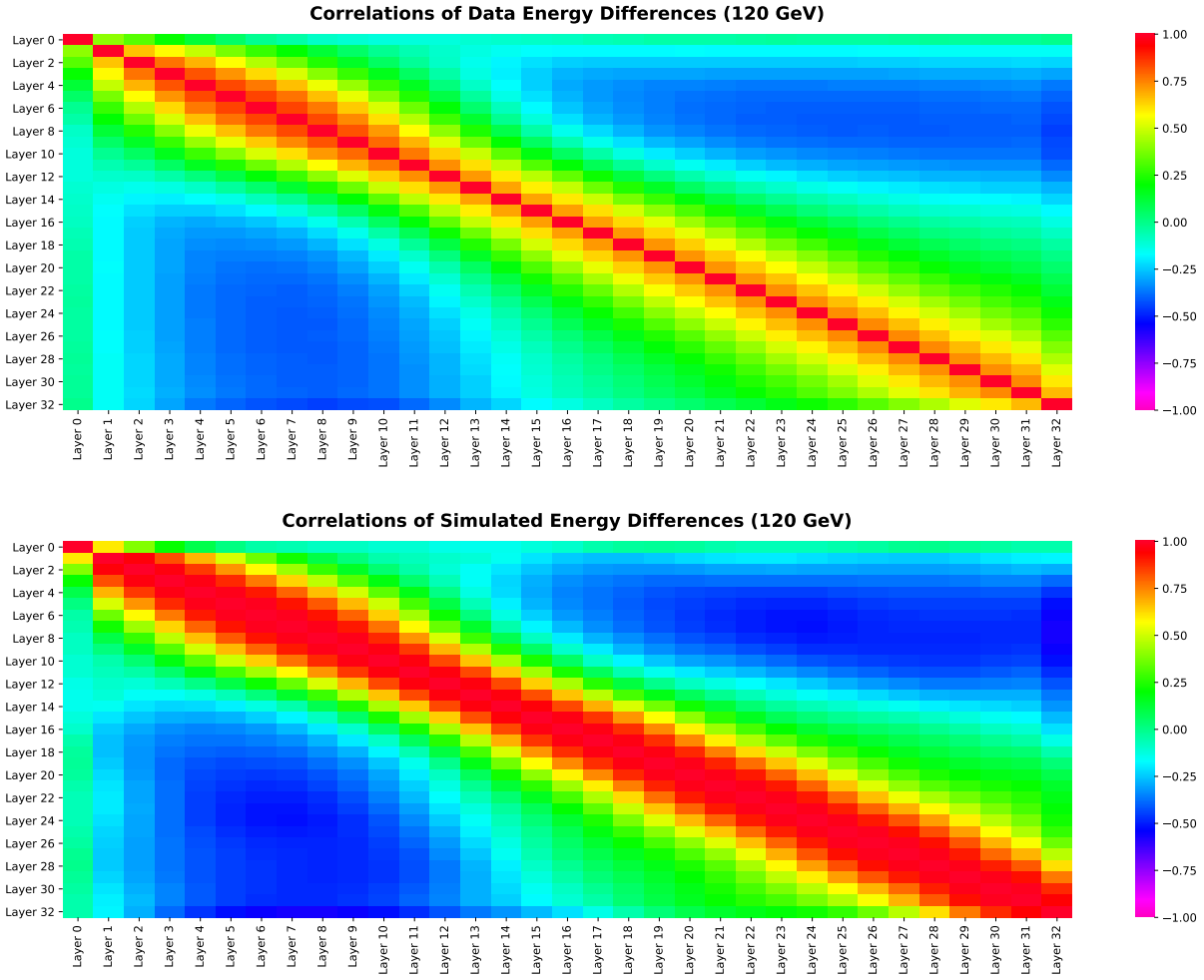


Figure 4.17.: Comparison of correlation factors between data (upper plot) and simulation (lower plot) for 120 GeV pions. The simulated (anti-)correlations are much stronger than their expectations from data.

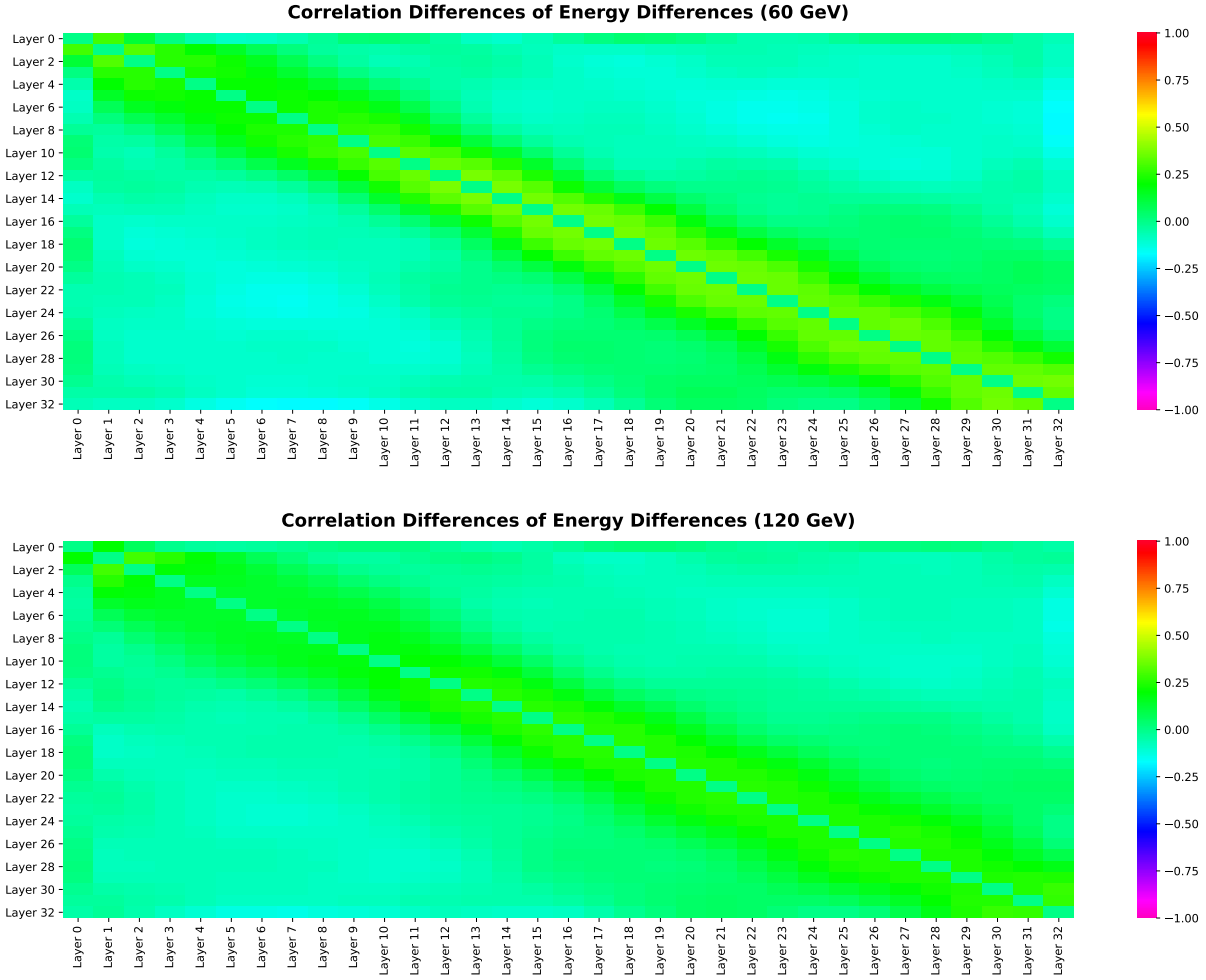


Figure 4.18.: Correlation differences, ΔC, calculated via Equation (4.16), between simulation and data for 60 GeV (upper plot) and 120 GeV (lower plot) pions. These heatmaps show that for low initial energies simulated correlations differ more from data than for higher energies.

5. Longitudinal Simulation of Pion Showers using Kernel Density Estimators

Next to the PCA presented in the previous Chapter, a second method for simulating pion showers longitudinally was investigated, namely the application of kernel density estimators (KDEs) to the already computed energy differences. From the PDFs that were obtained from this procedure, 100 000 events were again simulated. Each event contains the same 33 energy differences, which were then compared with data.

The theoretical background behind KDEs is introduced in Section 5.1. After this, Section 5.2 explains how KDEs were applied to energy differences obtained from data, as already introduced in Chapter 4, and how estimated PDFs were then used to simulate energy difference distributions. The simulations themselves and their correlation factors were then compared to data, as well as histograms of two kinematic shower variables: the total energy of the pion shower and the centre of gravity along the z-axis (COGZ) of the detector (the shower axis).

For this analysis, the pion dataset was split into equally sized parts. One part (the training sample) was used as basis upon which KDEs were built. The other part (the validation sample) was then used for comparisons of simulation with data. Therefore, all following figures that show such comparisons only include data from the validation sample.

5.1. Kernel Density Estimators

When dealing with large datasets, one may, for many reasons, not always be able to describe randomly distributed data points with an analytical PDF. Usually, when this is the case, it is because the actual underlying mathematical shape of the distribution is unknown or does not even exist. Instead, one has to rely on methods of estimating PDFs of such random variables. This can be done with the help of KDEs.

Consider a set of n data points (x_1, x_2, ..., x_n), for example results of a measurement of a certain physical variable, x, repeated n-times whose underlying PDF is unknown. In order to estimate this unknown distribution, its KDE is defined in the following way:

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) . \tag{5.1}$$

Here, the left-hand side of the equation represents the estimated distribution of the dataset. The right-hand side includes a parameter $h > 0$ called the bandwidth and a sum of kernels, K, running over all data points x_i. The kernel function K can be any non-negative density function that describes a single data point adequately. For this thesis, a Gaussian normal distribution was used:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) . \tag{5.2}$$

Hence, the right-hand side of Equation (5.1) is a sum of normal distributions centred around each data point x_i, normalised to the bandwidth h, and all of it scaled to the product of bandwidth and number of data points. Figure 5.1 shows an example of how this estimation works graphically. In this plot, ten data points and their corresponding centred and normalised Gaussian kernels are shown (red curves). The areas under the kernels are all equal to one. Furthermore, the estimated PDF, computed via Equation (5.1), is shown in blue, also normalised to unity. In this example, the bandwidth was chosen to be equal to five.

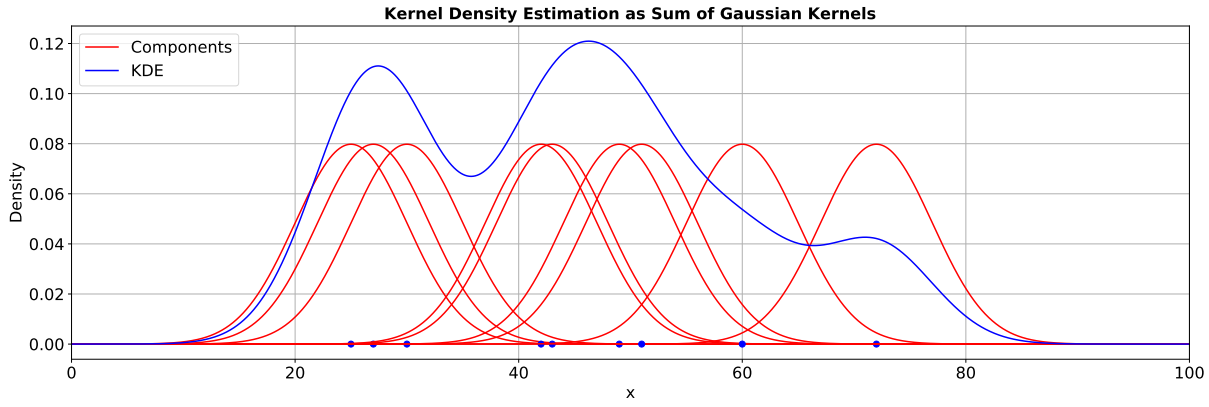


Figure 5.1.: An example of a PDF estimation using Gaussian kernels. First, ten data points were generated randomly. Then, Gaussian distributions were centred around them and normalised to the bandwidth (red curves). Finally, all Gaussian kernels were added up, yielding the final PDF (blue curve). For this example, $h = 5$ was chosen.

The bandwidth is an important parameter because it has significant impact on the smoothness of the estimation. If the bandwidth is chosen to be too small, every data point of the dataset becomes visible as a peak in the final distribution, which results in a curve that is not smooth enough. On the contrary, if h is too large, then the bandwidth obscures much of the underlying structure of the PDF by flattening it too much. Examples of how strongly results might differ, depending on the choice of bandwidth, are shown in Appendix A.

Equation (5.1) can be easily generalised to d dimensions. Instead of n single values, one is now dealing with a set of n d-dimensional vectors ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$). The corresponding KDE is then defined as

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} |\mathbf{H}|^{-\frac{1}{2}} K\left(\mathbf{H}^{-\frac{1}{2}}(\mathbf{x}-\mathbf{x}_i)\right) , \tag{5.3}$$

where $\mathbf{H}$ is the symmetric, positive definite $d \times d$ bandwidth matrix and $|\mathbf{H}|$ is its corresponding determinant. Furthermore, $\mathbf{H}^{-\frac{1}{2}}$ is the inverse of the square root of $\mathbf{H}$. For this thesis, a standard multivariate distribution has been used as kernel function in this generalised case:

$$K(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}\right) . \tag{5.4}$$

Here, $\mathbf{x}^{\mathrm{T}}$ is the transposed of vector $\mathbf{x}$. In principle, $\mathbf{H}$ can be chosen to be any symmetric, positive definite matrix. However, the amount of parameters that need to be chosen grows with $\frac{d(d+1)}{2}$. That is why it is often convenient to chose a more simplified form for $\mathbf{H}$. For this thesis, the bandwidth matrix

$$\mathbf{H} = h^2\mathbf{C} \tag{5.5}$$

has been used, where h is an arbitrarily chosen bandwidth and $\mathbf{C}$ is the covariance matrix of the training sample as defined by Equation (4.5). Substituting Equation (5.5) into Equation (5.3) then yields the KDE definition that has been applied to the training sample of this analysis:

$$f(\mathbf{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} |\mathbf{C}|^{-\frac{1}{2}} K\left(\frac{\mathbf{C}^{-\frac{1}{2}}(\mathbf{x}-\mathbf{x}_i)}{h}\right) . \tag{5.6}$$

5.2. Simulation of Individual Shower Energies

In order to simulate energy differences, as defined in Chapter 4, Equation (5.6) was applied layerwise to histograms of energy differences. As the bandwidth, $h = 0.01$ was chosen. The resulting PDFs were then used to generate 100 000 events, each containing 33

simulated energy differences (one for each layer). To make comparisons with data, these simulated events were visually displayed. The resulting PDFs are shown in Figure 5.2 for 60 GeV and in Figure 5.3 for 120 GeV pions. Both the data as well as the simulation curves are in very good agreement, for all detector layers and for all initial energies.

Theoretically, these results allow for a layerwise investigation of the energy deposition within layers 32 to 38. For this purpose, one would have to infer the energy difference PDF of each layer from the PDF of the combined variable. This can be done as follows. For each layer i ($32 \leq i \leq 38$), Equation (4.1) has to be integrated from i to $i+1$, giving seven areas, A_i, in total, corresponding to the seven layers that are being considered. After that, every layer has to be divided by the total area of Equation (4.1) between layers 32 and 38. Each of these results then represents the fraction of energy, f_i, deposited within layer i, of the total energy that is distributed between layers 32 and 38 according to the longitudinal profile of an average shower. By multiplying the combined energy difference ΔE_{32-38} with each fraction f_i, one obtains the energy difference ΔE_i in layer i. By doing this for every event, individual energy difference PDFs of layers 32 to 38 can be created, whose shapes are similar to the PDFs of all other layers.

In addition to Figures 5.2 and 5.3, one can compare (anti-)correlations between data and simulation in Figure 5.4 for 60 GeV and in Figure 5.5 for 120 GeV pions. Both Figures demonstrate that the KDE is also able to recreate the correlation factors between detector layers almost exactly. To emphasise this even more, one can calculate differences in correlation factors between data and simulation, according to Equation (4.16). The results of this procedure are also shown for 60 GeV and 120 GeV pions in Figure 5.6. Since all possible correlation differences are very close to zero, these plots clearly show that all linear correlation factors are correctly simulated.

Lastly, kinematic shower variables were compared between data and simulation to ensure correct kinematic behaviour of the simulation. In particular, the shower's total energy (in MIPs) and its COGZ (measured from the shower start and given in layers) were computed. The total energy was calculated via

$$E_{\text{tot}} = \sum_{\text{layers } i} \{\Delta E_i + E_A(i)\}\,, \tag{5.7}$$

where the sum runs over all detector layers for a certain event. Furthermore, ΔE_i is the energy difference in layer i and $E_A(i)$ is the value of Equation (4.1) in the same layer. The COGZ, on the other hand, was then determined via

$$\text{COGZ} = \frac{1}{E_{\text{tot}}} \sum_{\text{layers } i} \{\Delta E_i + E_A(i)\} \cdot i\,. \tag{5.8}$$

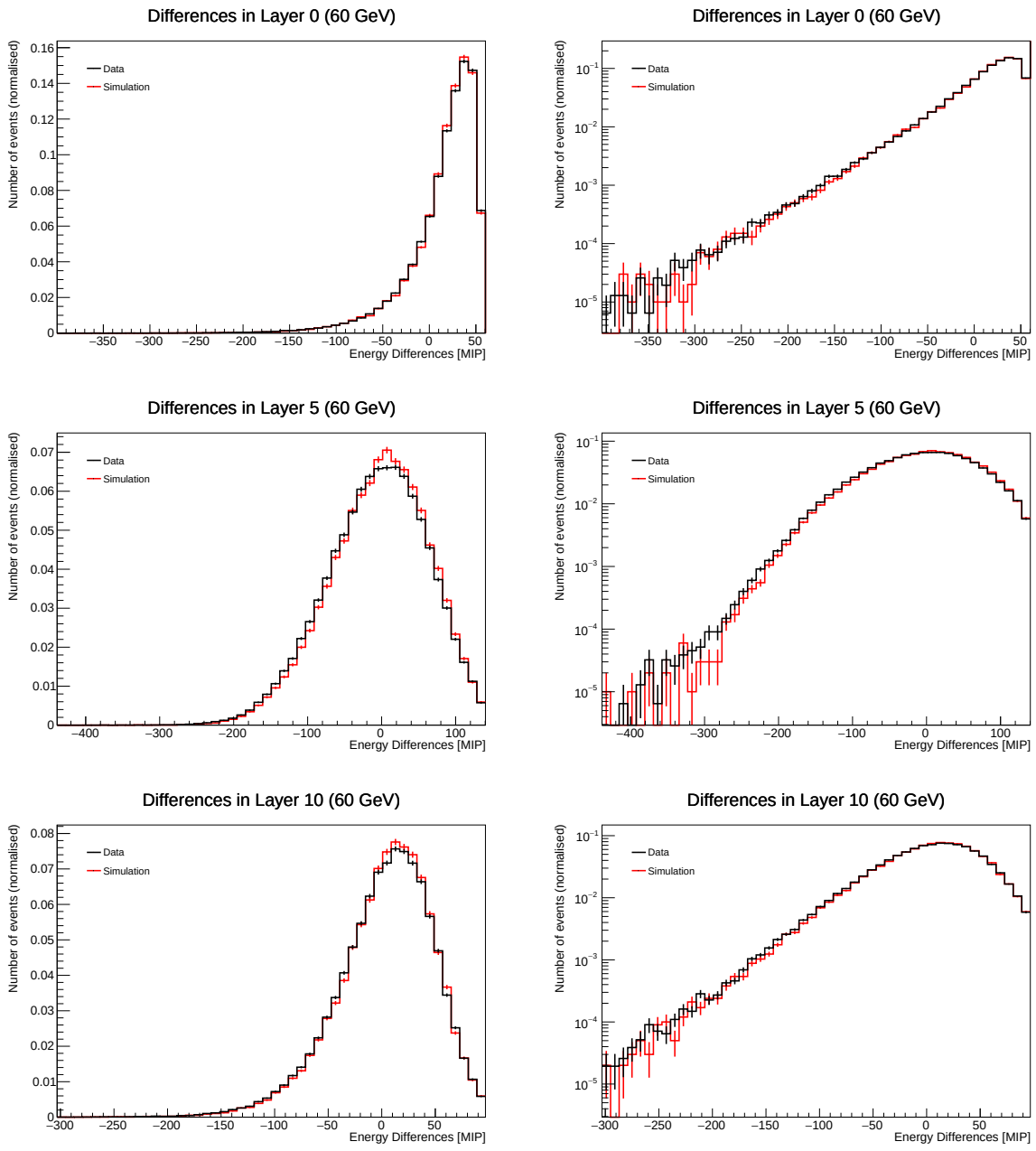


Figure 5.2.: Comparison of 60 GeV energy difference distributions between data (black) and simulations obtained from KDEs (red) for layers 0, 5, and 10. All layers exhibit very good agreement between data and simulation, both on linear (left) as well as on logarithmic scale (right).

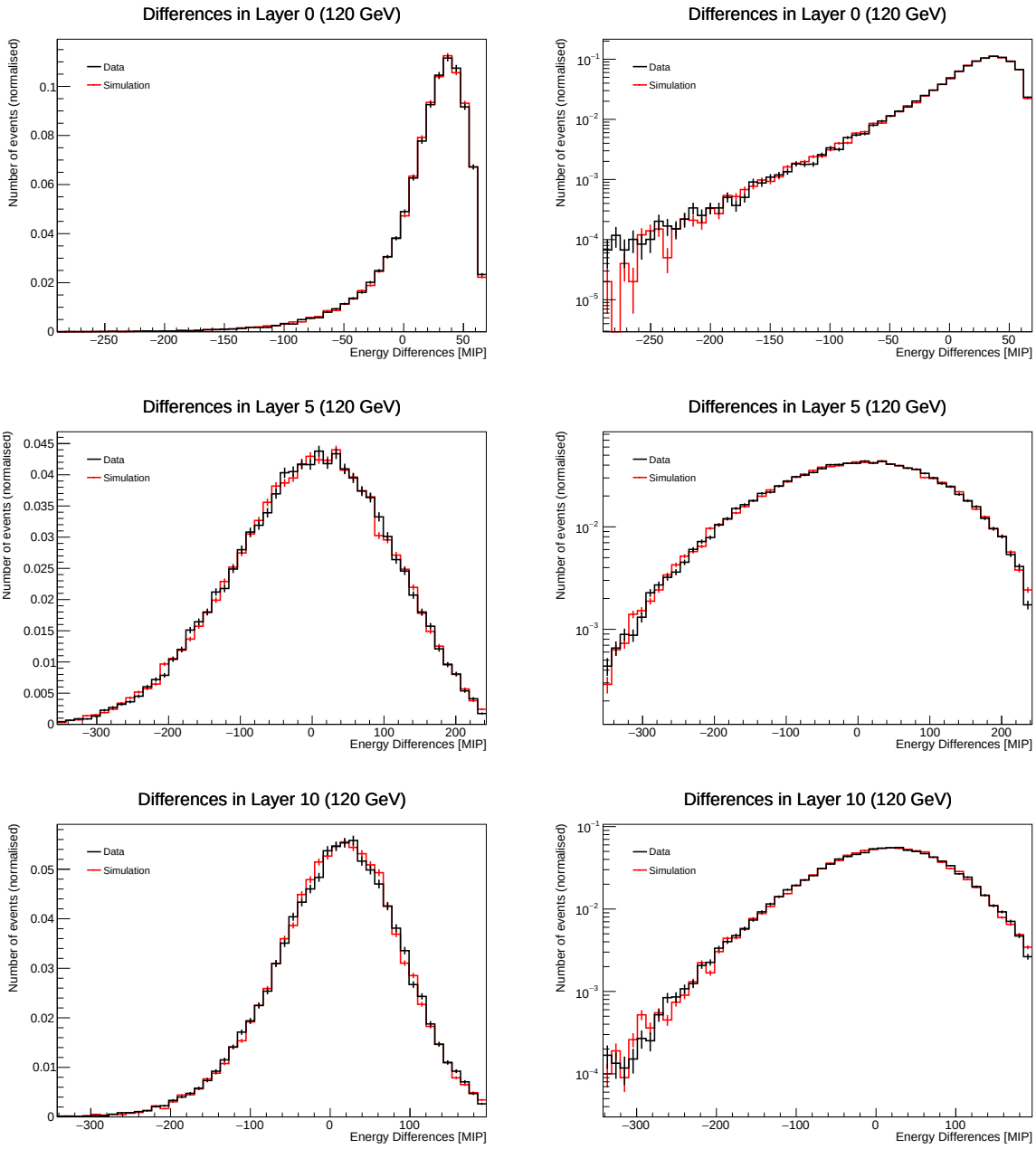


Figure 5.3.: Comparison of 120 GeV energy difference distributions between data (black) and simulations obtained from KDEs (red) for layers 0, 5, and 10. All layers exhibit very good agreement between data and simulation, both on linear (left) as well as on logarithmic scale (right).

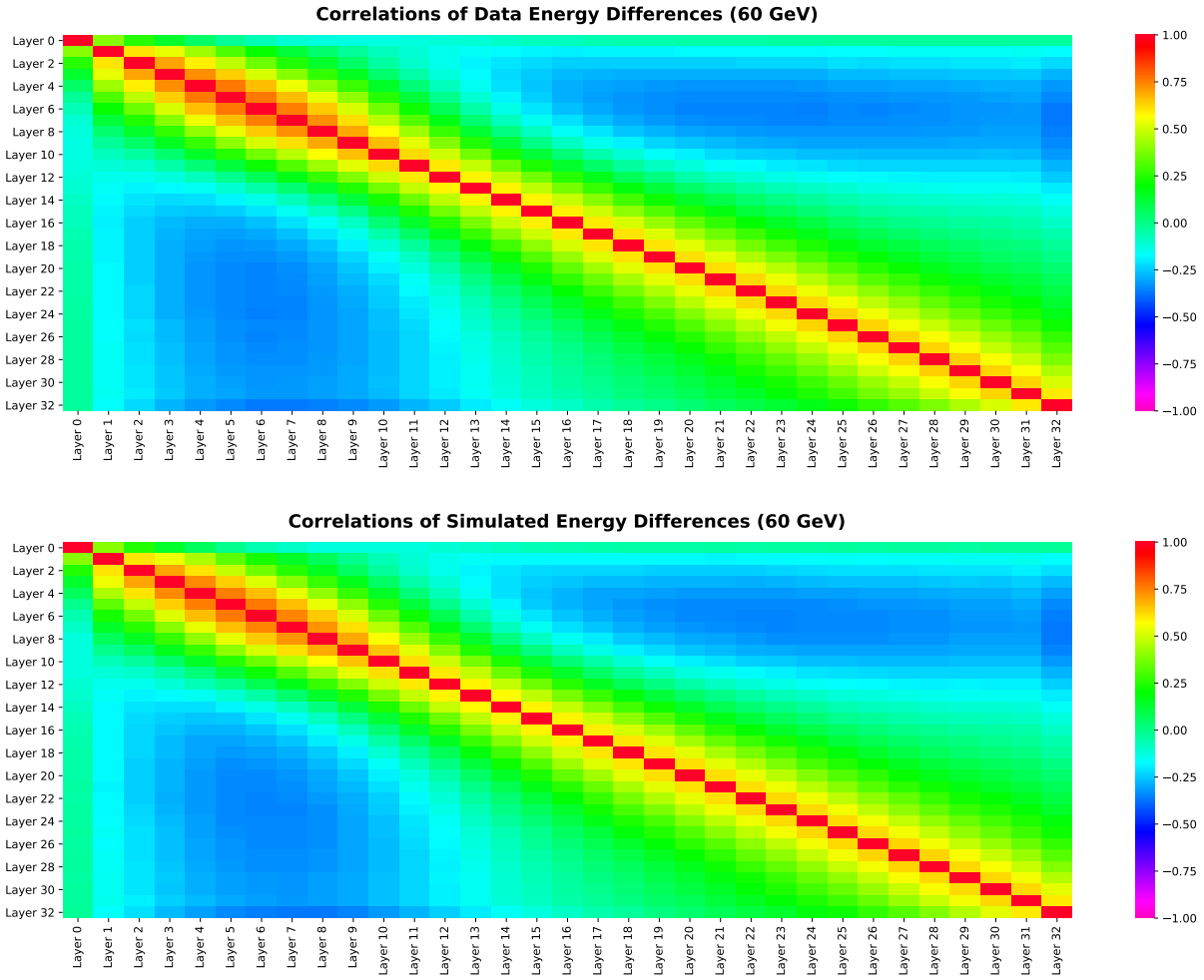


Figure 5.4.: Comparison of correlation factors between data (upper plot) and simulations obtained from KDEs (lower plot) for 60 GeV pions. Both plots show very good agreement between their correlation factors.

It is hence the weighted sum of all absolute energies per layer, weighted by their respective layer number and divided by the total energy of the event. PDFs of these kinematic variables are shown in Figure 5.7 (total energies) and in Figure 5.8 (COGZs) for various initial energies. The distributions of both kinematic variables exhibit very good agreement between data and simulation and show that the simulation reproduces a real shower's behaviour accurately.

Apart from the very good agreement between data and simulation, one can notice shoulders in all plots on the right-hand side in Figure 5.7 (logarithmic scale). These only appear at energies far above the beam energy and are of the order of approximately 0.1 % of all events. Furthermore, they are much more pronounced at medium energies, such as for example 40 GeV, than at very small or very large pion energies. There may be

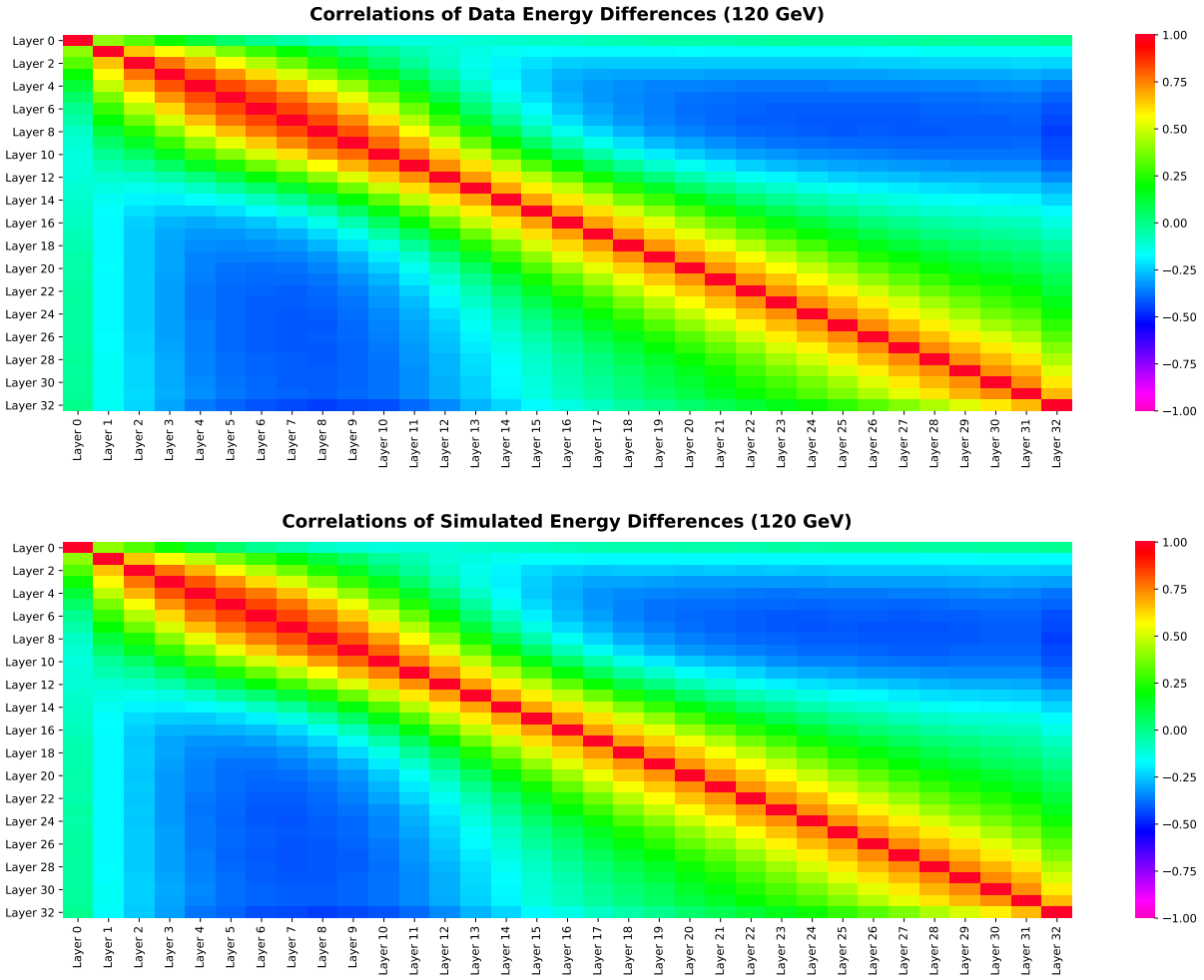


Figure 5.5.: Comparison of correlation factors between data (upper plot) and simulations obtained from KDEs (lower plot) for 120 GeV pions. Both plots show very good agreement between their correlation factors.

many causes for these additional peaks, the most significant are probably two or more simultaneously detected particles, beam contamination, or the rate at which test beam particles enter the detector. What might have caused these peaks, however, was not further investigated for this thesis.

In summary, a very accurate, data-based simulation of energy differences was designed. The simulation exhibits very good agreement with the data, and correlation factors between different calorimeter layers are correctly simulated too. The latter becomes apparent by looking at correlation differences between data and simulation which, as has been shown, are all close to zero. Furthermore, kinematic variables of simulated pion showers are also in very good agreement with their expectations. In particular, the total energy and the centre of gravity along the beam axis of simulated pion showers have been inves-

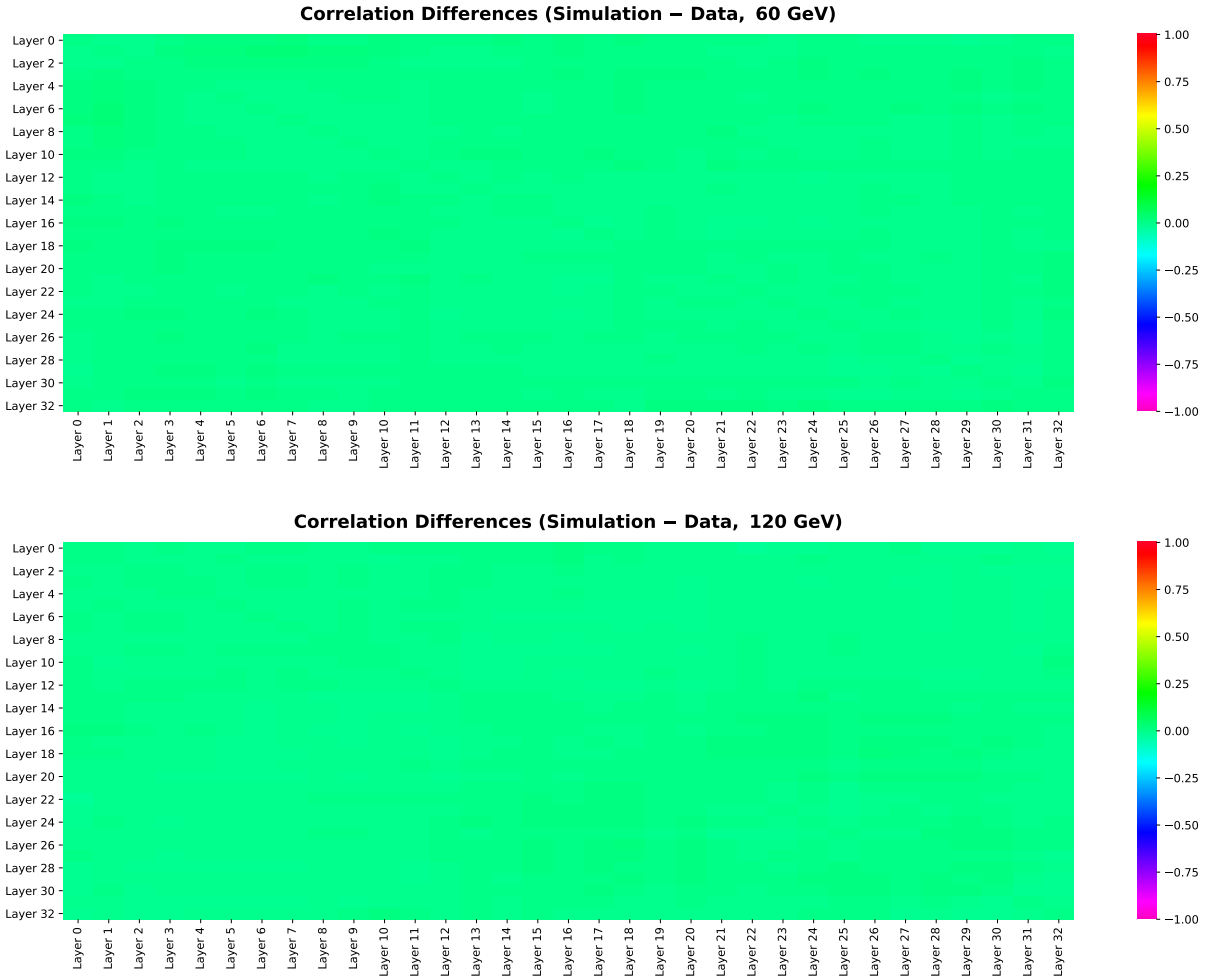


Figure 5.6.: Correlation differences, ΔC, between simulation and data for 60 GeV (upper plot) and 120 GeV (lower plot) pions. Both heatmaps show correlation differences very close to zero for all possible layer combinations.

tigated. Their distributions match their data counterparts almost exactly, and deviations are only minor.

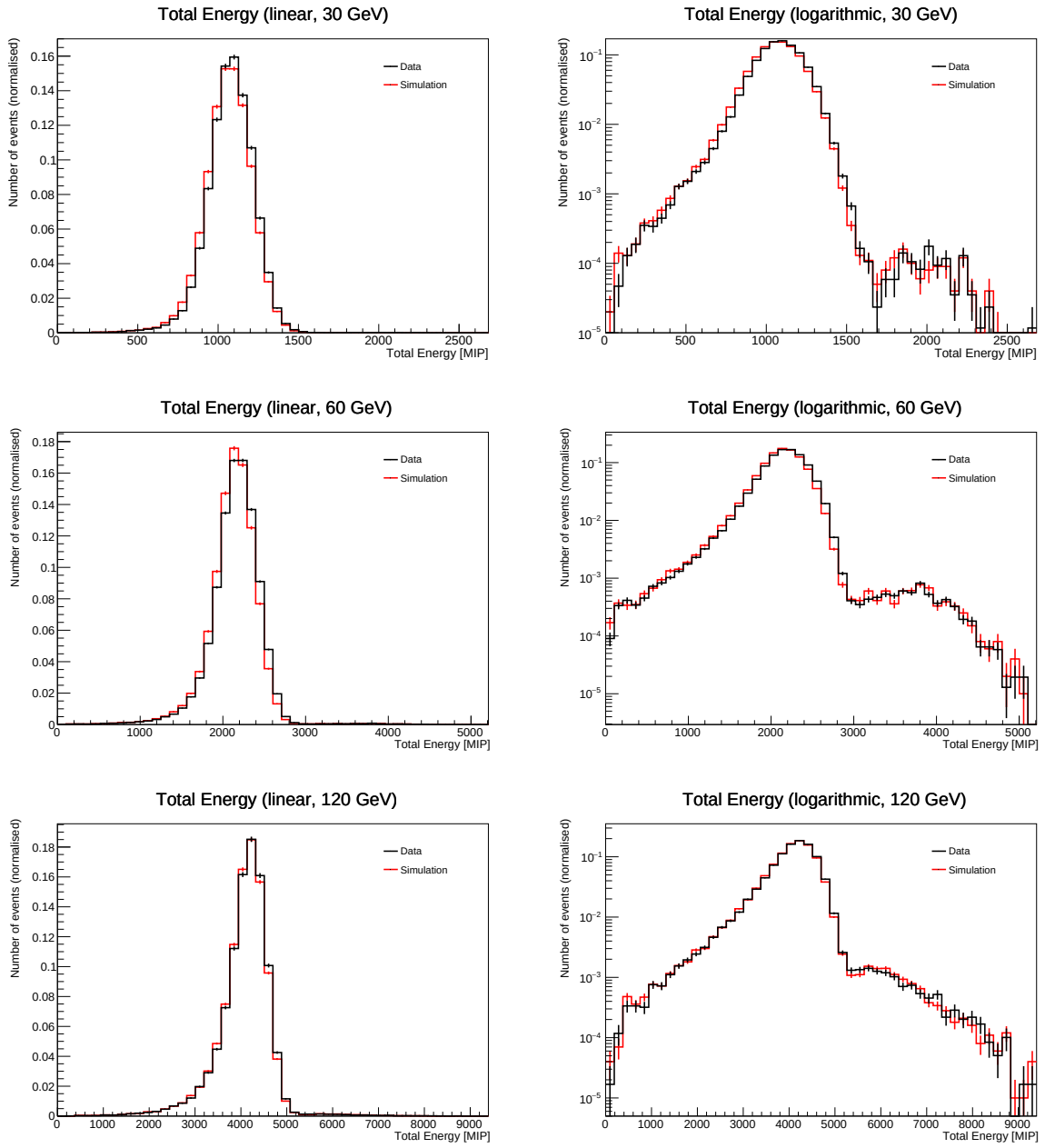


Figure 5.7.: Comparison of total energy distributions between data (black) and simulation (red) for 30 GeV, 60 GeV, and 120 GeV pions. All energies exhibit very good agreement between data and simulation.

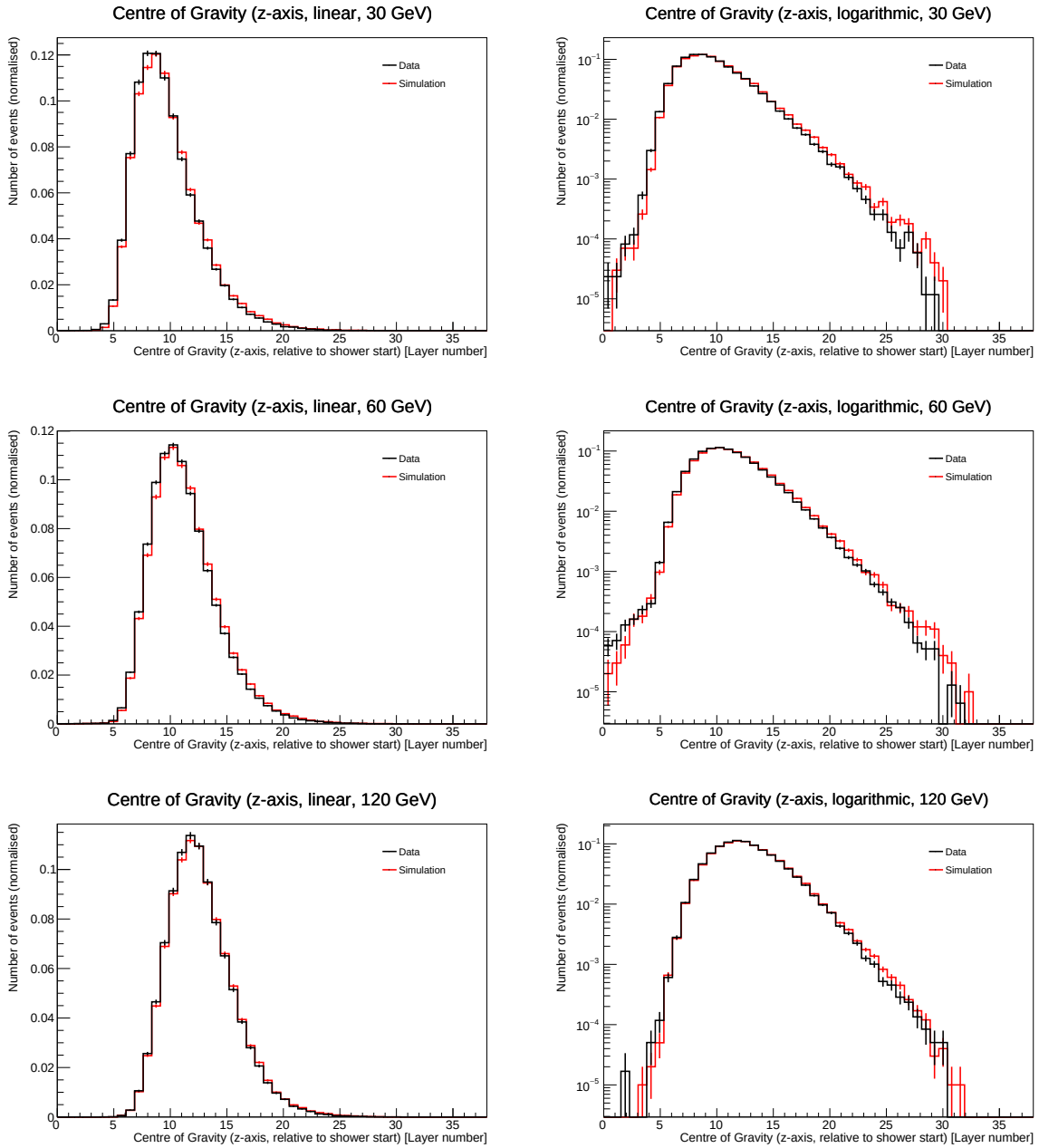


Figure 5.8.: Comparison of COGZ distributions between data (black) and simulation (red) for 30 GeV, 60 GeV, and 120 GeV pions. All energies exhibit very good agreement between data and simulation.

6. Interpolation Studies of Longitudinal Energy Distributions of Pion Showers

Since the simulations presented in the previous Chapter showed very good agreement with their expectations, interpolations of simulated pion showers between different initial energies are now investigated. Interpolations are an important part of a fast simulation of pion showers because once it is finished, this data-based fast simulation should be able to predict the behaviour of pion showers at any initial energy, not only those from the pion shower dataset. However, since one has to ensure that the interpolation works as expected, it was only conducted between energies of the pion shower dataset for this thesis.

This Chapter presents the results of how the KDEs introduced in Chapter 5 were used to interpolate simulated longitudinal energy distributions of single pion showers between various initial energies. The method and the mathematical procedures behind the interpolation are first introduced in Section 6.1. After that, interpolated longitudinal energy distributions are shown and compared with simulated longitudinal energy distributions, that were obtained from KDEs, in Section 6.2.

6.1. Mathematical Approach for Energy Interpolations

The objective of an interpolation is not only to predict distributions of energy differences correctly, but also to preserve the (anti-)correlations between the simulated energy differences for each calorimeter layer as precisely as possible. For this reason, the interpolation was done in the following way.

In order to interpolate simulated longitudinal energy distributions to a certain target energy, $E_{\text{interpolate}}$, the cumulative longitudinal energy difference distributions of two energies, E_{small} and E_{large}, which are equidistant from $E_{\text{interpolate}}$ ($E_{\text{small}} < E_{\text{interpolate}} < E_{\text{large}}$),

were used. A single event, containing 33 energy differences ($\Delta E_{\text{small, 0}}$, $\Delta E_{\text{small, 1}}$, ..., $\Delta E_{\text{small, 32}}$) was first randomly generated, according to the KDE of E_{small}. Then, for each layer i, the value of the cumulative longitudinal energy difference PDF at $\Delta E_{\text{small, }i}$ was determined. Since this was done for all 33 PDFs, a set of 33 real numbers between zero and one (A_0, A_1, ..., A_{32}) was obtained. This procedure corresponds to integrating the longitudinal energy difference PDF in layer i from the left (the smallest bin) to $\Delta E_{\text{small, }i}$, as schematically shown in the upper plot in Figure 6.1.

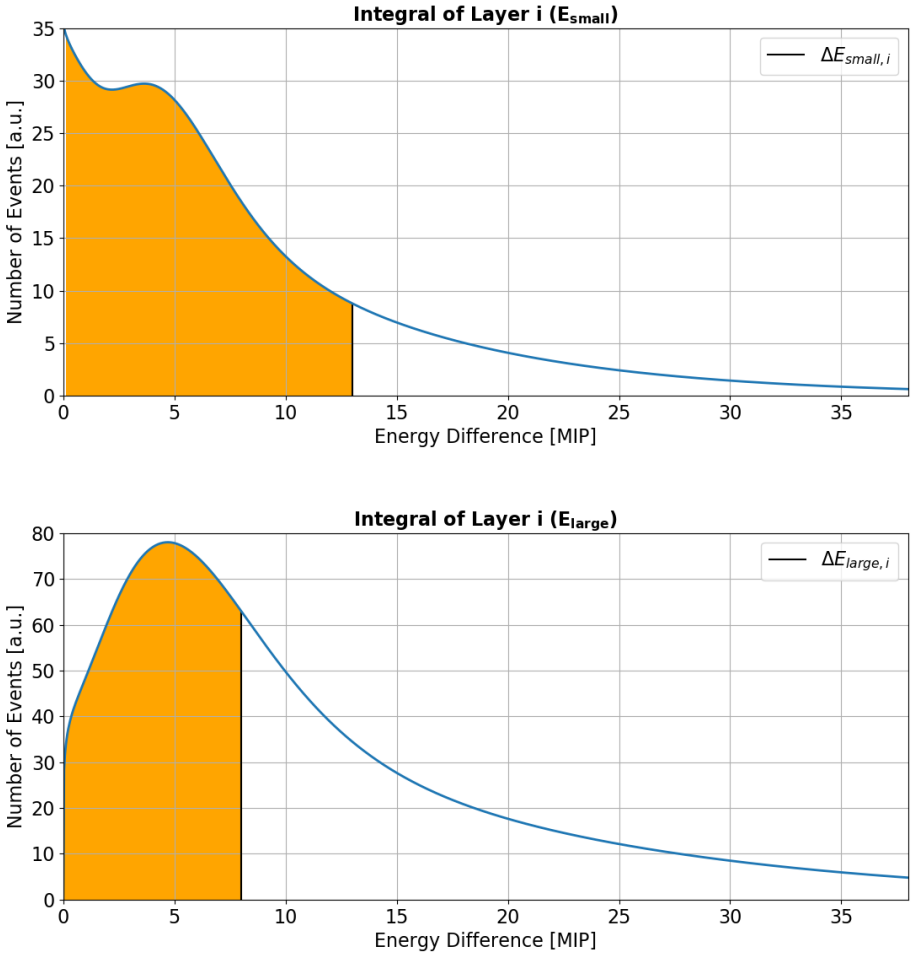


Figure 6.1.: Example plots of integrated arbitrary energy difference distributions. The upper curve corresponds to the integration of the PDF of E_{small} in layer i, until $\Delta E_{\text{small, }i}$ is reached. The lower curve corresponds to the integration of the PDF in the same layer but for E_{large}, until both orange shaded areas are equal, which is the case at $\Delta E_{\text{large, }i}$.

Next, the cumulative longitudinal energy difference distributions of E_{large} were considered. The cumulative distribution of layer i was used to determine the energy difference at which its y-value (i.e. the area integrated from the left of the respective energy differ-

ence PDF) equals A_i. This was also done for each layer, yielding another set of 33 energy differences ($\Delta E_{\text{large, 0}}$, $\Delta E_{\text{large, 1}}$, ..., $\Delta E_{\text{large, 32}}$). This process corresponds to integrating the energy difference PDF of E_{large} in layer i from the left, until the covered area equals A_i, as schematically shown in the lower plot in Figure 6.1.

To obtain interpolated energy differences, the two sets of energy differences were averaged pairwise:

$$\Delta E_{\text{interpolate, } i} = \frac{\Delta E_{\text{small, } i} + \Delta E_{\text{large, } i}}{2} \,. \tag{6.1}$$

Doing this for all 33 layers yields one complete, interpolated event for the target energy $E_{\text{interpolate}}$. The whole procedure was then also repeated vice versa, meaning that the indices "small" and "large" were now interchanged. In the end, another event was obtained for $E_{\text{interpolate}}$. Both procedures were done 100 000 times, resulting in a total of 200 000 events.

Of course, all of the previous steps can be conducted analogously if E_{small} and E_{large} are not equidistant from $E_{\text{interpolate}}$. Equation (6.1) is then generalised to be

$$\Delta E_{\text{interpolate, } i} = w_{\text{small}} \cdot \Delta E_{\text{small, } i} + w_{\text{large}} \cdot \Delta E_{\text{large, } i} \,, \tag{6.2}$$

where w_{small} and w_{large} satisfy

$$w_{\text{small}} + w_{\text{large}} = 1 \,. \tag{6.3}$$

In this generalised case, w_{small} and w_{large} are now unequal weights. Depending on the difference to $E_{\text{interpolate}}$, larger weights are chosen for closer energies, and vice versa. For this thesis,

$$w_i = 1 - \frac{|E_{\text{interpolate}} - E_i|}{E_{\text{large}} - E_{\text{small}}} \tag{6.4}$$

has been used, where the index i can be either "small" or "large". With this methodology, the (anti-)correlations between calorimeter layers are preserved, since each energy difference $\Delta E_{\text{large, } i}$ is determined as a function of all energy differences $\Delta E_{\text{small, } i}$, and vice versa, which means that all energy difference dependencies are considered during the interpolation.

6.2. Distributions of Interpolated Individual Shower Energies

First interpolation attempts were conducted on three neighbouring, equidistant initial energies, which excluded 10 GeV and 200 GeV as target energies, since they mark the low and

high energy "ends", respectively, of the whole dataset. Only the energies $E_{\text{interpolate}} = \{20, 30, 40, 60, 80, 120, 160\}$ GeV were used as target energies, following the procedure described in the previous Section. Results of the interpolation procedure are shown in Figure 6.2 for 60 GeV (interpolated from 40 GeV and 80 GeV) and in Figure 6.3 for 120 GeV (interpolated from 80 GeV and 160 GeV) pions which compare interpolated PDFs of simulated energy differences with their directly simulated expectations. Furthermore, the corresponding histograms of E_{small} and E_{large} are also shown. By comparing the interpolation with expectations, one can notice very good agreement between the curves. Minor fluctuations around the maxima of the distributions can be seen, however. Apart from these fluctuations, though, the shapes of the interpolation curves match those that have been directly simulated from KDEs.

In addition to distributions of interpolated, simulated energy differences, one can also compare correlation factors between interpolation and expectation by looking at Figures 6.4 (60 GeV pions) and 6.5 (120 GeV pions). These Figures also show very good agreement between each other, and the correct implementation of the interpolation becomes even more apparent by considering differences of correlation factors. These were calculated via

$$\Delta C = \text{Corr}_{\text{interpolation}}(x, y) - \text{Corr}_{\text{KDE}}(x, y) \tag{6.5}$$

and are depicted in Figure 6.6, also for 60 GeV and 120 GeV. In this Figure, both plots show correlation differences close to zero, which demonstrates that the interpolation is not only able to recreate distributions of simulated energy differences correctly, but also preserves correlations and anticorrelations of these energy differences between different layers.

The interpolation also does not seem to deteriorate too much if the target energy is kept constant, but E_{small} and E_{large} are altered (while keeping them equidistant). This can be seen by comparing Figure 6.3 with Figure 6.7, the latter showing interpolated PDFs for 120 GeV pions with $E_{\text{small}} = 60$ GeV and $E_{\text{large}} = 200$ GeV. Even though the difference between $E_{\text{interpolate}}$ and E_{small} (as well as E_{large}) was increased from 40 GeV in Figure 6.3 to 60 GeV in Figure 6.7, the performance of the interpolation in Figure 6.7 is still acceptable. However, the shapes around the maxima do not match exactly between interpolation and expectation, and all interpolated distributions are slightly shifted to larger energy differences. This confirms that it is better to base the interpolation on energies as close to $E_{\text{interpolate}}$ as possible.

Similar results for simulated energy difference distributions and correlation heatmaps can be obtained if non-equidistant initial energies are used for the interpolation. For example, Figures 6.8 (linear scale) and 6.9 (logarithmic scale) compare results of two

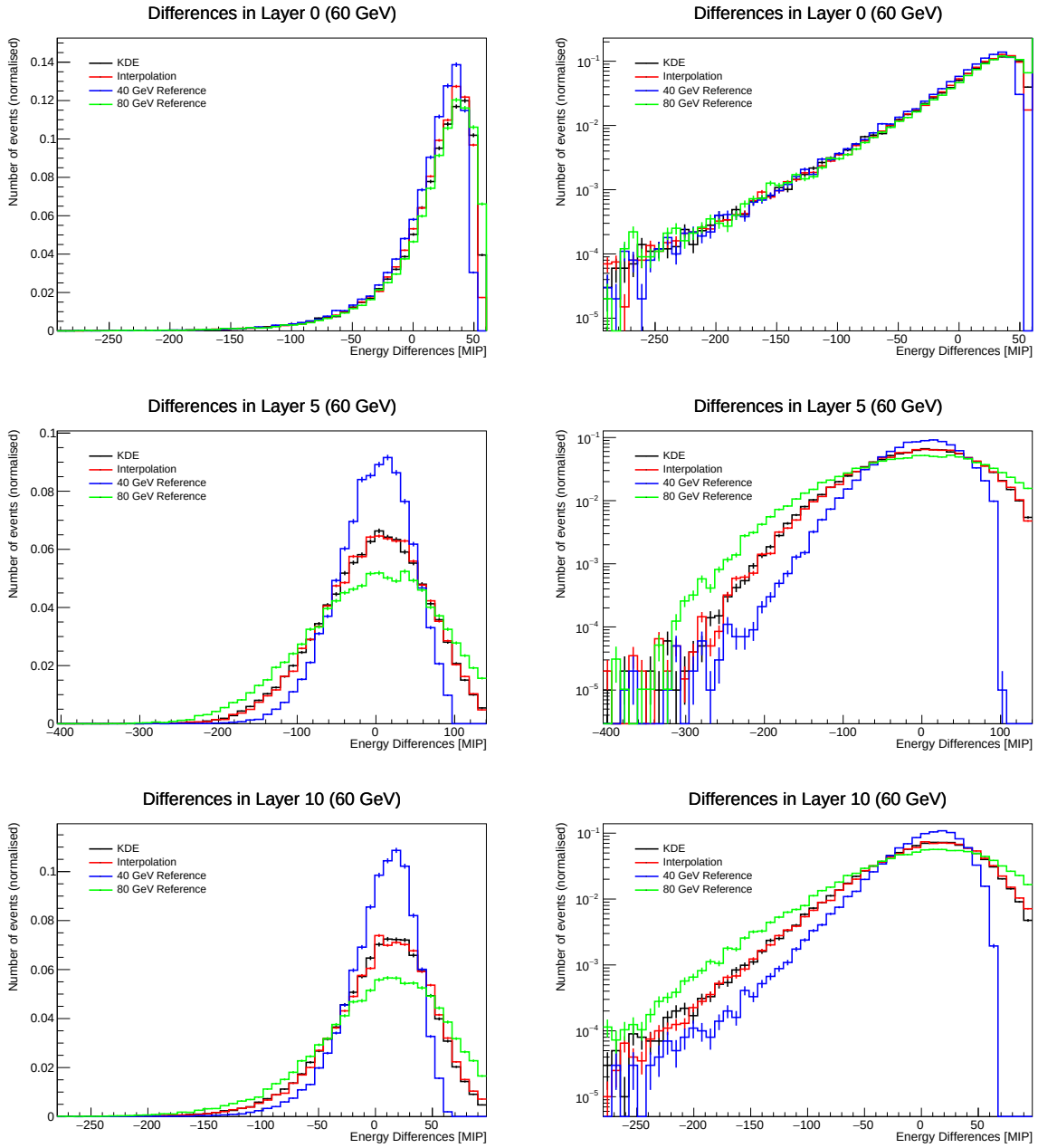


Figure 6.2.: Comparisons of interpolated simulation PDFs (black) with expected simulation PDFs (red) in layers 0, 5, and 10 for 60 GeV pions with linear scale on the left- and logarithmic scale on the right-hand side. The distributions of $E_{\text{small}} = 40\,\text{GeV}$ and $E_{\text{large}} = 80\,\text{GeV}$, from which interpolations were conducted, are also shown in blue and green, respectively.

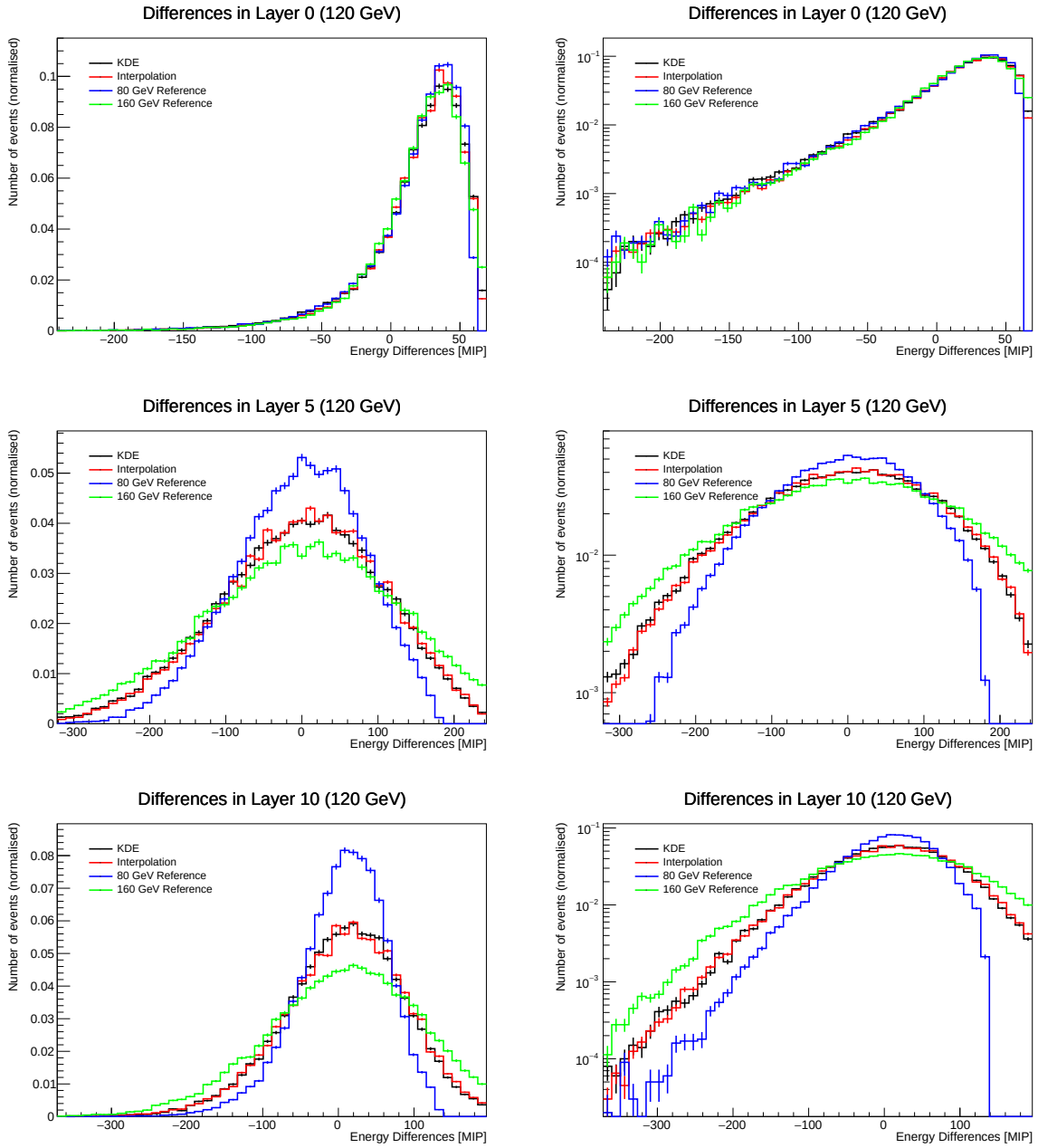


Figure 6.3.: Comparisons of interpolated simulation PDFs (black) with expected simulation PDFs (red) in layers 0, 5, and 10 for 120 GeV pions with linear scale on the left- and logarithmic scale on the right-hand side. The distributions of $E_{\text{small}} = 80\,\text{GeV}$ and $E_{\text{large}} = 160\,\text{GeV}$, from which interpolations were conducted, are also shown in blue and green, respectively.

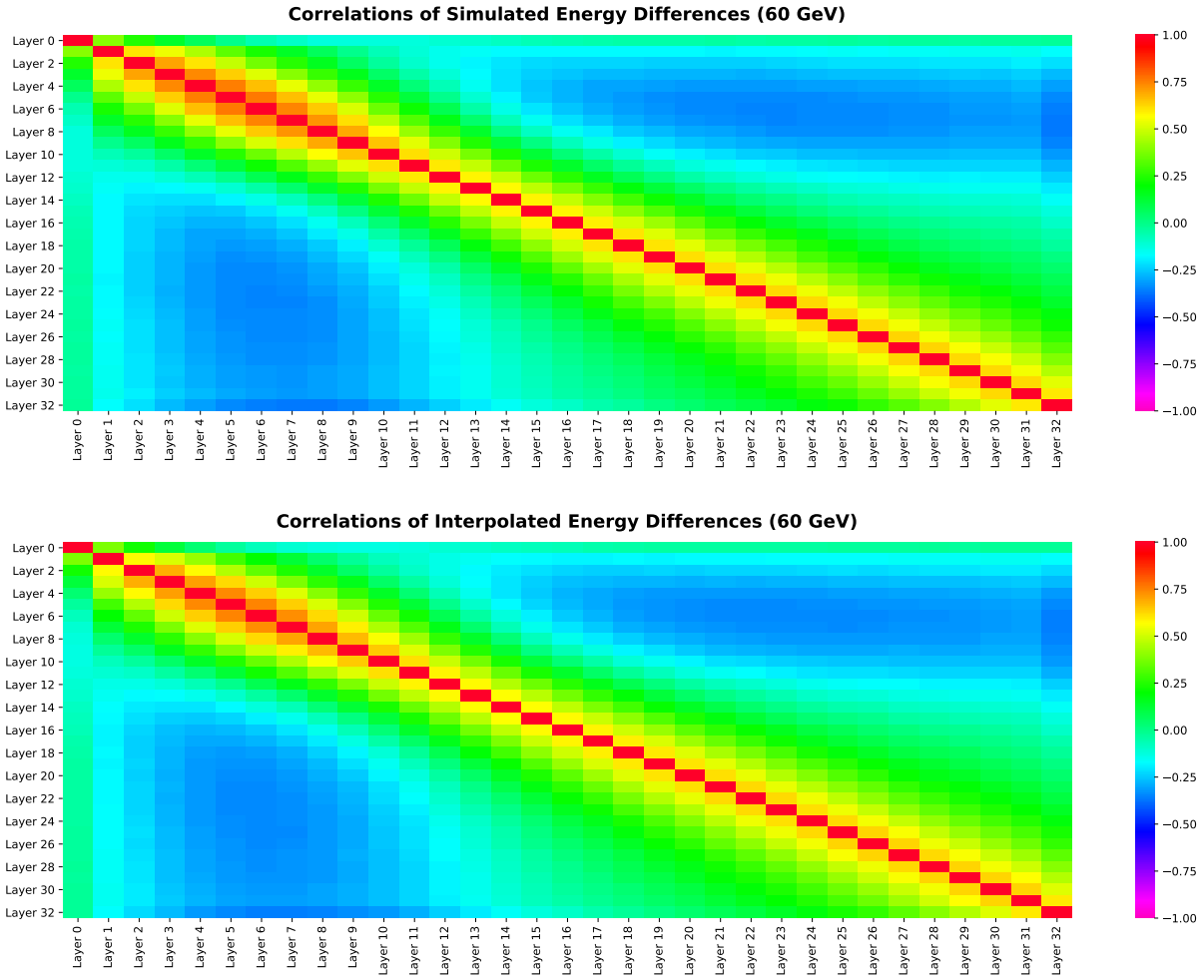


Figure 6.4.: Comparison of correlation factors between expectation (upper plot) and interpolation (lower plot) for 60 GeV pions. Both plots show very good agreement between each other.

80 GeV interpolations where one uses $E_{\text{small}} = 40\,\text{GeV}$ and $E_{\text{large}} = 120\,\text{GeV}$, whereas the other uses $E_{\text{small}} = 60\,\text{GeV}$ and $E_{\text{large}} = 120\,\text{GeV}$. For both equidistant energies, $w_{\text{small}} = w_{\text{large}} = \frac{1}{2}$ was assumed, whereas for the non-equidistant interpolation, $w_{\text{small}} = \frac{2}{3}$ and $w_{\text{large}} = \frac{1}{3}$. On both linear as well as logarithmic scale, one can see that the KDE and the interpolated PDFs are in very good agreement, and deviations between the equidistant and non-equidistant curves are small. Furthermore, the left-hand sides of Figures 6.8 and 6.9, together with Figure 6.3, also show that the interpolation still performs well for constant distances from $E_{\text{interpolate}}$, while the target energy is decreased or increased.

In Figures 6.10 and 6.11, which show comparisons of correlation factors between interpolation and expectation for 80 GeV pions, one can notice that both the equidistant ($E_{\text{small}} = 40\,\text{GeV}$, $E_{\text{large}} = 120\,\text{GeV}$) and non-equidistant ($E_{\text{small}} = 60\,\text{GeV}$, $E_{\text{large}} =$

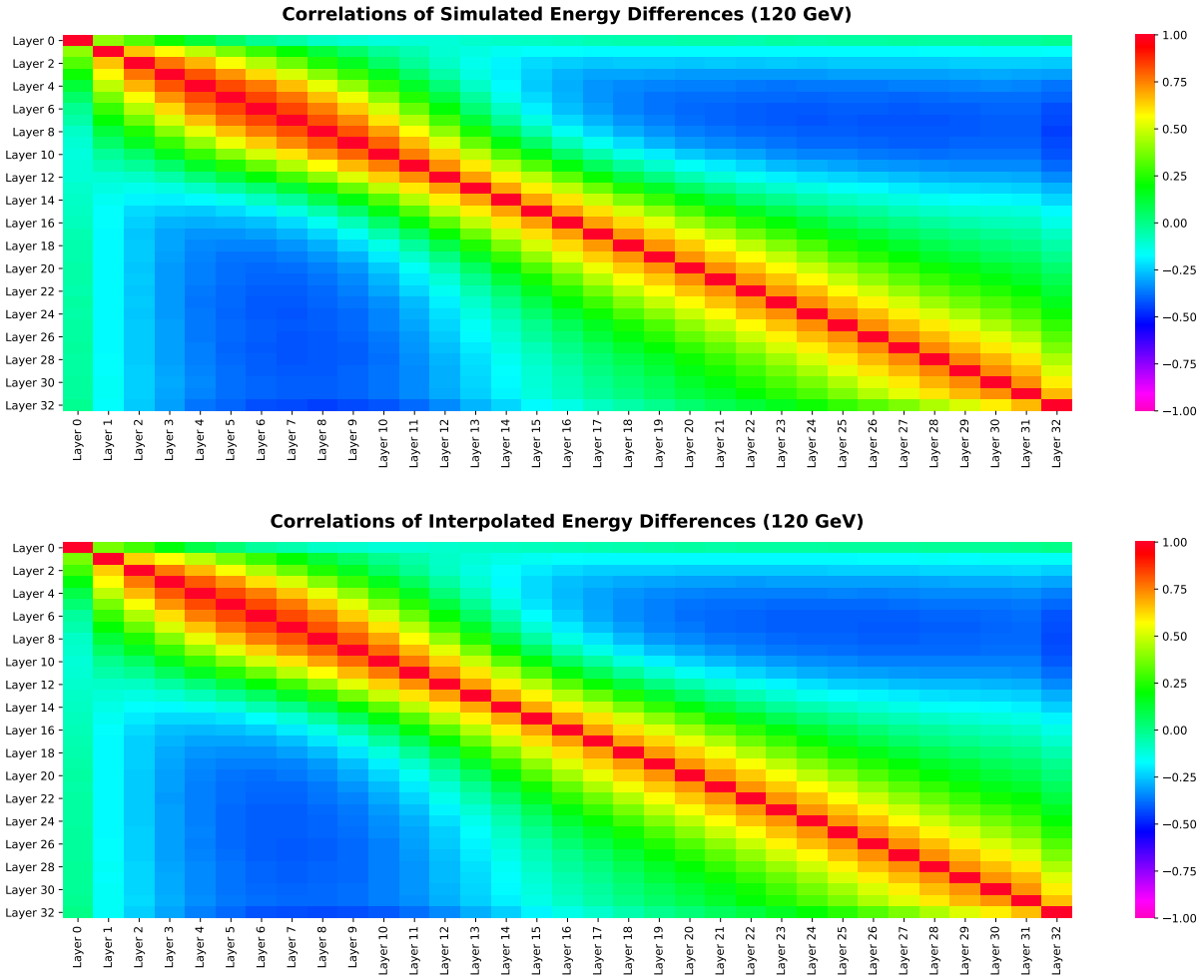


Figure 6.5.: Comparison of correlation factors between expectation (upper plot) and interpolation (lower plot) for 120 GeV pions. Both plots show very good agreement between each other.

120 GeV) interpolations are able to reproduce the correlation factors correctly. One can also see, however, that the non-equidistant interpolation performs slightly better than the equidistant one, recognisable by, for example, the broadening of the red diagonal in Figure 6.11, which is closer to expectation than Figure 6.10.

In addition to the previous paragraph, Figure 6.12 compares correlation factor differences (computed according to Equation (6.5)) of the equidistant and non-equidistant interpolations. In both cases, all differences are close to zero. However, even though both heatmaps depict correctly interpolated, simulated correlation factors, the equidistant heatmap is not as evenly coloured as its non-equidistant counterpart, recognisable by light blue shades around the diagonal, indicating negative correlation differences. Therefore, in this specific example the non-equidistant interpolation seems to perform better at preserv-

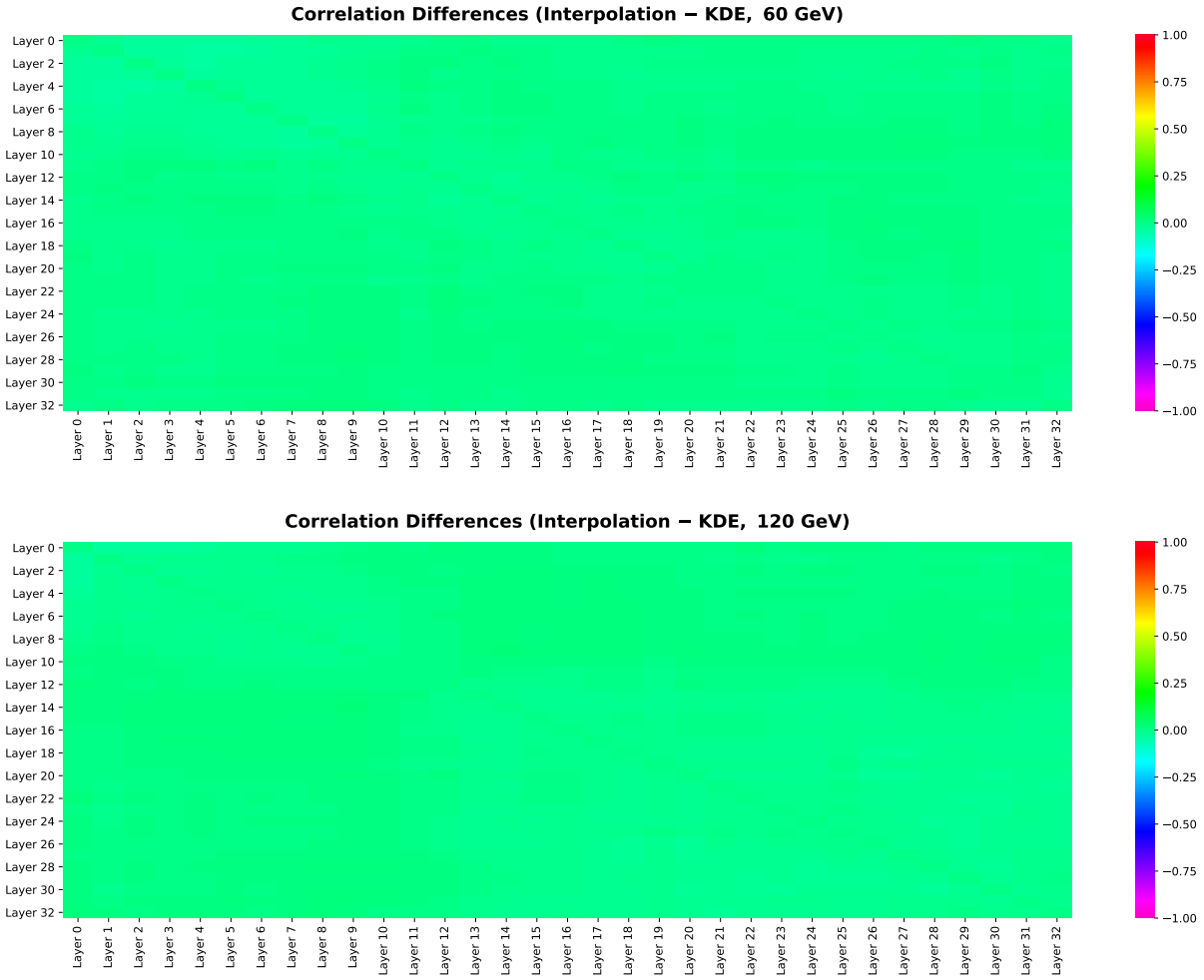


Figure 6.6.: Correlation differences, ΔC, between interpolation and expectation for 60 GeV (upper plot) and 120 GeV (lower plot) pions. Both plots show correlation differences very close to zero for all possible layer combinations.

ing correlation factors than the equidistant interpolation, which is due to $E_{\text{small}} = 60\,\text{GeV}$ having a larger influence on the final distributions than $E_{\text{small}} = 40\,\text{GeV}$. In general, though, the performance of the interpolation depends on the choice of E_{small} and E_{large}, not on whether an equidistant or non-equidistant interpolation was chosen. Nevertheless, a better performance of the non-equidistant interpolation is not recognisable in Figures 6.8 and 6.9, and since the overall improvement is only minor, one can safely assume that both interpolations work almost equally well.

To summarise, the presented interpolation method performs very well. Interpolated distributions of simulated energy differences show very good agreement with their expectations, though this is only the case if the distance between $E_{\text{interpolate}}$ and E_{small} as well as E_{large} is as small as possible. For larger distance, the interpolation worsens but only slowly.

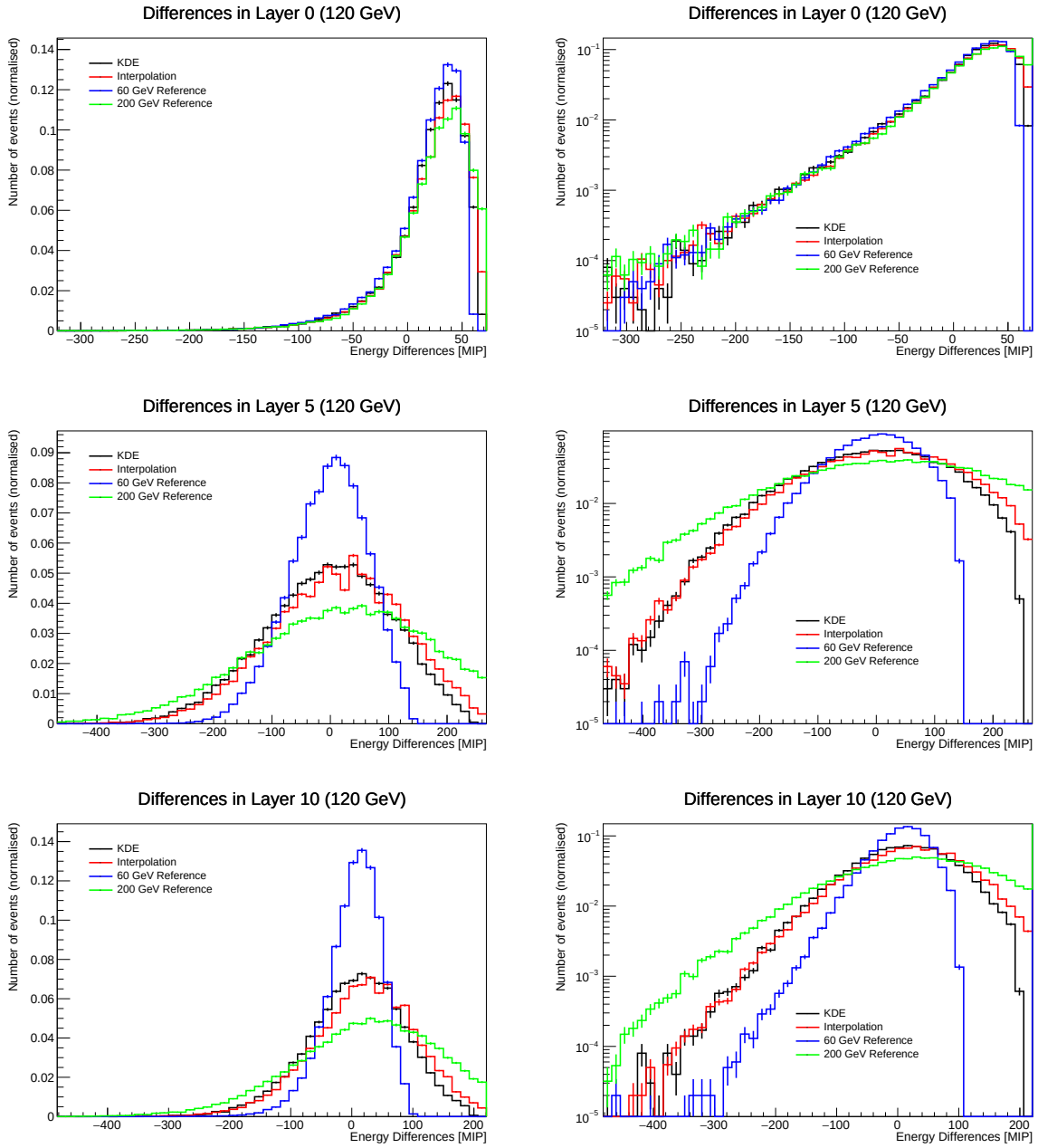


Figure 6.7.: Comparisons of interpolated simulation PDFs (black) with expected simulation PDFs (red) in layers 0, 5, and 10 for 120 GeV pions with linear scale on the left- and logarithmic scale on the right-hand side. The distributions of $E_{\text{small}} = 60\,\text{GeV}$ and $E_{\text{large}} = 200\,\text{GeV}$, from which interpolations were conducted, are also shown in blue and green, respectively.

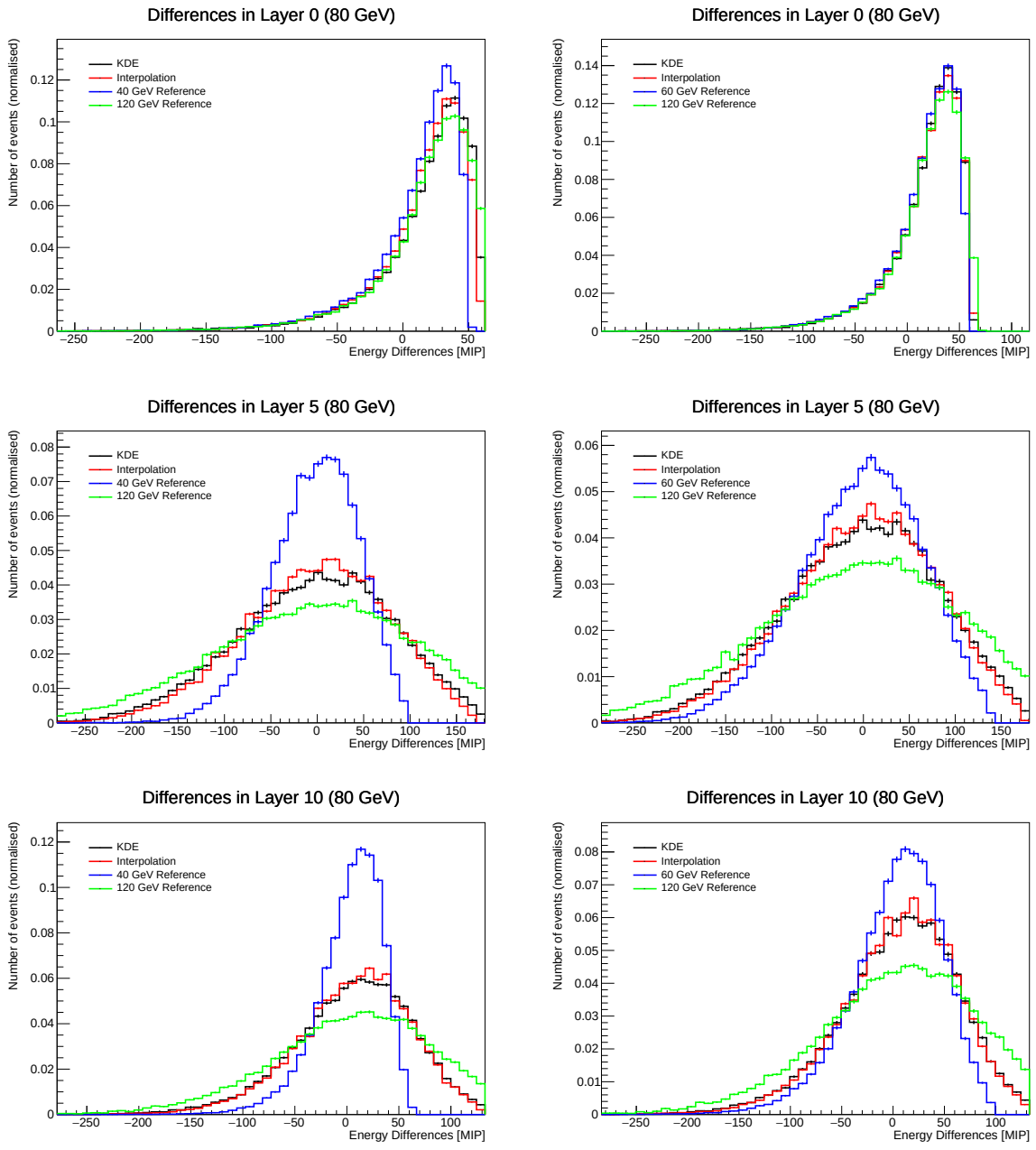


Figure 6.8.: Comparison of simulated energy difference distributions between equidistant and non-equidistant interpolations in layers 0, 5, and 10 for 80 GeV pions. The left-hand side shows simulated energy difference PDFs from the equidistant interpolation ($E_{\text{small}} = 40\,\text{GeV}$, $E_{\text{large}} = 120\,\text{GeV}$), and the right-hand side shows the same results but for the non-equidistant case ($E_{\text{small}} = 60\,\text{GeV}$, $E_{\text{large}} = 120\,\text{GeV}$). All plots are shown with linearly scaled y-axes.

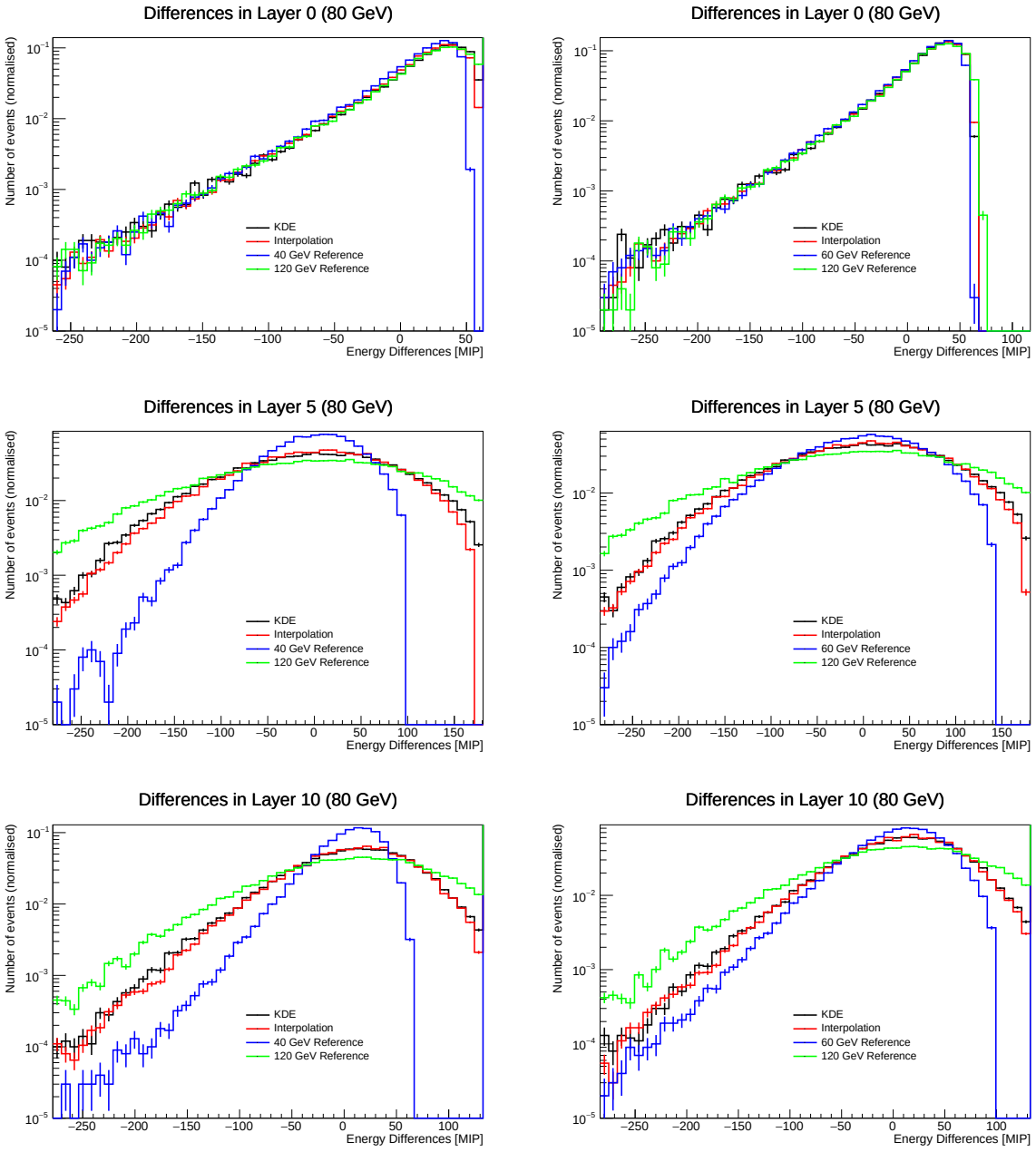


Figure 6.9.: Comparison of simulated energy difference distributions between equidistant and non-equidistant interpolations in layers 0, 5, and 10 for 80 GeV pions. The left-hand side shows simulated energy difference PDFs from the equidistant interpolation ($E_{\text{small}} = 40\,\text{GeV}$, $E_{\text{large}} = 120\,\text{GeV}$), and the right-hand side shows the same results but for the non-equidistant case ($E_{\text{small}} = 60\,\text{GeV}$, $E_{\text{large}} = 120\,\text{GeV}$). All plots are shown with logarithmically scaled y-axes.

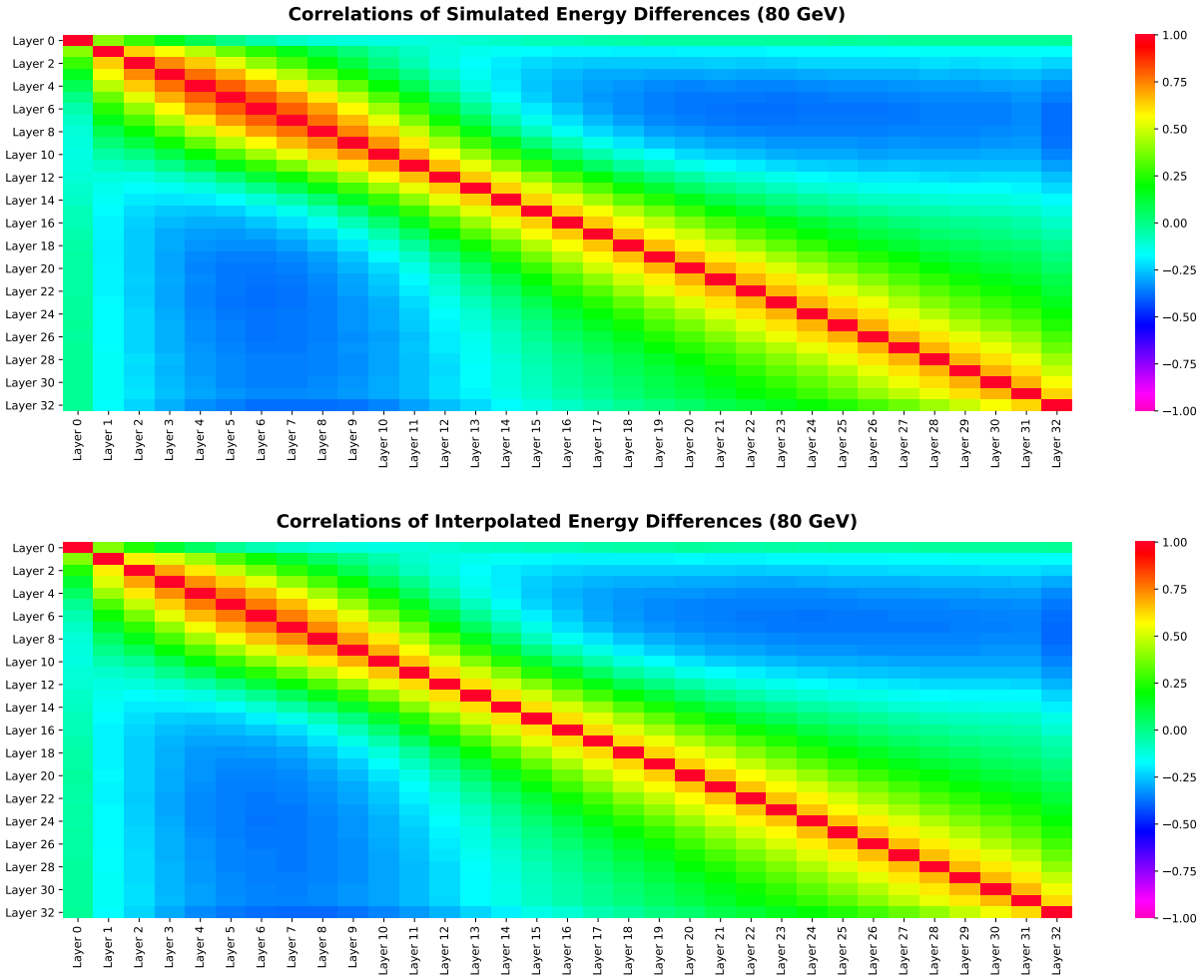


Figure 6.10.: Comparison of correlation factors between expectation (upper plot) and interpolation (lower plot) for 80 GeV pions. The interpolation was conducted with equidistant initial energies ($E_{\text{small}} = 40\,\text{GeV}$ and $E_{\text{large}} = 120\,\text{GeV}$). Both plots show good agreement between each other.

As has been shown, even distances ($|E_{\text{small}} - E_{\text{interpolate}}|$ as well as $|E_{\text{large}} - E_{\text{interpolate}}|$) up to 60 GeV still gave acceptable results. Interpolated correlation factors also showed very good agreement with expected correlation factors. Furthermore, the correlation factors showed that interpolations with non-equidistant energies also work as expected and that Equation (6.2), together with Equation (6.4), seems to be the correct, generalised method for interpolating simulated energy differences. Therefore, in order to be able to chose energies as close to the target energy as possible, a non-equidistant interpolation can be used as well.

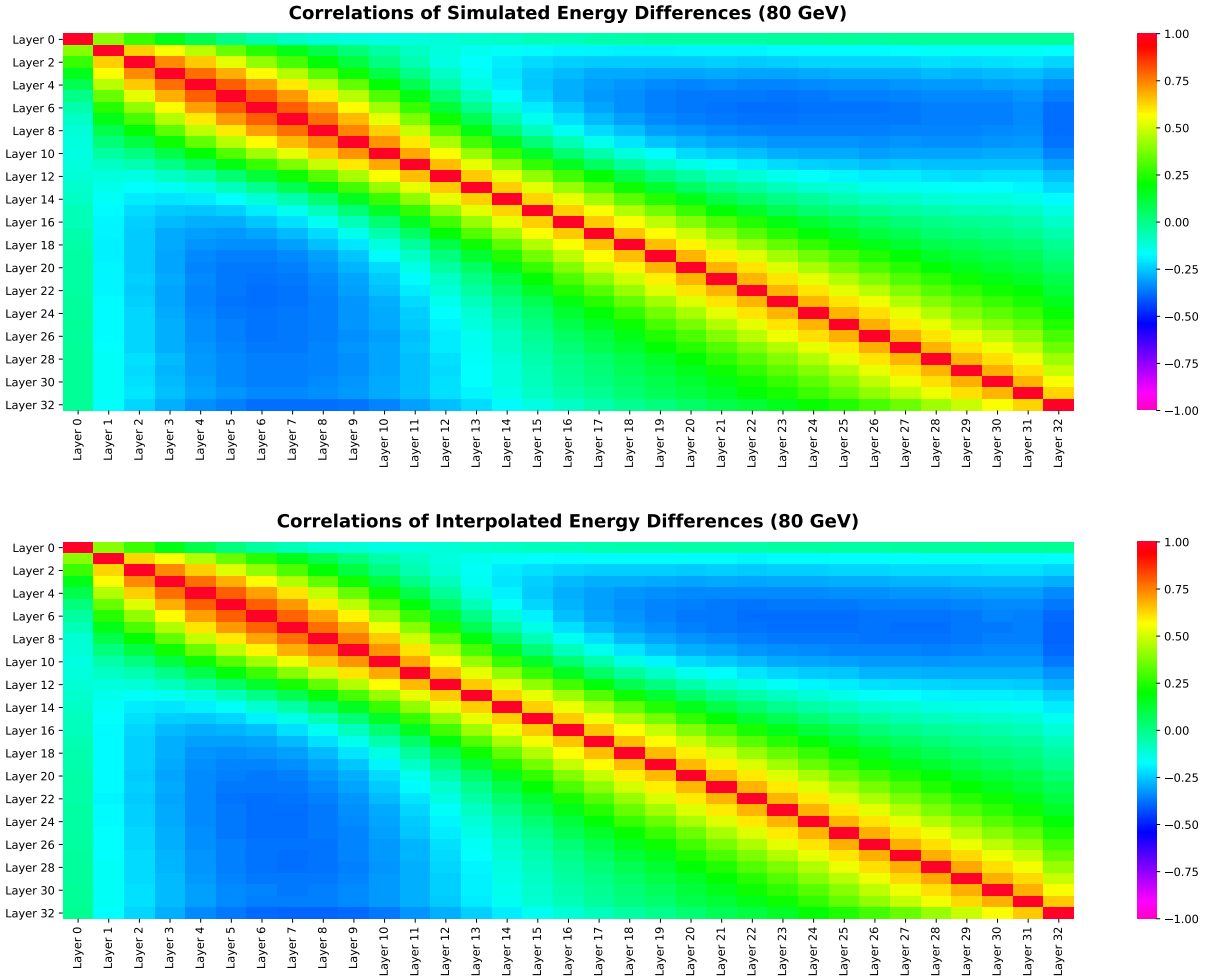


Figure 6.11.: Comparison of correlation factors between expectation (upper plot) and interpolation (lower plot) for 80 GeV pions. The interpolation was conducted with non-equidistant initial energies ($E_{\text{small}} = 60\,\text{GeV}$ and $E_{\text{large}} = 120\,\text{GeV}$). Both plots show even better agreement between each other compared to those depicted in Figure 6.10.

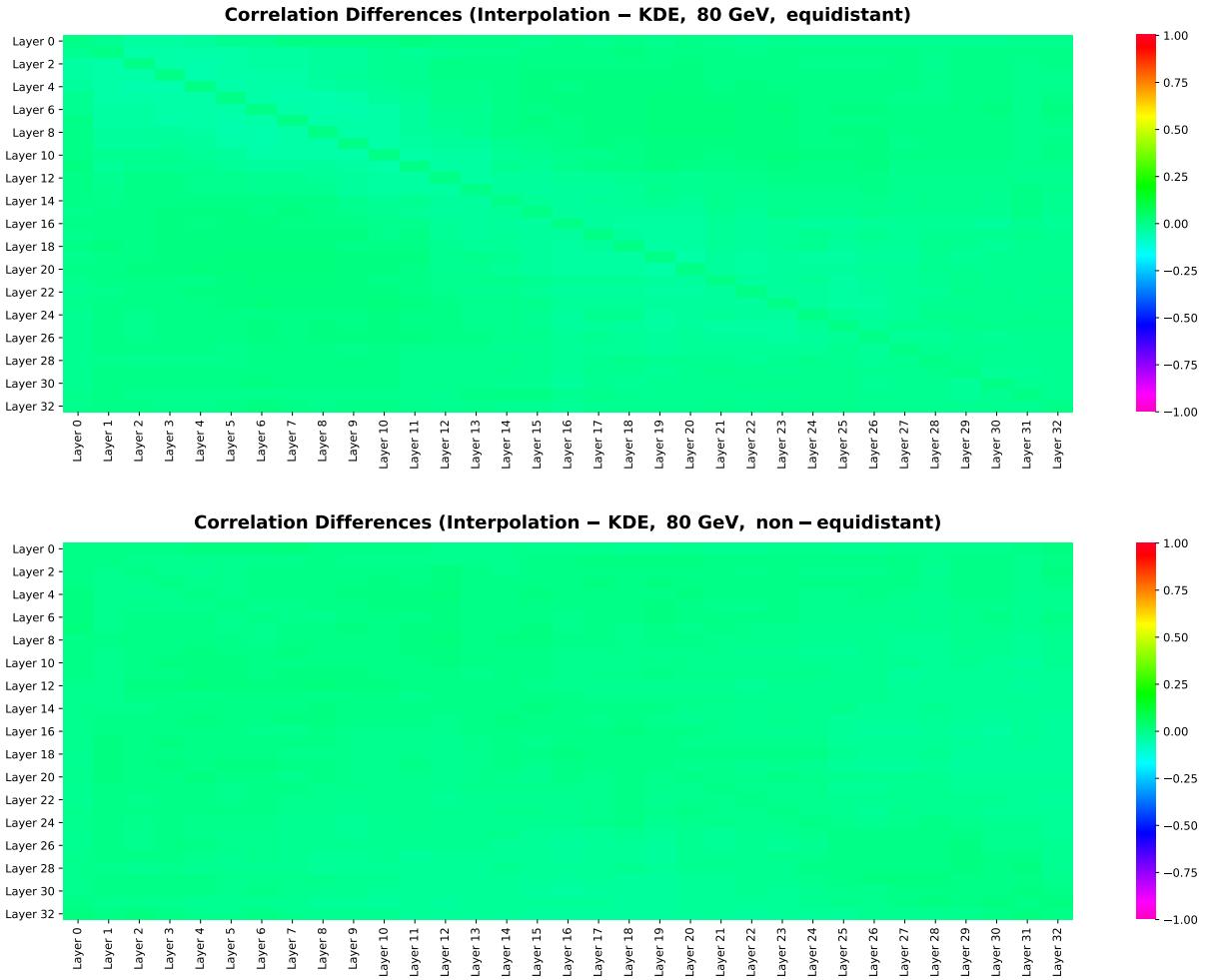


Figure 6.12.: Correlation differences, ΔC, between interpolation and expectation for the equidistant (upper plot; $E_{\text{small}} = 40\,\text{GeV}$, $E_{\text{large}} = 120\,\text{GeV}$) and non-equidistant (lower plot; $E_{\text{small}} = 60\,\text{GeV}$, $E_{\text{large}} = 120\,\text{GeV}$) interpolation for 80 GeV pions. Both heatmaps show correlation differences very close to zero for all possible layer combinations. Yet, the non-equidistant plot is more evenly coloured, indicating slightly better performance.

7. Conclusion

An investigation of fast hadron shower simulation methods was presented in this thesis. The fast simulation was implemented based on a pion shower dataset recorded in June 2018 at CERN by the AHCAL group of the CALICE Collaboration with the AHCAL Technological Detector Prototype. In total, the dataset comprised nine initial pion energies, ranging from 10 GeV to 200 GeV. From this dataset, differences in longitudinal energy distributions between single pion showers and a parameterisation of average pion showers were calculated.

A PCA was then conducted, for which all energy differences were transformed into uncorrelated principal components of which only the first eight were kept for every energy. The remaining ones were discarded. Distributions of principal components were then created and used as input for a random number generator. With this generator, principal components were simulated for each initial energy. These simulated principal components were finally transformed back into simulated energy differences and compared with data. The comparison showed that the PCA did not yield ideal results. Distributions of simulated energy differences exhibited shapes that were too broad around their maxima, too few negative energy differences, as well as unphysical values, corresponding to negative absolute energies. Furthermore, correlation factors between layers deviated too much between simulation and data. The results also did not improve significantly if no principal components were rejected or if fit functions, which agreed well with principal component PDFs, were used as input for the random number generator. Based on this, the conclusion that a PCA would not be further used for simulating pion showers was drawn.

A second method of simulating energy differences was examined, namely the application of Gaussian KDEs to the aforementioned energy differences. For this, multidimensional Gaussian normal distributions, centred at each data point, were summed and normalised to a chosen bandwidth. After that, the resulting PDFs were used for simulating energy differences. The simulations showed very good agreement with data, and correlation factors between energy differences were accurately preserved too. Furthermore, the KDE simulations were also able to recreate a pion shower's kinematic behaviour correctly. In

particular, distributions of the total shower energy and the centre of gravity along the z-axis of the detector showed little to no deviations between simulation and expectation, both around their maxima and their tails.

Based on the results of the KDE application, interpolations of simulated energy difference distributions were conducted on three initial pion energies, first for three equidistant, neighbouring initial energies, which yielded very good agreement between interpolation and expectation, both for the distributions of simulated energy differences as well as for their correlation factors. The distance between the initial energies was then increased, while keeping the target energy $E_{\text{interpolate}}$ constant, which showed that an increased distance deteriorated the interpolation, but not significantly. Doing this vice versa (constant distance, but altering the target energy) showed as good results as the equidistant, neighbouring case did (for both the PDFs as well as the correlation factors). Lastly, the interpolation was conducted with unequal weights on three neighbouring, but non-equidistant initial energies. The results of this method are in agreement with their respective expectations too, suggesting that the closer the initial energies are, the better the interpolation between them will be. Thus, the chosen interpolation method also seems suitable for interpolations to energies for which neither data nor simulations are available. Therefore, in summary, the simulation and interpolation of single pion showers with the help of KDEs works well and gives the expected results.

In order to improve this simulation even further, pre-shower energies also have to be considered, and bias due to the event selection mentioned in Chapter 4 needs to be minimised. If these two conditions are fulfilled, the analysis can be analogously conducted for radial energy distributions of pion showers, which would allow both simulations to be merged into one. Correlations between longitudinal and radial variables could then be investigated, and in the near future, the implementation of a complete, properly functioning fast simulation of pion showers, able to simulate individual cell hits, for instance, should become feasible. Moreover, the fast simulation is not restricted to pions only, but can be similarly extended to other particle types, such as electrons, for example, by considering test beam data of the corresponding particle. In a similar manner to what has been presented in this thesis, energy difference distributions and correlation factor plots can be simulated via KDEs and compared with data.

With calorimeters becoming more and more granular, simulations of particle showers have to become as precise as never before. With these high standards, the requirement of computational resources is growing steadily, as well as the need for resource-saving alternatives. Data-based fast simulations provide such alternatives by encapsulating the most crucial shower information, while neither relying on long computation times nor

large amounts of working storage. Their importance grows larger and larger with each year, and with more test beam runs and more recorded data, it should therefore soon become possible to predict the behaviour of electromagnetic as well as hadronic showers with unprecedented high precision and comparatively low effort.

Future plans for the design and construction of an International Linear Collider, where electron-positron collisions are going to take place, are already in development. It is supposed to reach a centre-of-mass energy of 500 GeV at a total length of approximately 34 kilometres. Such a collider would provide a great opportunity for CALICE, for example, to fully deploy the potential of its highly granular detector prototypes. Detecting particles and measuring their properties with very high precision requires finer and finer detectors, such as the AHCAL Technological Prototype, and with more and more data collected, scientists might soon gain insight into the physics of the smallest building blocks of nature as deeply as never before.

Bibliography

[1] A. Einstein, *Die Grundlage der allgemeinen Relativitätstheorie*, Annalen der Physik **354**, 769 (1916)

[2] M. Gell-Mann, *A Schematic Model of Baryons and Mesons*, Phys. Lett. **8**, 214 (1964)

[3] H. Fritzsch, M. Gell-Mann, H. Leutwyler, *Advantages of the color octet gluon picture*, Phys. Lett. B **47**, 365 (1973)

[4] S. L. Glashow, *Partial-symmetries of weak interactions*, Nuclear Physics **22**, 579 (1961)

[5] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19**, 1264 (1967)

[6] A. Salam, *Weak and Electromagnetic Interactions*, Conf. Proc. C **680519**, 367 (1968)

[7] P. W. Higgs, *Broken symmetries, massless particles and gauge fields*, Phys. Lett. **12**, 132 (1964)

[8] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13**, 508 (1964)

[9] F. Englert, R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13**, 321 (1964)

[10] G. S. Guralnik, C. R. Hagen, T. W. B. Kibble, *Global Conservation Laws and Massless Particles*, Phys. Rev. Lett. **13**, 585 (1964)

[11] ATLAS Collaboration, *Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC*, Phys. Lett. B **716**, 1 (2012)

[12] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B **716**, 30 (2012)

[13] H. Kolanoski, N. Wermes, *Teilchendetektoren: Grundlagen und Anwendungen*, Springer-Verlag Berlin Heidelberg (2016)

[14] Y.-S. Tsai, *Pair production and bremsstrahlung of charged leptons*, Rev. Mod. Phys. **46**, 815 (1974)

[15] R. L. Workman et al., *Review of Particle Physics*, PTEP **2022**, 083C01 (2022)

[16] E. Longo, I. Sestili, *Monte Carlo calculation of photon-initiated electromagnetic showers in lead glass*, Nuclear Instruments and Methods **128**, 283 (1975)

[17] M. Livan, R. Wigmans, *Calorimetry for Collider Physics, an Introduction*, Springer International Publishing (2019)

[18] N. Marinelli, P. Merkel, *LHC Collisions Every 25 Nanoseconds*, `https://cms.cern/news/lhc-collisions-every-25-nanoseconds`, Accessed: 2022-07-27

[19] S. Lee, *On the limits of the hadronic energy resolution of calorimeters*, J. Phys.: Conf. Ser. **1162**, 012043 (2019)

[20] CALICE Collaboration, *Hadron shower decomposition in the highly granular CALICE analogue hadron calorimeter*, JINST **11**, P06013 (2016)

[21] O. Pinto, *Operation and Calibration of a Highly Granular Hadron Calorimeter with SiPM-on-Tile Read-out* (2020), arXiv:2004.00370 [physics.ins-det]

[22] CALICE Collaboration, *SiPM-on-tile HCAL R&D*, `http://flc.desy.de/hcal/index_eng.html`, Accessed: 2022-07-29

[23] F. Sefkow, S. Frank, *A highly granular SiPM-on-tile calorimeter prototype*, J. Phys.: Conf. Ser. **1162**, 012012 (2019)

[24] B. Wang, J. Mu, *High-speed Si-Ge avalanche photodiodes*, PhotoniX **3** (2022)

[25] P. Eckert, H.-C. Schultz-Coulon, W. Shen, R. Stamen, A. Tadday, *Characterisation Studies of Silicon Photomultipliers*, Nucl. Instrum. Methods Phys. Res. A **620**, 217 (2010)

[26] N. Anfimov et al., *Study of Silicon Photomultiplier Performance at Different Temperatures*, Nucl. Instrum. Methods Phys. Res. A **997**, 165162 (2021)

[27] CMS HGCAL Collaboration, *Construction and commissioning of CMS CE prototype silicon modules* (2020), arXiv:2012.06336 [physics.ins-det]

[28] CMS HGCAL Collaboration, *The DAQ system of the 12,000 Channel CMS High Granularity Calorimeter Prototype*, JINST **16**, T04001 (2021)

[29] K. Kawagoe et al., *Beam test performance of the highly granular SiW-ECAL technological prototype for the ILC.*, Nucl. Instrum. Methods Phys. Res. A **950**, 162969 (2020)

[30] O. Pinto, *Studies of Average Shower Shapes with 2018 Testbeam Data*, `https://agenda.linearcollider.org/event/9326/contributions/48756/attachments/37108/58099/OP_CALICE_Meeting_10092021.pdf`, Presented at the CALICE AHCAL Main Meeting on December 8th, 2021, at DESY

[31] D. Heuchel, *Particle Flow Studies with Highly Granular Calorimeter Data*, Ph.D. thesis, University of Heidelberg, Heidelberg (2022), DOI: 10.11588/heidok.00031794

[32] O. Pinto, *Shower Shapes in a Highly Granular Analog Hadron Calorimeter*, Ph.D. thesis, University of Hamburg, Hamburg (2022), `https://ediss.sub.uni-hamburg.de/handle/ediss/9855`

[33] M. J. Oreglia, *A Study of the Reactions* $\psi' \to \gamma\gamma\psi$, Ph.D. thesis, SLAC, Stanford University, Stanford (1980), Reference Number: SLAC-R-236

[34] T. Skwarnicki, *A study of the radiative CASCADE transitions between the Upsilon-Prime and Upsilon resonances*, Ph.D. thesis, Cracow, INP, Cracow (1986), Reference Number: DESY-F31-86-02

A. Visualisation of different Bandwidth Choices

In order to visualise the impact of the bandwidth on the final distribution of a KDE, a set of 1000 values was generated according to the sum of two normal distributions centred at $x = 2$ and $x = 7$, respectively:

$$f(x) = \frac{1}{\sqrt{8\pi}} \cdot \left[\exp\left(-\frac{1}{2}(x-2)^2 \right) + \exp\left(-\frac{1}{2}(x-7)^2 \right) \right] . \tag{A.1}$$

Equation (A.1) is normalised to unity which is the reason why the scaling factor is equal to $\frac{1}{\sqrt{8\pi}}$ instead of $\frac{1}{\sqrt{2\pi}}$. Three KDEs with three different bandwidths ($h = \{5, 0.1, 0.0001\}$) were then applied to the 1000-points dataset, yielding the estimates shown in Figure A.1. One can see that $h = 0.1$ seems to be the best of the three choices, reproducing the true function the most accurately. $h = 5$, on the other hand, yields, as expected, a very flat, almost constant distribution, whereas $h = 0.0001$ exhibits many more peaks than the other two KDEs, clearly undersmoothing the PDF.

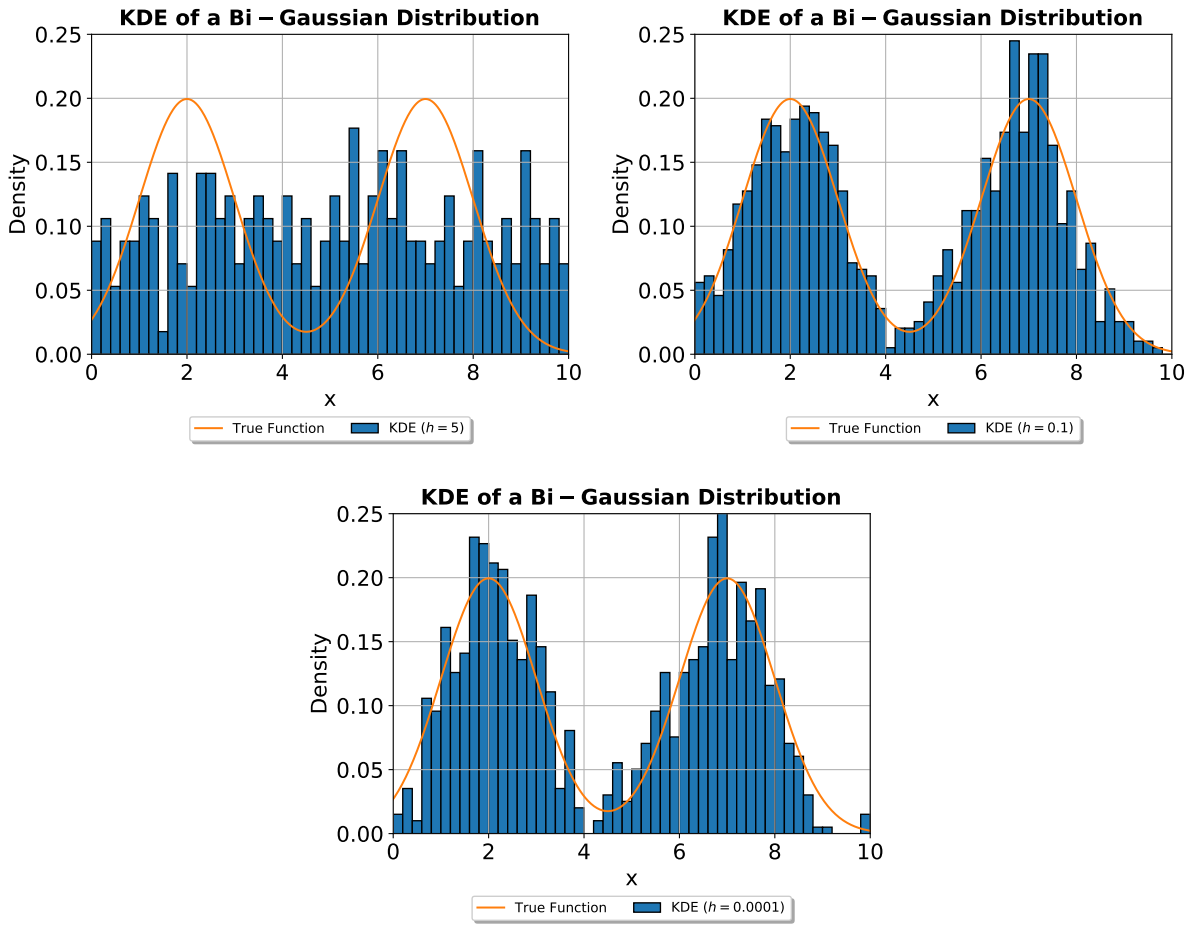


Figure A.1.: Examples of KDEs with different bandwidths. Three KDEs with different bandwidths were applied to a set of 1000 values, generated according to Equation (A.1), yielding the distributions shown above. Clearly, $h = 5$ ($h = 0.0001$) flattens (undersmooths) the distribution too much, whereas $h = 0.1$ reproduces the original function to a high accuracy and therefore seems to be the closest to the optimal choice of bandwidth.

Acknowledgements

Writing a Master's thesis is a very demanding process. It is neither accomplished easily nor can it be done single-handedly but only with a lot of support and help from outside. Yet, carrying out research and seeing my own effort bear fruits is a very fulfilling consequence, too, of being a scientist. Either way, successfully finishing this thesis would not have been possible without many people who stood by my side throughout my time at the II. Institute of Physics in Göttingen.

First of all, thanks a lot to you, Julian, for being my go-to person for any kind of technical question, for reading through my Master's thesis multiple times, as well as for being a great office partner. Time flew during our joint work at the faculty, and I really hope that we will be able to spend more time together in the near future.

I would also like to thank the AHCAL group of the CALICE Collaboration for their warm welcome at the beginning of my research, for the fruitful discussions about my progress during the weekly AHCAL meetings, and for providing me the test beam dataset which I had been working with during my time of research. In particular, I want to thank Katja Krüger who introduced me to the group, who was always there for me for any kind of detector-related question, and who gave me the opportunity to participate in the 2022 test beam run at CERN. The latter really was an outstandingly interesting first time that I spent in Geneva, and I am also thankful for being able to make new contacts and find new friends and colleagues at this special occasion.

Special thanks also go to Olin Pinto whose average shower shape studies are an inherent part of the foundation upon which my own research is built. In addition, he also provided me self-written code which greatly simplified the event selection of my analysis. Moreover, I want to thank Jack Rolph and Erik Buhmann for the many, very helpful discussions about the application of kernel density estimators to my research and for providing me code too.

I also owe a great debt of gratitude to Prof. Dr. Stanley Lai. It has been more than two years from now since I joined your working group for the first time, and I have never since regretted this decision. I love the working atmosphere in your group, the effort you daily put into looking and caring for your students, and the fact that you never let your

students feel as if they were working for you but with you as a fellow scientist. Doing research for my Master's thesis was a time I greatly enjoyed, and I am looking forward to continue my work, in particular with you, at the II. Institute of Physics in Göttingen.

Finally, I want to thank my parents and sisters for their continuous moral and financial support. Everyday, I feel blessed to have a family like you, and I surely do not take your support for granted. Without you I would definitely not have been able to finish studying physics, and I cannot put into words how thankful I am for having you as my family. Furthermore, I also want to thank all of my friends, both here in Göttingen and at home, for the great times we spent together and for distracting me from my (quite often) very stressful studies.

Erklärung nach §17(9) der Prüfungsordnung für den Bachelor-Studiengang Physik und den Master-Studiengang Physik an der Universität Göttingen:
Hiermit erkläre ich, dass ich diese Abschlussarbeit selbstständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe.
Darüberhinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, im Rahmen einer nichtbestandenen Prüfung an dieser oder einer anderen Hochschule eingereicht wurde.

Göttingen, den 24. November 2022

(André Wilhahn)